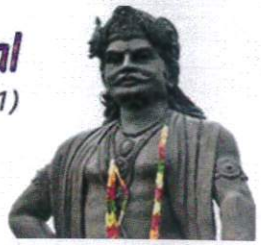
Madhya Pradesh Bhoj (Open) University, Bhopal
(Established under an Act of State Assembly in 1991)
मध्यप्रदेश भोज (मुक्त) विश्वविद्यालय, भोपाल



SELF - LEARNING MATERIAL

MBA, First Year
Paper – II

MANAGERIAL ECONOMICS

MBA First Year

Paper II

MANAGERIAL ECONOMICS

मध्यप्रदेश भोज (मुक्त) विश्वविद्यालय – भोपाल

MADHYA PRADESH BHOJ (OPEN) UNIVERSITY - BHOPAL

COURSE WRITER

D.N. Dwivedi, Professor of Economics Maharaja Agrasen Institute of Management Studies, Delhi Former Reader, Ramjas College University of Delhi

Published by Registrar, MP Bhoj (open) University, Bhopal in 2020

VIKAS®

Vikas® is the registered trademark of Vikas® Publishing House Pvt. Ltd.

VIKAS® PUBLISHING HOUSE PVT. LTD.
E-28, Sector-8, Noida - 201301 (UP)
Phone: 0120-4078900 • Fax: 0120-4078999
Regd. Office: A-27, 2nd Floor, Mohan Co-operative Industrial Estate, New Delhi 1100 44
• Website: www.vikaspublishing.com • Email: helpline@vikaspublishing.com

SYLLABI-BOOK MAPPING TABLE

Managerial Economics

Syllabi	Mapping in Book
UNIT - I 1. Managerial Economics–Nature, Scope and Concepts 2. Demand Analysis	**Unit-1:** The Fundamentals of Managerial Economics and Demand Analysis **(Pages 3-85)**
UNIT - II 3. Production Analysis 4. Cost Analysis	**Unit 2:** Production Analysis and Cost Analysis **(Pages 87-148)**
UNIT - III 5. Price-Output and Market Structure 6. Pricing Strategies	**Unit 3:** Market Structure and Pricing Strategies **(Pages 149-240)**
UNIT - IV 7. Investment Decisions 8. Firm Objectives and Constraints	**Unit 4:** Investment Decisions and Firm Objectives **(Pages 241-274)**
UNIT - V 9. Advertising 10. Welfare Economics	**Unit 5:** Advertising and Welfare Economics **(Pages 275-294)**

CONTENTS

INTRODUCTION

NOTES

The main functions of business managers are to take appropriate decisions on business matters, to manage and organize resources and to make optimum use of available resources with the objective of achieving predetermined business goals. In today's world, business decision-making has become an extremely complex task due to the ever-growing complexity of the business world and business environment. It is in this context that modern economics—howsoever defined—contributes a great deal towards business decision-making and the performance of managerial duties and responsibilities. Just as biology contributes to the medical profession and physics to engineering, economics contributes to the managerial profession.

This book, *Managerial Economics*, aims at equipping management students with concepts in economics, economic theories and tools and techniques of economic analysis applied to business decision-making. The reason is that making an appropriate business decision requires a clear understanding of market conditions, market fundamentals and the business environment. This requires extensive and intensive analysis of market conditions for products, inputs and financial markets

This book has been written in the Self-Instructional Mode (SIM) wherein each unit begins with an Introduction to the topic followed by an outline of the Objectives. The detailed content is then presented in a simple and an organized manner, interspersed with Check Your Progress questions to test the understanding of the students. A Summary along with a list of Key Terms and a set of Self-Assessment Questions and Exercises is also provided at the end of each unit for effective recapitulation.

UNIT 1 THE FUNDAMENTALS OF MANAGERIAL ECONOMICS AND DEMAND ANALYSIS

Structure

1.0 INTRODUCTION

Businesses need to make crucial decisions on a day-to-day basis. These decisions can be about an investment opportunity, a new product, a new competitor or the direction of a company. For such important decisions, businesses need to rely on experts. These experts come from the background of Managerial Economics. Managerial economists get to sit at the table with the executives, rather than be a part of the executive branch of the company. They are the experts who provide monetary value to the different opportunities and then urge the company to proceed. Managerial economics is a stream of management studies that emphasizes primarily solving business problems and decision-making by applying the theories and principles of microeconomics and macroeconomics. It is a specialized stream dealing with an organization's internal issues by using various economic theories. Economics is an indispensable part of any business. All the business assumptions, forecasting, and investments are derived from this single concept. It requires a lot of logical

thinking and creative skills for decision-making or problem-solving. The other important part of a business is demand. A firm must research for a certain product in the market with the aim of profit maximisation. This analysis is known as demand analysis. Since a business would not be able to exist if there was improper demand estimate or demand forecast, hence demand analysis is one of the most important aspects of managerial economics and it is studied in great detail. This unit will explain the nature and scope of managerial economics. It will also discuss the concept of demand analysis in detail.

1.1 OBJECTIVES

After going through this unit, you will be able to:

- Explain the nature and scope of managerial economics
- Discuss various concepts related to managerial economics
- Describe the law of demand and its elements
- Examine the concept of demand elasticities
- Explain demand forecasting techniques

1.2 MANAGERIAL ECONOMICS—NATURE, SCOPE AND CONCEPTS

In this section, you will learn about the meaning, nature, scope and basic concepts of managerial economics.

1.2.1 Meaning and Nature of Managerial Economics

Managerial economic can be defined as *the study of economic theories, logic, concepts and tools of economic analysis applied in the process of business decision-making.* In general practice, economic theories and techniques of economic analysis are applied to diagnose the business problems and to evaluate alternative options and opportunities open to the firm for finding an optimum solution to the problems. Look at some other definitions of managerial economics offered by some economists.

Mansfield: "Managerial economics is concerned with the application of economic concepts and economics to the problem of formulating rational decision making".

Spencer and Seigelman: "Managerial economics is the integration of economic theory with business practice for the purpose of facilitating decision making and forward planning by management."

Davis and Chang: "Managerial Economics applies the principles and methods of economics to analyse problems faced by management of a business, or other types of organizations, and to help find solutions that advance the best of such organizations."

Douglas: "Managerial economics is concerned with the application of economic principles and methodologies to the decision making process within the firm or

organization. It seeks to establish rules and principles to facilitate the attainment of the desired goal of management."

As these definitions reveal, managerial economics is an integration of economic science with decision making process of business management. The integration of economic science with management has become inevitable because application of economic theories and analytical tools make significant contribution to managerial decision-making.

As we know, the basic managerial functions are planning, organizing, staffing, leading and controlling business related factors. The ultimate objective of these managerial functions is to ensure maximum return from the utilization of firm's resources. To this end, managers have to take decisions at each stage their functions in view of business issues and implement decisions effectively to achieve the goals of the organization. As we will see later, almost all managerial decision issues involve economic analysis and analytical techniques. Therefore, economic theories and analytical tools are applied as a means to find solution to the business issue. This is how economics gets integrated to managerial functions and gives emergence of managerial economics. The integration of economics with business management is illustrated in Figure 1.1.

In Figure 1.1, Block 1 shows the major areas of business decisions making. Taking decision on all of these business problems involves economic consideration. For example, choice of the product requires assessment of demand and supply conditions of the perspective products. This requires application of theories of demand and supply. Similarly, all other decision problems require the application of relevant *economic concepts, theories and analytical tools* to find ways and means to arrive at an appropriate solution to the problem.

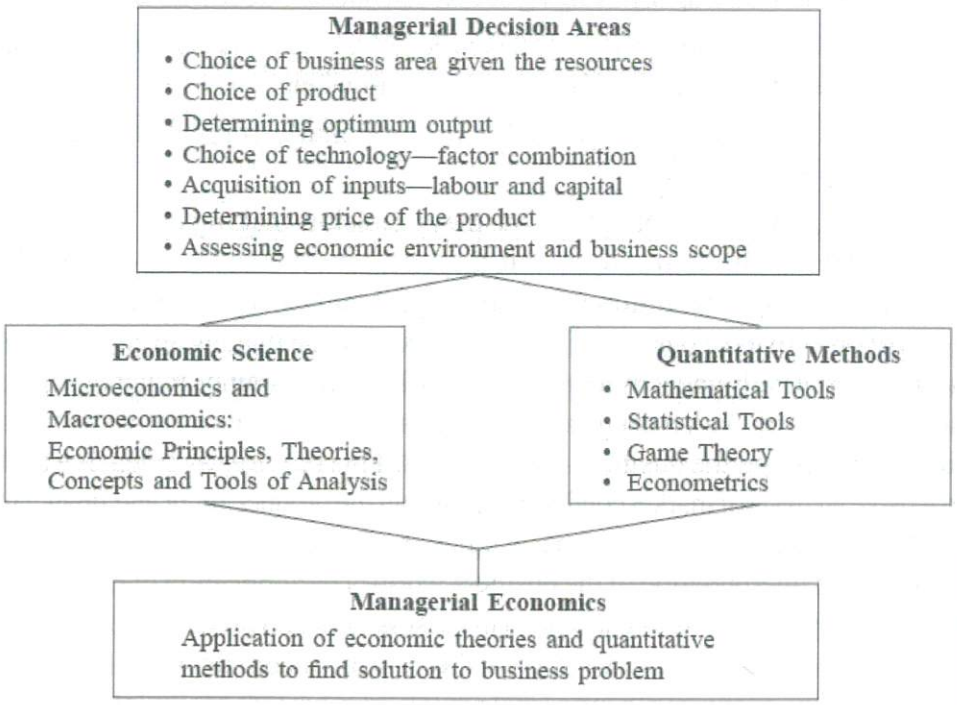


***Fig. 1.1** Integration of Economics with Managerial Decisions*

However, application of economic concepts and theories alone is not sufficient to make a specific decision. It has to be combined with *quantitative methods* to find a numerical solution to the decisions problems. For example, once the choice of product is finalized, next question arises 'how much to produce' to optimize the output. To find answer to this question, quantitative methods (mathematical models and statistics as shown in Block 2) have to be combined with the theories of production and cost. It means that to make a sound decision, economic concepts and theories have to be integrated with quantitative methods and models. The integration of economic theories and concepts with quantitative methods creates *managerial economics.*

It may be added at the end that economic science has a very wide perspective. All economic theories are neither applicable nor are applied to business decision-making. Most business management issues are of internal nature and a significant part of microeconomics deals with internal decision-making issues of the business firms—what to produce, how to produce, how much to produce, and what price to charge, etc. That is why most microeconomic theories and analytical tools are generally applied to managerial decision-making. Therefore, *managerial economics is* treated as *applied microeconomics.* Macroeconomics *deals with environmental issues*—how is the economic condition of the country; what is the likely trend; what are government's economic policies; how government policies might affect business environment of the country; what kind of business policy will be required, and so on.

The emergence of managerial economics as a separate branch of economic study can be attributed to the following three factors:

(i) Growing complexity of business decision-making due to rapid expansion of business and rapidly changing market conditions and business environment

(ii) An increasing use of economic theory, logic, concepts and tools of economic analysis in the process of business decision-making

(iii) Rapid increase in demand for professionally trained managerial manpower

The growing complexity of business decision-making has inevitably increased the application of economic theories, concepts and tools of analysis to find an appropriate solution to business problems, in conformity with the goals of objectives of business firms. The reason is that making an appropriate decision, especially under the conditions of risk and uncertainty, requires a clear understanding of market conditions, nature and degree of competition, market fundamentals and the business environment. This requires an intensive and extensive analysis of market conditions regarding the product, input and financial markets. On the other hand, the basic function of economics is to offer logical analysis and interpretation of the business world. The application of economic theories, logic and polls of analysis to the assessment and prediction of market conditions has proved to be a great help in business decision-making. Consequently, economic theories and analytical tools that are widely used in business decision-making have crystallized as a separate and specialized branch of study, called 'managerial economics'.

Let us now see how managerial economics can be defined.

Definition of managerial economics

Managerial economics is essentially that part of economics which is applied to business decision-making. As noted above, economic science is a body of knowledge consisting of economic concepts, logic and reasoning, economic laws and theories, and tools and techniques for analysing economic phenomena, evaluating economic options and for optimizing allocation of resources. The part of economic science that is applied to business analysis and decision-making constitutes managerial economics.

Managerial economics can be defined more comprehensively. Managerial economics is the study of economic theories, concepts, logic and tools of analysis that are used in analysing business conditions with the objective of finding an appropriate solution to business problems. As regards its nature, managerial economics is essentially applied micro economics.

1.2.2 Scope of Managerial Economics

In general, the scope of managerial economics comprehends all those economic concepts, theories and tools of analysis which can be used to analyse the business environment and to find solutions to practical business problems. In other words, managerial economics is economics applied to the analysis of business problems and decision-making. Broadly speaking, it is applied economics.

The areas of business issues to which economic theories can be directly applied may be broadly divided into two categories: (*a*) operational or internal issues, and (*b*) environment or external issues.

(A) Microeconomics applied to operational issues

Operational problems are of an internal nature. They include all those problems which arise within the business organization and fall within the purview and control of the management. Some of the basic internal issues are: (*i*) choice of business and the nature of product, i.e., what to produce, (*ii*) choice of size of the firm, i.e., how much to produce, (*iii*) choice of technology, i.e., choosing the factor-combination, (*iv*) choice of price, i.e., how to price the commodity, (*v*) how to promote sales, (*vi*) how to face price competition, (*vii*) how to decide on new investments, (*viii*) how to manage profit and capital, and (*ix*) how to manage inventory, i.e., stock of both finished goods and raw materials. These problems may also figure in forward planning. Microeconomics deals with these questions and the like confronted by managers of the business enterprises. The microeconomic theories which deal with most of these questions are the following:

1. **Theory of demand:** Demand theory explains the consumer's behaviour. It answers the questions: How do consumers decide whether or not to buy a commodity? How do they decide on the quantity of a commodity to be purchased? When do they stop consuming a commodity? How do consumers behave when price of the commodity, their income and tastes and fashions, change? At what level of demand, does changing price become inconsequential in terms of total revenue? The knowledge of demand theory can, therefore, help in choosing commodities for production.

2. **Theory of production and production decisions:** Production theory, also called 'Theory of Firm', explains the relationship between inputs and output. It also explains under what conditions costs increase or decrease; how total output increases when units of one factor (input) are increased keeping other factors constant, or when all factors are simultaneously increased; how output can be maximized from a given quantity of resources; and how optimum size of output can be determined. Production theory, thus, helps in determining the size of the firm, size of the total output and the amount of capital and labour to be employed.
3. **Analysis of market-structure and pricing theory:** Price theory explains how prices are determined under different market conditions; when price discrimination is desirable, feasible and profitable; to what extent advertising can be helpful in expanding sales in a competitive market. Thus, price theory can be helpful in determining the price policy of the firm. Price and production theories together, in fact, help in determining the optimum size of the firm.
4. **Profit analysis and profit management:** Profit making is the most common objective of all business undertakings. But making a satisfactory profit is not always guaranteed because a firm has to carry out its activities under conditions of uncertainty with regard to (*i*) demand for the product, (*ii*) input prices in the factor market, (*iii*) nature and degree of competition in the product market, and (*iv*) price behaviour under changing conditions in the product market. Therefore, an element of risk is always there even if the most efficient techniques are used for predicting future and even if business activities are meticulously planned. The firms are, therefore, supposed to safeguard their interest and avert, as far as possible, the possibilities of risk or try to minimize it. Profit theory guides firms in the measurement and management of profit, in making allowances for the risk premium, in calculating the pure return on capital and pure profit and also for future profit planning.
5. **Theory of capital and investment decisions:** Capital, like all other inputs, is a scarce and expensive factor. Capital is the foundation of business. Its efficient allocation and management is one of the most important tasks of managers and a determinant of the success level of the firm. The major issues related to capital are (*i*) choice of investment project, (*ii*) assessing the efficiency of capital, and (*iii*) most efficient allocation of capital. Knowledge of capital theory can contribute a great deal in investment decision-making, choice of projects, maintaining capital intact, capital budgeting, etc.

(B) Macroeconomics applied to business environment

Environmental issues pertain to the general business environment in which a business operates. They are related to the overall economic, social and political atmosphere of the country. The factors which constitute the economic environment of a country include the following:

(i) The type of economic system of the country

(ii) General trends in production, employment, income, prices, saving and investment

NOTES

(iii) Structure of and trends in the working of financial institutions, e.g., banks, financial corporations and insurance companies

(iv) Magnitude of and trends in foreign trade

(v) Trends in labour and capital markets

(vi) Government's economic policies, e.g., industrial policy, monetary policy, fiscal policy and price policy

(vii) Social factors like the value system of the society, property rights, customs and habits

(viii) Social organizations like trade unions, consumers' cooperatives and producers unions

(ix) Political environment which comprises such factors as political system—democratic, authoritarian, socialist, or otherwise,—state's attitude towards private business, size and working of the public sector and political stability

(x) The degree of openness of the economy and the influence of MNCs on the domestic markets

The environmental factors have a far-reaching bearing upon the functioning and performance of firms. Therefore, business decision-makers have to take into account the changing economic, political and social conditions in the country and give due consideration to the environmental factors in the process of decision-making. This is essential because business decisions taken in isolation of environmental factors may not only prove infructuous, but may also lead to heavy losses. For instance, a decision to set up a new alcohol manufacturing unit or to expand the existing ones ignoring impending prohibition—a political factor—would be suicidal for the firm; a decision to expand the business beyond the paid-up capital permissible under the Monopoly and Restrictive Trade Practices Act (MRTP Act) amounts to inviting legal shackles and hammer; a decision to employ a highly sophisticated, labour-saving technology ignoring the prevalence of mass open unemployment—an economic factor—may prove to be self-defeating; a decision to expand the business on a large scale, in a society having a low per capita income and hence a low purchasing power stagnated over a long period, may lead to wastage of resources. The managers of a firm are, therefore, supposed to be fully aware of the economic, social and political conditions prevailing in the country while taking decisions on the wider issues of the business. let us study only the macroeconomic issues pertaining to business decision-making.

The major macroeconomic or environmental issues which figure in business decision-making, particularly with regard to forward planning and formulation of the future strategy, may be described under the following three categories:

1. **Issues related to macro variables:** There are issues that are related to the trends in macro variables, e.g., the general trend in the economic activities of the country, investment climate, trends in output and employment and price trends. These factors not only determine the prospects of private business, but also greatly influence the functioning of individual firms. Therefore, a firm planning to set up a new unit or expand its existing size

would like to ask itself: What is the general trend in the economy? What would be the consumption level and pattern of the society? Will it be profitable to expand the business? Answers to these questions and the like are sought through macroeconomic studies.

2. **Issues related to foreign trade:** An economy is also affected by its trade relations with other countries. The sectors and firms dealing in exports and imports are affected directly and more than the rest of the economy. Fluctuations in the international market, exchange rate and inflows and outflows of capital in an open economy have a serious bearing on its economic environment and, thereby, on the functioning of its business undertakings. The managers of a firm would, therefore, be interested in knowing the trends in international trade, prices, exchange rates and prospects in the international market. Answers to such problems are obtained through the study of international trade and international monetary mechanism.
3. **Issues related to government policies:** Government policies designed to control and regulate economic activities of the private business firms affect the functioning of private business undertakings. Besides, firms' activities as producers and their attempt to maximize private gains or profits lead to considerable social costs, in terms of environmental pollution, congestion in the cities, creation of slums, etc. Such social costs not only bring a firm's interests in conflict with those of the society, but also impose a social responsibility on the firms. The government's policies and its various regulatory measures are designed, by and large, to minimize such conflicts. Managers should, therefore, be fully aware of the aspirations of the people and give such factors a due consideration in their decisions. The forced closure of polluting industrial units set up in the residential areas of Delhi city and the consequent loss of business worth billions of rupees in 2000, is a recent example of the result of ignoring the public laws and the social responsibility of the businessmen. The economic concepts and tools of analysis help in determining such costs and benefits.

1.2.3 Significance of Managerial Economics Decision-Making

It is a well-known fact that with the growing complexity of the business environment, the usefulness of economic theory as a tool of business analysis has increased. Its contribution to the process of decision-making is a widely recognized fact today. In managerial economics, microeconomic analysis is a common tool for specific business decisions. This is why it bridges economic theory and economics in practice. Managerial economics makes good use of quantitative techniques like regression analysis and correlation techniques and Lagrangian linear calculus. Marginalization and incremental principle are also important techniques of managerial economics. Marginal analysis uses marginal changes in the dependent variable resulting from a unit change in its determinant and the independent variable. The incremental principle is applied to business decisions which involve a large increase in total cost and total revenue. Most

economic managers strive to optimize business decisions given their firm's objectives as well as constraints imposed by scarcity. Operations research and programming prove to be very handy in this.

The techniques of managerial economics can be used to analyse any business decision. However,these techniques are most frequently applied in case of the following:

(i) **Risk analysis:** It refers to a technique for identifying and assessing those factors or elements which have the potential to mar the success of a business enterprise. Through risk analysis, we can determine preventive measures can be determined so that these factors do not occur. Such an analysis can also help us decide counter measures to successfully deal with such probable obstacles. To determine or assess the riskiness of a business decision,we have the options to use several uncertainty models and risk quantification techniques.

(ii) **Production analysis:** Managerial economics techniques are used to analyse several factors relating to production of an enterprise, such as production efficiency, enterprise's cost function and optimum factor allocation.

(iii) **Pricing analysis:** It refers to examination and evaluation of a proposed price. It does not include the evaluation of its separate cost elements and proposed profit. Managerial economics techniques are very useful in analysing the various pricing decisions by policy decision-makers and business managers, such as transfer pricing, joint product pricing, price discrimination, price elasticity estimations. These techniques are also helpful in choosing the optimum pricing method.

(iv) **Capital budgeting:** Also called investment proposal, it refers to the planning process which is used to assess whether a firm's long-term investments are worth pursuing.These long-term investments could range from replacement of machinery to R&D projects. Business managers take the help of investment-related theories to determine an enterprise's capital puchasing decisions. Capital budgeting involves various methods such as net present value (NPV), internal rate of return (IRR), equivalent annuity, profitability index, and modified internal rate of return (MIRR).

Time perspective in business decisions

All business decisions are taken with a certain time perspective. The time perspective refers to the duration of time period extending from the relevant past and foreseeable future taken in view while taking a business decision. Relevant past refers to the period of past experience and trends which are relevant for business decisions with long-run implications. All business decisions do not have the same time perspective. Some business decisions have short-run repercussions and, therefore, involve short-run time perspective. A decision to buy explosive materials, for example, for manufacturing crackers involves short-run demand prospects. Similarly, a decision regarding building inventories of finished products involves a short-run time perspective.

There are, however, a large number of business decisions which have long-run repercussions, e.g., investment in a plant, building, machinery, land; spending on labour welfare activities; expansion of the scale of production; introduction of a new product; advertisement; bribing a government officer and investment abroad. The decision about such business issues may not be profitable in the short run but may prove very profitable in the long run. The introduction of a new product, for example, may not be profitable in the short run but may prove very profitable in the long run, e.g., the introduction of a new product may not catch up in the market quickly and smoothly. It may be difficult to even cover the variable costs. But in the long run it may result in a roaring business. Also, spending on labour welfare may enhance costs in the present scenario and may lead to a decline in profit. But in the long run, it may increase labour productivity in a much greater proportion as compared to an increase in cost. Therefore, while taking a business decision with long-term implications, it is immensely important to keep a well-worked out time perspective in view.

Business decision-makers must assess and determine the time perspective of business propositions well in advance and make decisions accordingly. Determination of time perspective is of great significance, especially where projections are involved. The decision-makers must decide on an appropriate future period for projecting the value of a variable. Otherwise, projections may prove meaningless from the analysis point of view and decisions based thereon may result in poor pay-offs. In a business decision, for example, regarding the establishment of a management institute, projecting a short-run demand and taking a short-run time perspective will be unwise, and in buying explosive materials for manufacturing crackers for Deepawali, a long-run perspective is unwise.

1.2.4 Role and Responsibility of a Managerial Economist

Economics contributes a great deal towards the performance of managerial duties and responsibilities; just as biology contributes to the medical profession and physics to engineering, economics contributes to the managerial profession. All other quali-fications being the same, managers with a working knowledge of economics can perform their functions more efficiently than those without it. The basic function of the managers of a business firm is to achieve the objective of the firm to the maximum possible extent with the limited resources placed at their disposal. The emphasis here is on the maximization of the objective and limitedness of the resources. Had the resources been unlimited, like sunshine and air, the problem of economizing on resources or resource management would have never arisen. But resources, howsoever defined, are limited. Resources at the disposal of a firm, whether finance, men or material, are by all means limited. Therefore, the basic task of the management is to optimize the use of the resources.

As mentioned above, economics, though variously defined, is essentially the study of logic, tools and techniques of making optimum use of the available resources to achieve the given ends. Economics thus provides analytical tools and

techniques that managers need to achieve the goals of the organization they manage. Therefore, a working knowledge of economics, not necessarily a formal degree, is essential for managers. Managers are essentially practising economists.

NOTES

In performing his functions, a manager has to take a number of decisions in conformity with the goals of the firm. Many business decisions are taken under the condition of uncertainty and risk. Uncertainty and risk arise mainly due to uncertain behaviour of the market forces, changing business environment, emergence of competitors with highly competitive products, government policy, external influence on the domestic market and social and political changes in the country. The complexity of the modern business world adds complexity to business decision-making. However, the degree of uncertainty and risk can be greatly reduced if market conditions are predicted with a high degree of reliability. The prediction of the future course of the business environment alone is not sufficient. What is equally important is to take appropriate business decisions and to formulate a business strategy in conformity with the goals of the firm.

Taking appropriate business decisions requires a clear understanding of the technical and environmental conditions under which business decisions are taken. Application of economic theories to explain and analyse the technical conditions and the business environment contributes a good deal to the rational decision-making process. Economic theories have, therefore, gained a wide range of application in the analysis of practical problems of business. With the growing complexity of the business environment, the usefulness of economic theory as a tool of analysis and its contribution to the process of decision-making has been widely recognized.

Baumol has pointed out three main contributions of the economic theory to business ecomomics.

First, 'one of the most important things which the economic theories can contribute to the management science' is building analytical models which help to recognize the structure of managerial problems, eliminate the minor details which might obstruct decision-making, and help to concentrate on the main issue.

Second, the economic theory contributes to the business analysis 'a set of analytical methods' which may not be applied directly to specific business problems, but they do enhance the analytical capabilities of the business analyst.

Third, economic theories offer clarity to the various concepts used in business analysis, which enables the managers to avoid conceptual pitfalls.

1.2.5 Basic Concepts

Lets discuss some of the basic concepts related to the understanding of managerial economics.

1. **Short and long-run:** The reference to time period involved in production process is another important concept used in production analysis. The two reference periods are short-run and long-run. The short-run refers to a period of time in which the supply of certain inputs (e.g., plant, building,

NOTES

machinery, etc.) is fixed or inelastic. In the short-run, therefore, production of a commodity can be increased by increasing the use of only variable inputs like labour and raw materials.

It is important to note that short-run and long-run are economists' jargon. They do not refer to any fixed time period. While in some industries short-run may be a matter of few weeks or few months, in some others (e.g., electric and power industry), it may mean three or more years.

Long-run refers to a period of time in which the supply of all inputs is elastic, but not enough to permit a change in technology. That is, in the long run, all inputs are variable. Therefore, in the long-run, production of a commodity can be increased by employing more of both variable and fixed inputs.

Economists use another term, i.e., very long-run which refers to a period in which the technology of production is supposed to change. In the very long-run, the production function also changes. The technological advances result in a larger output from a given quantity of inputs.

2. **Firm, plant and industry:** The difference between plant, firm and industry is not too difficult to understand. Before we try to understand how they are different, let us first try and define each of these terms.
   - A firm would be any business unit or enterprise— in the form of a partnership, sole proprietorship or a corporation—which is referred to by the name or title under which it transacts business.
   - An industry can be defined as a group of people, organizations or companies engaged in the same (or similar) commercial enterprise.
   - A plant is a site containing a structure housing various tools, machines, equipment, furniture, material, accessories and other items used in manufacturing products or providing services.

   In other words, the plant is where the product is actually manufactured while the firm is the company that is involved in business dealings related to that product. The industry is the collection of firms dealing in the same/ similar product. To explain this clearly, let us suppose that SuperPlast is a firm that deals in plastic goods. It has its manufacturing plant in Ghaziabad and is a well-established firm in the Indian plastic industry.
3. **Opportunity Cost:** The concept of opportunity cost is very important in economic and business analyses as it helps in making a choice between the alternative opportunities available to a person or to a business firm. Here, let us explain the concept of opportunity cost.

The concept of opportunity cost is related to scarcity of resources and their alternative uses. As pointed out earlier, resources available to any person, firm or society are limited. But resources have alternative uses with different productivity, i.e., income or returns from the alternative uses of resources are different. While one kind of use yields a higher income and return, the alternative uses yield a lower return or income. Due to their income maximizing behaviour, people (individuals, households and firms) put their scarce resources to the use that yields the highest income. When they put their resources to the use yielding the highest

income, they sacrifice the income expected from the next, the second best use of the resources. In economics terminology, this sacrifice is called opportunity cost of earning from the best use of resources. Such as it is, opportunity cost may be defined as income expected from the second best use of the resource which is sacrificed for the best use of the resource. Thus, *opportunity cost is opportunity lost*. From a firm's point view, the opportunity cost of using a resource is what the firm must give up to use the resource as it is used.

Opportunity cost is also called alternative cost because it arises due to the possibility of alternative uses of a resource. If a resource has only one use, i.e., it has no alternative use, there would not be any alternative or opportunity cost.

4. Present value of money and discounting principle

Money has a time value. The concept of time value of money is well represented by the proverb, 'a bird in hand is preferable to two in the bush'. That money has a time value can be understood by the fact that a rupee received today is preferable to a rupee receivable tomorrow. In general, present money is preferable to future money for the following reasons:

- It gives convenience of spending as the need arises.
- It gives liquidity and confidence.
- Given time, money can earn more money: it can be invested to earn a return or interest.

The concept of the time value of money is very often applied to investment decisions. Generally, there is a time lag between investment and return. When an investment is made today, it begins to yield returns at some future date. The time gap between investment and return is called time lag. During the time lag, the investor loses interest on the return receivable in future. It means that a rupee expected in future, say, a year hence, is worth less than a rupee today. This implies that present money has a future value, if invested, and future money has a *present value*. The concept of future value and present value of money can be clarified by using the method of knowing the present and future values.

Future value of present money

The future value of a sum of money invested is simply the *money invested plus interest*. In simple words, the future value, called *amount, equals principle plus* interest, i.e., *amount = principle + interest*. Given this definition, the future value can be calculated by the method of calculating the amount. For example, a sum of ₹ 100 is deposited with a bank of a period one year at the interest rate of 10 per cent p.a. The formula for finding the future value, i.e., the amount, is given as follows:

$$\text{Amount} = X + X(r)$$

where X is the principal (₹ 100) and r is the rate of interest (10%)

By using this formula, the 'amount' (the future value of ₹ 100), can be found as follows.

$$\begin{aligned}\text{Future value} &= 100 + 100\ (0.10)\\ &= 100 + 10\\ &= 110\end{aligned}$$

NOTES

Thus, given the rate of interest, the future value of a sum of money can be easily found.

Discounting principle

The present value (*PV*) of future income, an income expected at some future date, can be obtained by using the following discounting formula:

$$PV = \frac{X}{(1+r)}$$

where X is the future income and r is the market rate interest

Given the formula, the *PV* of ₹ 110 at 10 per cent rate of interest can be obtained as follows:

$$PV = \frac{X}{(1+r)} = \frac{110}{(1+0.10)} = ₹\ 100$$

Thus, the present value of ₹ 110 expected one year hence is ₹ 100, at the interest rate of 10 per cent. The discounting formula for working out *PV* of income in any future year (n) is given below.

$$PV = \frac{X}{(1+r)^n} \quad (n \text{ being the } nth \text{ year})$$

5. Externalities

Externalities refer to the costs that people have to bear without consuming goods and services produced by the firms and benefits which people reap without paying for them. The concept of externalities is related to the exclusion principle. The exclusion principle requires that:

(i) Those who do not pay for goods and services are excluded from the benefit or consumption of the goods and services.

(ii) Those who do not benefit from the production and sale of goods and services are not required to bear any cost.

In reality, however, the exclusion principle does not apply in the true sense of the term. Application of the exclusion principle requires that those who do not pay for a good should be excluded from its consumption and those who derive benefit from a good should bear its cost. In a modern complex society, there are numerous productive activities which harm those who do not benefit from them and benefit those who do not pay for such items of production. For instance, smoke-emitting factories, automobiles plying in the cities and use of loudspeakers on marriage ceremonies harm people by causing atmosphere and noise pollution. Even the people who do not benefit from these pollutants bear the cost in terms of loss of welfare. Such costs are known as 'spill-over costs'. Similarly, planting trees on road-sides, creation of parks and gardens, spread of education, creation of hygienic surroundings, etc. benefit even those who do not pay for them. Such benefits are known as 'spill-over benefits'. The spill-over costs and benefits are jointly called externalities. The market mechanism does not compensate or charge those who are affected by externalities.

Because of these shortcomings, the free market mechanism has failed in achieving optimum distribution of goods and services, optimum allocation of

resources, maximum efficiency and maximum social welfare. It has instead caused the growth of monopolies, inequitable income distribution, unemployment and poverty. Besides, though the free enterprise system is capable of bringing about economic growth, it does not ensure a stable, sustained and balanced growth. It, therefore, becomes necessary for the government to intervene in the market mechanism and apply necessary controls and measures to regulate the market mechanism.

6. Trade-off

Conceptually, trade-off refers to a situation in which two opposite variables moving in the same or opposite direction are in balance. The most important example of trade-off used in business analysis is the trade-off between the total revenue (*TR*) and the total cost (*TC*). *TR* and *TC* are defined, as follows:

$$TR = P \cdot Q$$

and $TC = AC \cdot Q$

where P = price and AC = average cost.

If price (*P*) remains constant, *TR* increases at a constant rate with increase in the sale of the output (*Q*). The behaviour of *TR* with increasing sales is shown by the *TR* line in Figure 1.2. However, as theory of cost reveals, *AC* first decreases and then increases as output (*Q*) increases. Therefore, *TC* takes the form of a curve, as shown by *TC* curve in Figure 1.2.

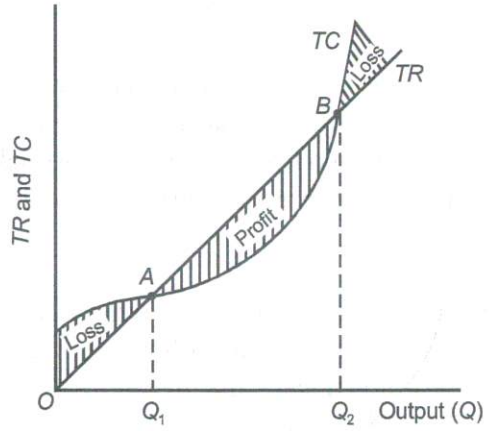


Fig. 1.2 *Trade-offs between TR and TC*

As Figure 1.1 shows, *TR* line and *TC* curve intersect at two points, *A* and *B*. These points mark the *trade-off points*, also called as *break-even points*, i.e., the points where *TR* and *TC* break even. Note that prior to point *A, TC* is higher than *TR*. Therefore, the producer makes losses. As output and sales increase, the gap between *TR* and *TC* decreases and losses disappear. Thus, point *A* shows the output (Q_1) at which there is trade-off between the *TR* and *TC*. Beyond output Q_1, the producer makes profit. Profit continues to increase until output increases to Q_2. At output Q_2, there is again a trade-off between the *TR* and *TC* as shown by point *B*. At point *B, TR* and *TC* break even again. Thus, the trade-off points *A* and *B* mark the range of profit-yielding production and sales.

7. Risk, uncertainty and profit

The concept of risk and uncertainty can be better explained and understood in contrast to the concept of certainty. Here, let us have a closer look at the concept of certainty and then proceed to explain the concepts of risk and uncertainty. *Certainty* is the state of perfect knowledge about the market conditions. In the state of certainty, there is only one rate of return on the investment and that rate is known to the investors. That is, in the state of certainty, the investors are fully aware of the outcome of their investment decisions. For example, if you deposit your savings in 'fixed deposit' bearing 10 per cent interest, you know for certain that the return on your investment in time deposit is 10 per cent, and FDR can be converted into cash any day at a lower rate of interest, of course. Or, if you buy government bonds, treasury bills, etc., bearing an interest of 11 per cent, you know for sure that the return on your investment is 11 per cent per annum, your principal remaining safe. In either case, your are sure that there is little or no possibility of the bank or the government defaulting on interest payment or on refunding the money. This is called the state of certainty.

- **Risk**

In common parlance, risk means a low probability of an expected outcome. From business decision-making point of view, risk refers to a situation in which a business decision is expected to yield more than one outcome and the probability of each outcome is known to the decision-makers or can be reliably estimated.

- **Uncertainty**

Uncertainty refers to a situation in which there is more than one outcome of a business decision and the probability of no outcome is known or can be meaningfully estimated. The unpredictability of outcome may be due to lack of reliable market information, inadequate past experience and high volatility of the market conditions.

- **Profit**

Profit means different things to different people. 'The word "profit" has different meaning to businessmen, accountants, tax collectors, workers and economists and it is often used in a loose polemical sense that buries its real significance.' In a general sense, 'profit' is regarded as income accruing to the equity holders, in the same sense as wages accrue to the labour; rent accrues to the owners of rentable assets and interest accrues to the money lenders. To a layman, profit means all income that flow to the investors and to an accountant, profit means the excess of revenue over all paid-out costs including both manufacturing and overhead expenses. It is more or less the same as 'net profit'. For all practical purposes, profit or business income means profit in accountancy sense plus non-allowable expenses. Economist's concept of profit is of 'pure profit' called 'economic profit' or 'just profit'. Pure profit is a return over and above the opportunity cost, i.e., the income which a businessman might expect from the second best alternative use of his resources.

Accounting profit vs economic profit

The two important concepts of profit that figure in business decisions are 'economic profit' and 'accounting profit'. It will be useful to understand the

difference between the two concepts of profit. As already mentioned, in accounting sense, profit is surplus of revenue over and above all paid-out costs, including both manufacturing and overhead expenses. Accounting profit may be calculated as follows:

$$\text{Accounting profit} = TR - (W + R + I + M)$$

where W = wages and salaries

R = rent

I = interest

and M = cost of materials

Obviously, while calculating accounting profit, only explicit or book costs, i.e., the cost recorded in the books of accounts, are considered.

The concept of 'economic profit' differs from that of 'accounting profit'. Economic profit takes into account also the implicit or imputed costs. In the context of profit, *implicit cost* is essentially the *opportunity cost*. Opportunity cost is defined as the payment that would be 'necessary to draw forth the factors of production from their most remunerative alternative employment.' Alternatively, opportunity cost is the income foregone which a businessman could expect from the second best alternative use of his resources. For example, if an entrepreneur uses his capital in his own business, he foregoes interest which he might earn by purchasing debentures of other companies or by depositing his money with joint stock companies for a period. Furthermore, if an entrepreneur uses his labour in his own business, he foregoes his income (salary) which he might earn by working as a manager in another firm. Similarly, by using productive assests (land and building) in his own business, he sacrifices his market rent. These foregone incomes—interest, salary and rent—are called *opportunity costs* or *transfer costs*. Accounting profit does not take into account the opportunity cost.

It should also be noted that the *economic* or *pure profit* makes provision also for (*a*) insurable risks, (*b*) depreciation, and (*c*) necessary minimum payment to shareholders to prevent them from withdrawing their capital. Pure profit may thus be defined as 'a residual left after all contractual costs have been met, including the transfer costs of management, insurable risks, depreciation and payments to shareholders sufficient to maintain investment at its current level'. Thus,

$$\text{Pure profit} = \text{Total revenue} - (\text{explicit costs} + \text{implicit costs})$$

Pure profit so defined may not be necessarily positive for a single firm in a single year—it may be even negative, since it may not be possible to decide beforehand the best way of using the resources. Besides, in economics, pure profit is considered to be a short-term phenomenon—it does not exist in the long run, especially under perfectly competitive conditions.

8. Marginalism: The marginal analysis

One of the most important contributions that economics has made to business analysis and managerial decisions is the application of marginal analysis. Marginal analysis uses marginal change in the dependent variable resulting from a unit change in its determinant, the independent variable, given the functional relationship between the dependent and independent variable.

For further clarification, look at the following cost function, as an example:

$$TC = f(Q)$$

It reads, total cost of production (TC) is the function of (or depends on) the quantity produced (Q). Here, TC is the dependent variable and Q is the independent variable. The relevant marginality concept here is marginal cost (MC). Marginal cost can be defined as the change in TC due to one unit change in Q. There are two ways of expressing the marginal value: in our case here, it is marginal cost (MC). When a series of Q and TC data is given, MC can be expressed and calculated as:

$$MC = TC_n - TC_{n-1}$$

For example, suppose a cost-output table shows that TC of 100 units of output (Q) of a commodity is ₹ 2500 and TC of output 101 units is ₹ 2550. That is,

$$\begin{aligned} MC &= TC_n - TC_{n-1} \\ &= TC_{101} - TC_{101-1} \\ &= ₹\,2550 - ₹\,2500 \\ &= ₹\,50 \end{aligned}$$

And, when a cost function is given, MC is defined and measured in terms of partial derivative, as shown below:

$$MC = dTC/dQ$$

The concepts of marginal value and marginal analysis are used widely in analysing many behavioural issues, e.g., marginal utility (MU) in consumer behaviour; marginal cost (MC) in cost analysis and in decisions for using an additional input (labour or capital), and marginal revenue (MR) in revenue maximization issue, and marginal profit (MP) in assessing the effect of a business decision on a firm's profit.

The application of marginal principle will be shown ahead when we deal with the equilibrium of consumer and firm, price determination and determination of factor prices.

9. Incrementalism: The incremental analysis

Similar to the concept of 'marginal' value is the concept of 'incremental' value. As mentioned earlier, marginal principle is applied only where MC and MR are known or can be calculated precisely. In general, however, firms do not have the knowledge of MC and MR. The reason is that most business firms produce and sell their products in bulk, not unit by unit unless, of course, it is the case of producing and selling such large-unit goods as aeroplanes, ships, large buildings, turbines, and so on. Where production and sale activities are carried out on bulk basis, and where both *fixed* and *variable* costs are subject to change, business managers use the *incremental principle* or *incremental analysis* in their business decisions.

The incremental principle is applied to business decisions which involve a large increase in total cost and total revenue. Such increase in total cost and total revenue is called incremental cost and incremental revenue, respectively. Let us

first explain the concept of incremental cost. Incremental costs can be defined as costs that arise due to a business decision. For example, suppose a firm decides to increase production by adding a new plant to the existing capacity or by setting up a new production unit. This decision increases the firm's total cost of production from ₹ 100 million to ₹ 115 million. Then ₹ 115 million – ₹ 100 million = ₹ 15 million is the incremental cost. Thus, an increase in the total cost of production due to a business decision is incremental cost. Incremental cost includes both fixed and variable costs. However, it does not include the costs already incurred on the excess capacity or what is called the sunk cost or the cost of unused material.

There are three major components of incremental cost (*IC*): (*i*) present explicit costs, (*ii*) opportunity cost, and (*iii*) future costs. *Present explicit costs* include: (*a*) *fixed cost*, i.e., the cost of plant and building, (*b*) *variable costs* including cost of direct labour and materials and overheads like electricity and indirect labour. *Opportunity cost* as you know refers to expected income foregone from the second best use of the resources involved in the present decision. *Future costs* of a business decision include depreciation and advertising costs if the product does not sell as well as expected.

The *incremental revenue*, on the other hand, is the increase in revenue due to a business decision. A business decision is taken presumably in expectation of an increase in the firm's revenue. When a business decision is successfully implemented, it does result in a significant increase in its total revenue. The increase in the total revenue resulting from a business decision is called *incremental revenue*. Suppose that after the installation of a new plant, the total production increases and the firm is able to sell the incremental product. As a result, the firm's total sales revenue increases, let us suppose, from ₹ 130 million to ₹ 148 million. Thus, the post-decision total revenue of ₹ 148 million *minus* the pre-decision total revenue of ₹ 130 million = ₹ 18 million is the incremental revenue.

Incremental reasoning in business decision

The use of the incremental concept in business decisions is called *incremental reasoning*. Incremental reasoning is used in accepting or rejecting a business proposition or option. For instance, suppose that in our foregoing example, the firm considers whether or not to install a new plant. As noted above, the firm estimates an incremental cost of installing a new plant at ₹ 15 million and an incremental revenue of ₹ 18 million. The incremental revenue exceeds the incremental cost by ₹ 3 million which means a 20 per cent return (gross of overheads) on the investment in the new plant. The firm will accept the proposition of installing a new plant, provided there is no better business proposition available to the firm.

However, if the firm finds that it will be required to spend an additional amount of ₹ 1.5 million on advertising to sell the output from the new plant, it may reject the proposition to install a new plant, for in that case its rate of return would fall to 10 per cent (gross of overheads). The firm may find it advisable to put its money in mutual funds at 12.5 per cent annual return (net of overheads).

It may be added at the end, by way of comparison, that the *marginal concept* (especially when defined and measured by calculus) is used in economic

analysis where a high degree of precision is involved, whereas the incremental concept is used where large values of cost and revenue especially are involved. Besides, incremental concept and reasoning are used in business decisions more frequently than the marginality concept. There are at least two reasons for this. First, marginality concept used in business analysis is generally associated with *one* (marginal) unit of output produced or sold, whereas most business decisions involve large quantities and values. Second, the precise calculation of marginal change (defined in terms of the first derivative of a function) is neither practicable nor necessary in real life business considerations.

10. The equi-marginal principle

The equi-marginal principle was originally associated with the consumption theory and the law is called 'the law of equi-marginal utility'. The law of equi-marginal utility states that a utility maximizing consumer distributes his consumption expenditure between various goods and services he/she consumes in such a way that the marginal utility derived from each unit of expenditure on various goods and service is the same. This pattern of consumption expenditure maximizes a consumer's total utility.

The law of equi-marginal principle was over time applied to the allocation of resources between their alternative uses with a view to maximizing profit in case a firm carries out more than one business activity. This principle suggests that available resources (inputs) should be so allocated between the alternative options that the marginal productivity gains (*MP*) from the various activities are equalized. For example, suppose a firm has a total capital of ₹ 100 million which it has the option of spending on three projects, A, B and C. Each of these projects requires a unit expenditure of ₹ 10 million. Suppose also that the marginal productivity schedule of each unit of expenditure on the three projects is given as shown in Table 1.1.

Table 1.1 *Marginal Productivity (MP) Schedule of Projects A, B and C*

Units of Expenditure	*Marginal Productivity (MP)* (₹ *in million*)		
(₹ *10 million*)	*Project A*	*Project B*	*Project C*
1st	50_1	40_3	35_4
2nd	45_2	30_5	30_6
3rd	35_7	20_8	20_9
4th	20_{10}	10	15
5th	10	0	12

Note: Subscripts 1, 2, 3... indicate the order of the unit of expenditure on Projects *A*, *B* and *C*.

Going by the equi-marginal principle, the firm will allocate its total resources (₹ 100 million) among the projects A, B and C in such a way that marginal product of each project is the same, i.e., $MP_A = MP_B = MP_C$. It can be seen from Table 1.1 that, going by this rule, the firm will spend 1st, 2nd, 7th and 10th unit of finance on project A; 3rd, 5th and 8th unit on Project B; and 4th, 6th and 9th unit on Project C. In all, it puts 4 units of its finances in Project A, 3 units each in Projects B and C. In other words, of the total finances of ₹ 100 million, a profit maximizing firm would invest ₹ 40 million in Project A, ₹ 30 million each in Projects B and C.

This pattern of investment maximizes the firm's productivity gains. No other pattern of investment will ensure this objective.

Now, the equi-marginal principle can be formally stated. The equi-marginal principle suggests that a profit (gain) maximizing firms allocates its resources in a proportion such that

$$MP_A = MP_B = MP_C = \ldots = MP_N$$

If cost of project (*COP*) varies from project to project, then resources are so allocated that *MP* per unit of *COP* is the same. That is, resources are allocated in such proportions that

$$\frac{MP_A}{COP_A} = \frac{MP_B}{COP_B} = \frac{MP_C}{COP_C} = \ldots = \frac{MP_N}{COP_N}$$

The equi-marginal principle can be applied only where (*i*) firms have limited investible resources, (*ii*) resources have alternative uses, and (*iii*) the investment in various alternative uses is subject to diminishing marginal productivity or returns.

Check Your Progress

1. Write any two methods involved in capital budgeting.
2. Which relationship does the theory of firm explain?
3. Define time lag.

1.3 DEMAND ANALYSIS: LAW OF DEMAND

The term 'demand' refers to the *quantity demanded* of a commodity *per unit of time* at a given price. It also implies a *desire* backed by ability and willingness to pay. Mere desire of a person to purchase a commodity is not his demand. He must possess adequate resources and must be willing to spend his resources to buy the commodity. Besides, the *quantity demanded* always has a reference to 'a price' and 'a unity of time'. The quantity demanded referred to as 'per unit of time' makes it a *flow* concept. Apparently there may be some problems in applying this flow concept to the demand for durable consumer goods like a house, car, refrigerator, etc. But this apparent difficulty may be resolved by considering the fact that the total service of a durable good is not consumed at one point of time and its utility is not exhausted in a single use. The service of a durable good is consumed over time. At a time, only a part of its service is consumed. Therefore, the demand for the services of durable consumer goods may also be visualized as a demand per unit of time. However, this problem does not arise when the concept of demand is applied to total demand for a consumer durable. Thus, the demand for consumer goods also is a flow concept.

Elements of demand

- Quantity demanded per unit of time
- Demand of a given price
- One who demands has ability to pay
- One who has ability to pay has willingness to pay for a commodity

NOTES

Demand analysis as a subject includes a variety of topics. We'll restrict ourselves to only the fundamentals in this unit, beginning with law of demand in this section, followed by elasticity and forecasting in the subsequent sections.

1.3.1 Law of Demand

The quantity of a commodity that an individual or a household consumes per unit of time is determined by a number of factors including price of the commodity, price of its substitutes and complements, consumer's income, his/her wealth holding, taste and preference, expectations about future income and price and demonstration effect.

Let it be emphasized that the price of a commodity is the most important determinant of its demand. The relationship between price and demand is expressed by the *law of demand.* The law of demand states that *quantity of a product demanded per unit of time increases when its price falls, and decreases when its price increases, other factors remaining constant.* The assumption 'other factors remaining constant' implies that income of the consumers, prices of the substitutes and complementary goods, consumers' taste and preference, and number of consumers, remain unchanged.

The law of demand can be illustrated through a demand schedule. A demand schedule is a series of quantities which consumers would like to buy per unit of time at different prices. To illustrate the law of demand, an imaginary demand schedule for tea is given in Table 1.2. It shows seven alternative prices and the corresponding quantities (number of cups of tea) demand per day. Each price has a unique quantity demanded, associated with it. As price per cup of tea decreases, daily demand for tea increases, in accordance with the law of demand.

Demand curve

The law of demand can also be presented through a curve called *demand curve. A demand curve is a locus of points showing various alternative price-quantity combinations*. It shows the quantities of a commodity that consumers or users would buy at different prices per unit of time under the assumptions of the law of demand. An individual's demand curve for tea as given in Figure 1.2 can be obtained by plotting the data given in Table 1.2.

Table 1.2 *Demand Schedule for Tea*

Price per cup of tea (₹)	*No. of cups of tea demand per consumer per day*	*Symbols representing per price-quantity combination*
8	2	A
7	3	B
6	4	C
5	5	D
4	6	E
3	7	F
2	8	G

In Figure 1.2, the curve from point *A* to point *G* passing through points *B*, *C*, *D*, *E* and *F* is the demand curve *DD'*. Each point on the demand curve *DD'* shows a unique price–quantity combination. The combinations read in alphabetical

order show decreasing price of tea and increasing number of cups of tea demanded per day. Price–quantity combinations in reverse order of letters show increasing price of tea per cup and decreasing number of cups of tea per day consumed by an individual. The whole demand curve shows a functional relationship between the alternative price of a commodity and its corresponding quantities which a consumer would like to buy during a specific period of item—per day, per week, per month, per season, or per year. *The demand curve shows an inverse relationship between price and quantity demanded.* This inverse relationship between price and quantity demanded makes the demand curve slope downward to the right.

Why does the demand curve slope downward to the right?

As Figure 1.3 shows, the demand curve slopes downward to the right. The downward slope of the demand curve shows the law of demand, i.e., the quantity of a commodity demanded per unit of time increases as its price falls, and vice versa.

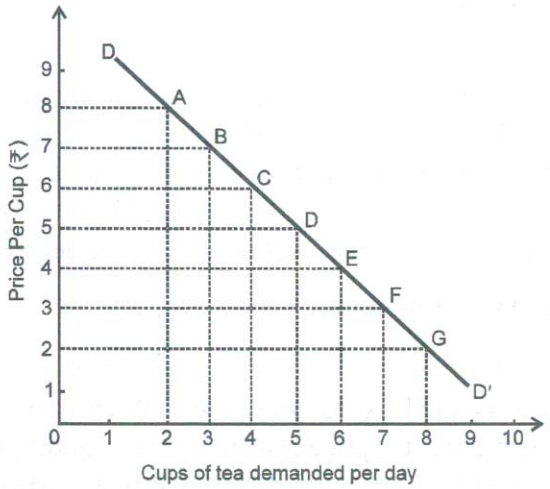


Fig. 1.3 *The Demand Curve*

The reasons behind the law of demand, i.e., inverse relationship between price and quantity demanded are the following:

(i) **Substitution effect:** When the price of a commodity falls it becomes relatively cheaper if the price of all other related goods, particularly of substitutes, remain constant or, in other words, substitute goods become relatively costlier. Since consumers substitute cheaper goods for costlier ones, demand for the relatively cheaper commodity increases. The increase in demand on account of this factor is known as *substitution effect.*

(ii) **Income effect:** As a result of fall in the price of a commodity, the real income of its consumer increases at least in terms of this commodity. In other words, his/her purchasing power increases since he is required to pay less for the same quantity. The increase in real income (or purchasing power) encourages demand for the commodity with reduced price. The increase in demand on account of an increase in real income is known as *income effect.*

It should however be noted that the *income effect* is negative in the case of inferior goods. In case price of an inferior good accounting for a considerable

proportion of the total consumption expenditure falls substantially, consumers' real income increases: they become relatively richer. Consequently, they substitute the superior good for the inferior ones, i.e., they reduce the consumption of inferior goods. Thus, the *income effect* on the demand for inferior goods becomes negative.

(iii) **Diminishing Marginal Utility:** *Diminishing marginal utility* is also responsible for increase in demand for a commodity when its price falls. When a person buys a commodity, he exchanges his money income with the commodity in order to maximize his satisfaction. He continues to buy goods and services so long as marginal utility of money (MU_m) is less then marginal utility of the commodity (MU_c). Given the price of a commodity, he adjusts his purchase so that $MU_c = MU_m$. This proposition holds good under both the Marshallian assumption of constant MU_m and the Hicksian assumption of diminishing MU_m. Under the Marshallian approach, MU_m remaining constant, $MU_c = P_c$ and a utility maximizing consumer reaches his equilibrium where

$$MU_m = P_c = MU_c$$

When price falls, $(MU_m = P_c) < MU_c$. Thus, the equilibrium condition is disturbed. To regain his equilibrium condition, i.e., $MU_m = P_c = MU_c$, he purchases more of the commodity. For, when the stock of a commodity increases, its MU decreases and once again $MU_m = MU_c$. That is why demand for a commodity increases when its price decreases.

Exceptions to the law of demand

The law of demand does not apply to the following cases:

(a) **Expectations regarding future price:** When consumers expect a continuous increase in the price of a durable commodity, they buy more of it despite increase in its price. They do so with a view to avoiding the pinch of still higher prices in future. Similarly, when consumers anticipate a considerable decrease in the price in future, they postpone their purchases and wait for the price to fall to the expected level rather than buy the commodity when its price initially falls. Such decisions of the consumers are contrary to the law of demand.

(b) **Status goods:** The law does not apply to the commodities which serve as a 'status symbol', enhance social prestige or display wealth and richness, e.g., gold, precious stones, rare paintings and antiques. Rich people buy such goods mainly because their prices are high.

(c) **Giffen goods:** An exception to this law is also the classic case of Giffen goods named after Robert Giffen (1837–1910). Giffen goods do not refer to any specific commodity. They may be any inferior commodities much cheaper than their superior substitutes, consumed mostly by poor households as essential consumer goods. If the price of such goods increases (price of their substitutes remaining constant), their demand increases instead of decreasing. For instance, let us suppose that the monthly minimum consumption of foodgrains by a poor household is 30 kgs including 20 kgs of bajra (an inferior good) at the rate of ₹10 per kg and 10 kgs of wheat (a superior good) at ₹20 per kg. It spends a fixed

amount of ₹400 on these items. Now, if the price of bajra increases to ₹12 per kg the household will be forced to reduce the consumption of wheat by 5 kgs and increase that of bajra by the same quantity in order to meet its minimum monthly consumption requirement within ₹400. Obviously, the household's demand for bajra increases from 20 to 25 kgs when its price increases.

1.3.2 Market Demand Curve

The quantity of a commodity which an individual is willing to buy at a particular price of the commodity during a specific time period, given his money income, his taste, and prices of substitutes and complements, is known as individual demand for a commodity. The total quantity which all the consumers of a commodity are willing to buy at a given price per time unit, other things remaining the same, is known as market demand for the commodity. In other words, the market demand for a commodity is the sum of individual demands by all the consumers (or buyers) of the commodity, per time unit, and at a given price, other factors remaining the same. For instance, suppose there are three consumers (viz., *A, B, C*) of a commodity *X*, and their individual demands at different prices is of *X* as given in Table 1.3. The last column presents the market demand, i.e., the aggregate of individual demand by three consumers at different prices. Graphically, the market demand curve is the *horizontal summation* of individual demand curves. The individual demand schedules plotted graphically and summed up horizontally gives the market demand curve as shown in Figure 1.4.

Table 1.3 *Price and Quantity Demanded*

Price of commodity X (per unit)	***Quantity of X demanded by***			***Market demand***
	A	***B***	***C***	
10	4	2	0	6
8	8	4	0	12
6	12	6	2	20
4	16	8	4	28
2	20	10	6	36
0	24	12	8	44

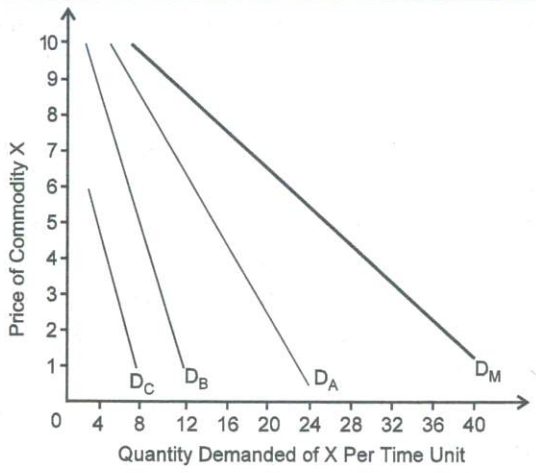


Fig. 1.4 *Derivation of Market Demand Curve*

The individual demands for commodity X are given by D_A, D_B and D_C, respectively. The horizontal summation of these individual demand curves results in the *market demand curve* (D_M) for the commodity X. The curve D_M represents *the market demand curve* for commodity X when there are only three consumers of the commodity.

Determinants of market demand

The market demand for a product is determined by a number of factors, viz. price of the product, price and availability of the substitutes, consumer's income, his own preference for a commodity, utility derived from the commodity, 'demonstration effect', advertisement, credit facility by the sellers and banks, off-season discounts, number of the uses of the commodity, population of the country, consumer's expectations regarding the future trend in the price of the product, consumers' wealth, past levels of demand, past levels of income, government policy, etc. But all these factors are not equally important. Besides, some of these factors are not quantifiable, e.g., consumer's preferences, utility, demonstration effect, and expectations, and hence are not usable in the demand estimation. Nevertheless, we will discuss here how some important quantifiable and non-quantifiable determinants determine the market demand for a product.

1. Price of the commodity

As stated above, price is the most important determinant of the quantity demanded of a commodity. The price–quantity relationship is the central theme of demand theory. The nature of relationship between price of a commodity and its quantity demanded has already been discussed under the 'Law of Demand'.

2. Price of substitutes and complementary goods

The demand for a commodity depends also on the prices of *its substitutes and complementary goods*. Two commodities are deemed to be *substitutes* for each other if change in the price of one affects the demand for the other in the same direction. For instance, commodities X and Y are, in an economic sense, *substitutes* for each other if a rise in the price of X increases the demand for Y, and vice versa. Tea and coffee, hamburger and hot dog, alcohol and drugs are some common examples of substitutes.

By definition, the relationship between demand of a product (say, tea) and the price of its substitute (say, coffee) is positive in nature. When price of the substitute (coffee) of a product (tea) falls (or increases), demand for the product falls (or increases). The relationship of this nature is given in Figure 1.5(a).

A commodity is deemed to be a *complement* of another when it complements the use of the other. In other words, when the use of any two goods goes together so that their demand changes (increases or decreases) simultaneously, they are treated as complements. For example, petrol is a complement of motor vehicles; butter and jam are complements of bread; milk and sugar are complements of tea and coffee. Technically, two goods are complements of one another if an increase in the price of one causes a decrease in the demand for another. By definition, there is an inverse relationship between the demand for a good and the price of its complement. For instance, an increase (or decrease) in the price of petrol causes a decrease (or an increase) in the

demand for car, other things remaining the same. The nature of relationship between the demand for a product and the price of its complement is given in Figure 1.5(b).

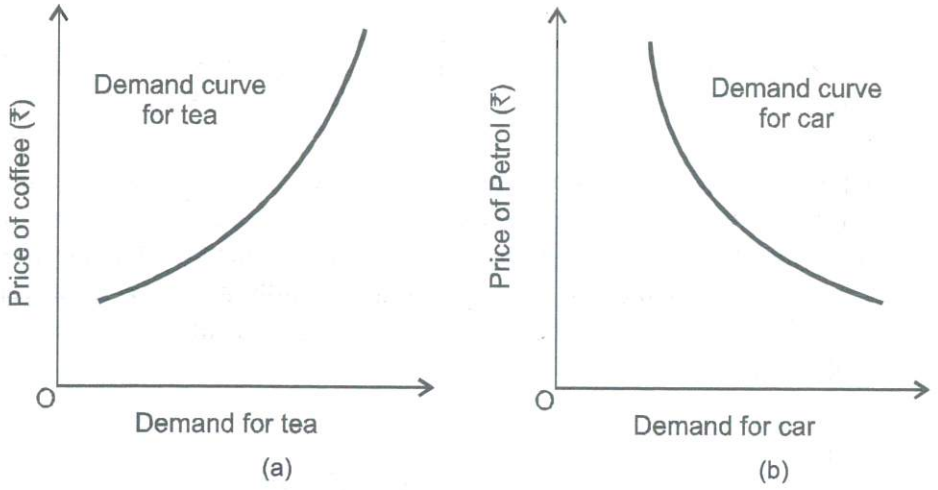


Fig. 1.5 *Demand for Substitutes and Complements*

3. Consumer's income

Income is the basic determinant of the quantity demanded of a product as it determines the purchasing power of the consumer. That is why the people with higher current disposable income spend a larger amount on normal goods and services than those with lower incomes. Income–demand relationship is of a more varied nature than that between demand and its other determinants.

For the purpose of income–demand analysis, goods and services may be grouped under four broad categories, viz. (a) essential consumer goods; (b) inferior goods; (c) normal goods; and (d) prestige or luxury goods. The relationship between income and the different kinds of consumer goods is presented through the Engel Curves in Figure 1.6.

(a) **Essential consumer goods (ECG):** The goods and services which fall in this category are consumed, as a matter of necessity, by almost all persons of a society, e.g., foodgrains, salt, vegetable oils, matches, cooking fuel, a minimum clothing and housing. Quantity demanded of such goods increases with increase in consumer's income only upto a certain limit, other factors remaining the same. The relation between demand of this category and consumer's income is shown by curve *ECG* in Figure 1.6. As the curve shows, consumer's demand for essential goods increases until his income rises to OY_2 and beyond this level of income, it does not.

(b) **Inferior goods:** Inferior and superior goods are generally known to the consumers by and large. For instance, every consumer knows that *bajra* is inferior to wheat and rice; *bidi* (an indigenous cigarette) is inferior to cigarette, cars without AC are inferior to AC cars, kerosene-stove is inferior to gas-stove; travelling by bus is inferior to travelling by taxi, and so on. In economic terminology, however, a commodity is deemed to be inferior if its demand decreases with the increase in consumers' income. The relation

between income and demand for an inferior good is shown by curve *IG* in Figure 1.6 under the assumption that other determinants of demand remain the same. Demand for such goods may initially increase with increase in income (say, upto Y_1) but it decreases when income increases beyond a certain level.

(c) **Normal goods:** Technically, normal goods are those which are demanded in increasing quantities as consumers' income rises. Clothing is the most important example of this category of goods. The nature of relation between income and demand for the normal goods is shown by curve *NG* in Figure 1.6. As the curve shows, demand for such goods increases with the increase in income of the consumer, but at different rates at different levels of income. Demand for normal goods initially increases rapidly, and later, at a lower rate. With the increase in the consumers' income, its income-elasticity decreases.

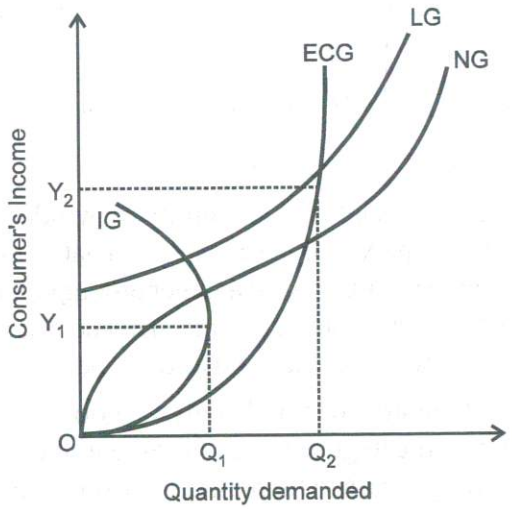


Fig. 1.6 Income–Demand Curves

It may be noted from Figure 1.5 that upto a certain level of income (Y_1) the relation between income and demand for all types of goods is positive. While demand for some NG_s increases at a faster rate, for others, it increase at a low rate. The difference is of degree only. The income–demand relationship becomes distinctly different beyond the level of income Y_1.

(d) **Prestige or luxury goods:** Prestige goods are those which are consumed mostly by the rich section of the society, e.g., luxury cars, stone-studded jewellery, costly cosmetics, decoration items (like antiques) etc. Demand for such goods arises only beyond a certain level of consumer's income. The income–demand relationship of this category of goods is shown by the curve *LG* in Figure 1.6.

4. Consumers' taste and preferences

Consumers' taste and preferences play an important role in determining the demand for a product. Taste and preferences depend, generally, on the social customs, religious values attached to a commodity, habits of the people, the general lifestyle of the society, and also the age and sex of the consumers. Changes in these factors change consumers' taste and preferences. As a result, consumers reduce or give

up the consumption of some goods and include some others in their consumption basket. Generally, if consumers' liking, taste and preference for certain goods and services change following the change in fashion, people switch their consumption pattern from cheaper, old-fashioned goods over to costlier 'mod' goods, so long as the price differentials commensurate with their preference. Consumers are prepared to pay higher prices for 'mod' goods even if their virtual utility is the same as that of old-fashioned goods. This fact reveals that tastes and preferences also influence the demand for goods and services.

5. Consumers' expectations

Consumers' expectations regarding the future course of economic events, particularly regarding changes in prices, income, and supply position of goods, play an important role in determining the demand for goods and service in the short-run. As mentioned above, if consumers expect a rise in the price of a commodity, they tend to buy more of it at its current price with a view to avoiding the pinch of price–rise in future. For example, when the automobile owners expect or Government of India announces a rise in petrol and diesel prices from a future date, automobile owners buy more of petrol and diesel at their current prices. On the contrary, if consumers expect a fall in the price of certain goods, they postpone their purchase of such goods with a view to taking advantage of lower prices in future, mainly in the case of non-essential goods. This behaviour of consumers reduces the current demand for the goods whose prices are expected to decrease in future.

Similarly, an expected increase in income on account of announcement of revision of pay scales, dearness allowance, bonus, etc., induces increase in current purchase, and vice versa. Besides, if consumers or users expect scarcity of certain goods in future on account of a reported fall in future production, labour strikes on a large scale, diversion of civil supplies towards the military use, etc., the current demand for such goods would increase, more so if their prices show an upward trend. Consumers demand more for future consumption; profiteers demand more to make money out of expected scarcity. In simple words, expectation regarding the shortage of a commodity in future increases its current demand at the prevailing price.

6. Demonstration Effect

When new commodities or new models of existing ones appear in the market, rich people buy them first. Some people buy new goods or new models of goods because they have a genuine need for them while others buy because they want to exhibit their affluence. Fashion goods make the most common case for this category of goods. But once new commodities come in vogue, many households buy them not because they have a genuine need for them but because others or neighbours have bought these goods. The purchases by the latter category of the buyers are made out of such feelings as jealousy, competition, equality in the peer group, social inferiority and the desire to raise their social status. Purchases made on account of these factors are the result of 'demonstration effect' or the 'Band-wagon effect'. These effects have a positive effect on the demand. On the contrary, when a commodity becomes a thing of common use, some people, mostly the rich, decrease or give up the consumption of such goods. This is known as 'snob

effect'. It has a negative effect on the demand for the related goods.

7. Consumer-credit facility

Availability of credit to the consumers from the sellers, banks, relations and friends or from any other source encourages the consumers to buy more than what they would buy in the absence of credit facility. That is why the consumers who can borrow more can consume more than those who can borrow less. Credit facility affects mostly the demand for consumer durables, particularly those which require bulk payment at the time of purchase.

8. Population of the country

The total domestic demand for a product depends also on the size of the population. Given the price, per capita income, taste, preferences, etc., the larger the population, the larger the demand for a product. With an increase (or decrease) in the size of population, employment percentage remaining the same, demand for the product will increase (or decrease). The relation between market demand for essential and normal goods and the size of population is similar to the income–demand relation.

9. Distribution of national income

Apart from the level of individual incomes, the distribution pattern of national income also affects the demand for a commodity. If national income is evenly distributed, market demand for normal goods will be the largest. If national income is unevenly distributed, i.e., if majority of population belongs to the lower income groups, market demand for essential goods (including inferior ones) will be the largest whereas the same for other kinds of goods will be relatively low. Furthermore, given a distribution of national income and a market demand for various types of goods, if national income gets distributed in favour of the rich so that this section becomes smaller, the demand for essential goods will increase and the same for other kinds of goods will decrease and vice versa.

1.3.3 Demand Function

In mathematical language, a function is a symbolic statement of relationship between a dependent and independent variables. Demand function states the relationship between demand for a product (the dependent variable) and its determinants (the independent variables).

A demand function may include a single independent variable (price) or many other independent variables, depending on the purpose of analysis. A demand function based on a single independent variable is called *simple demand function* and one including many independent variables is called *dynamic* or *multivariate* demand function. Let us first discuss the simple demand function.

(i) Simple demand function

Let us assume that the quantity demanded of a commodity X depends only on its price (P_x), other factors remaining constant. The demand function will then read as 'demand for a commodity (X) depends on its price (P_x)'. The same statement may be symbolically written as

$$D_x = f(P_x) \quad ...(1.1)$$

In Equation (1.1), D_x is a dependent and P_x is an independent variable. The demand function Equation (1.1) reads 'demand for commodity X (D_x) is the function of its price (P_x,). It implies that a change in P_x (the independent variable) will cause a change in D_x (the dependent variable). Equation (1.1) however neither specifies the change in D_x for a given change in P_x, nor does it reveal the nature of relationship between D_x and P_x. When quantitative relationship between D_x and P_x is known, the demand function is expressed in the form of an equation, as

$$D_x = a - bP_x \quad ...(1.2)$$

where a denotes the quantity demanded at zero price (P_x) and $-b$ gives the change in demand (D_x) for change in P_x by 1.

The form of equation depends on the nature of demand–price relationship. The two most common forms of demand–price relationship are *linear* and *non-linear*. Accordingly, the demand function may assume a *linear* or a *non-linear* form.

1. Linear demand function

A demand function is said to be linear when the slope of the demand curve (i.e., D*D*/D*P*) remains constant throughout its length. The simplest form of a linear demand function is given by Equation (1.2).

Given the demand function of Equation (1.2), if values of parameters a and b are known, the total demand (D_x) for any given price (P_x) can easily be obtained and if a series of alternative prices is given, a demand schedule can be easily prepared. Let us assume that $a = 100$ and $b = 5$. Now the demand function of Equation (1.2) can be written as

$$D_x = 100 - 5P_x \quad ...(1.3)$$

Given the Equation (1.3), the value of D_x can be easily obtained for any value of P_x. For example, if $P_x = 4$,

$$D_x = 100 - 5 \times 4 = 80$$

If $P_x = 10$

$$D_x = 100 - 5 \times 10 = 50$$

and if $P_x = 15$

$$D_x = 100 - 5 \times 15 = 25$$

Thus, a demand schedule can be prepared assigning different values for P. This demand schedule plotted on a graph paper gives a linear demand curve, as shown in Figure 1.7.

From the demand function, one can easily obtain the price function. For example, given the demand function Equation (1.2), the price function may be written as

$$P_x = \frac{a - D_x}{b} \quad ...(1.4)$$

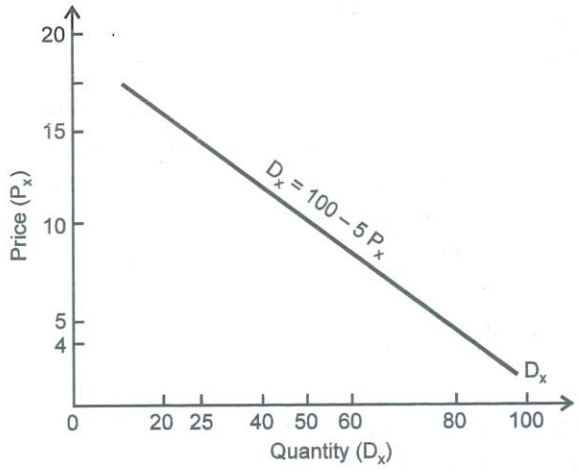


Fig. 1.7 Linear Demand Function

2. Non-linear demand function

A demand function is said to be non-linear or curvilinear when the slope of a demand curve, (D*D*/D*P*) changes all along the demand curve. Non-linear demand function yields a demand curve instead of a *demand line*, as shown in Figure 1.8. A non-linear demand function, generally, takes the form of a power function as

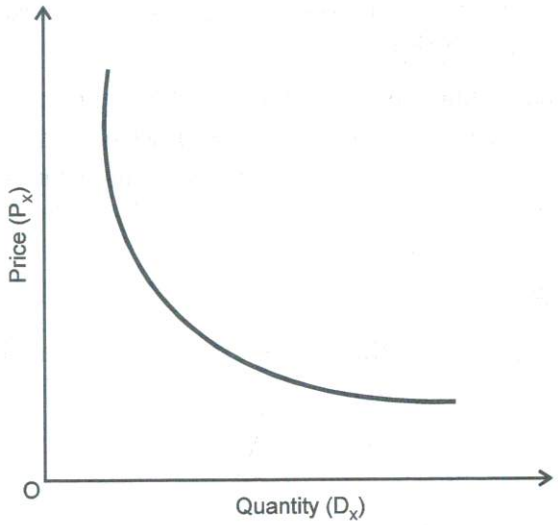


Fig. 1.8 Non-linear Demand Function

$$D_x = aP_x^{-b} \qquad ...(1.5)$$

or of a rectangular hyperbola of the form

$$D_x = \frac{a}{P_x + c} - b \qquad ...(1.6)$$

where $a, b, c > 0$.

It is noteworthy that the exponent $(-b)$ of the price variable in a non-linear demand function of Equation (1.5) is the coefficient of price elasticity of demand which is constant.

(ii) Dynamic or multivariate demand function

We have discussed above demand function with price as a single independent variable. This kind of demand function may be termed as *short-term demand function*. In the long-run, the market demand for a product depends on the composite impact of all the determinants operating simultaneously. Therefore, in a *long-run* or *dynamic demand function*, all the relevant determinants of demand for a product are included in the demand function. For instance, if individual demand (D_x) for a commodity X, depends on its price (P_x), consumer's income Y, price of its substitute (P_s), price of complementary goods (P_c), consumer's taste (T) and firms' advertisement expenditure (A), demand function can be expressed in a functional form as

$$D_x = f(P_x, Y, P_s, P_c, T, A) \quad ...(1.7)$$

The demand function of Equation (1.7) describes the relationship between demand for the commodity X, and its determinants (viz., P_x, Y, P_s, P_c, T and A). If relationship between D_x and the independent variables P_x, Y, P_s, P_c and A is of linear form, the estimable form of the demand function is expressed as

$$D_x = a + bP_x + cY + dP_s + jA \quad ...(1.8)$$

where a is a constant term and b, c, d and j are the coefficients of relation between D_x and the respective independent variables. In a *market demand function* for a product, other independent variables, viz. size of population (N) and a measure of income distribution, i.e., Gini-coefficient (G), may also be included. The dynamic demand function is useful for both theoretical analysis and empirical estimation of demand for a product.

Movement on and shift in demand curve

While using dynamic or long-term demand function, a distinction is often made between

(i) extension or contraction in demand causing movement along the demand curve, and

(ii) shift in demand. The two phrases are explained here briefly.

Movement along the demand curve: By the law of demand, when price of a commodity changes all other determinants of demand remaining constant, demand for the commodity changes. As a result, the consumer moves from one point to another on the *same demand curve*. This is called *movement on or along the demand curve*. For example, suppose the demand curve for a commodity, say X, is given as D_1 in Figure 1.8 and price of the commodity is given at OP_2. Given these conditions, the consumer is placed at point A, demanding OQ_1 units of the commodity. Now, let the price of commodity X fall from OP_2 to OP_1. With the fall in price, the utility maximizing consumer moves from point A to point B. His/her demand for the commodity X increases from OQ_1 to OQ_2. This kind of consumer's movement along the demand curve is also called as *extension in demand.*

On the contrary, if the price of the commodity X increases for some reason from OP_1 to OP_2, making the consumer move back to point A, this upward movement is called *contraction in demand.* Note that change in price, other

factors remaining constant, causes change in demand and change in demand causes movement along the same demand curve.

Shift in demand curve: When a demand curve changes its position, its shape remaining the same (though not necessarily), the change is known as *shift in demand curve*. Let us suppose that the demand curve, D_2 in Figure 1.8 is the original demand curve for commodity X. As shown in the figure, at price OP_2 the consumer buys OQ_2 units of X, other factors remaining constant. Now if any of the other factors (e.g., consumer's income) change, it will change the consumer's ability and willingness to buy commodity X. For example, if the consumer's disposable income decreases, say, due to an increase in income tax, he may be able to buy only OQ_1 units of X instead of OQ_2 at price OP_2. This is true for the whole range of the price of X—the consumers would be able to buy less of commodity X at all other prices. This will cause a *downward shift* in the demand curve from D_2 to D_1. Similarly, increase in disposable income of the consumer due to reduction in taxes may cause an *upward shift* from D_2 to D_2. Such changes in the position of the demand curve are known as *shifts in demand curve.*

Reasons for shift in demand curve

Shifts in a price–demand curve may take place owing to the change in one or more of other determinants of demand. Consider, for example, the decrease in demand for commodity X by Q_1Q_2 in Figure 1.9. Given the price OP_1, the demand for X might have fallen from OQ_2 to OQ_1 (i.e., by Q_1Q_2) for any of the following reasons:

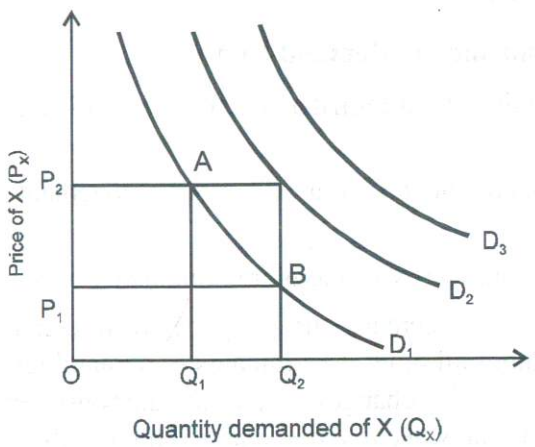


Fig. 1.9 *Shift in Demand Curves*

(i) fall in the consumer's income so that he can buy only OQ_1 of X at price OP_2—this is the income effect;

(ii) price of X's substitute falls so that the consumers find it beneficial to substitute Q_1Q_2 of X with its substitute—this is the substitution effect;

(iii) advertisement made by the producer of the substitute, changes consumer's taste or preference against commodity X so much that they replace Q_1Q_2 of X with its substitute, again a substitution effect;

(iv) price of complement of X increases so much that they can now afford only OQ_1 of X; and

(v) also for such reasons as commodity X is going out of fashion; its quality has deteriorated; consumer's technology has so changed that only OQ_1 of X can be used, and due to change in season if commodity X has only seasonal use.

1.3.4 Types of Demand

The demand for various goods is generally classified on the basis of the consumers of a product, suppliers of the product, nature of goods, duration of consumption of a commodity, interdependence of demand, period of demand and nature of use of the goods (intermediate or final). Let us discuss the major types of demand that figure in business decisions.

(a) Individual and market demand

As mentioned earlier, the quantity of a commodity which an individual is willing to buy at a particular price during a specific time period, given his income, his taste and prices of other commodities (particularly substitutes and complements), is called 'individual's demands for a commodity'.

The total quantity which all the consumers of a commodity are willing to buy at a given price per time unit, given their money income, taste and prices of other commodities (mainly substitutes) is known as 'market demand for the commodity'. In other words, the market demand for a commodity is the sum of individual demands by all the consumers (or buyers) of the commodity, over a time period and at a given price, other factors remaining the same. For example, the quantity of detergent purchased by an individual household, in a month, is termed as individual demand. Unlike market demand implies the sum total of all individual demand for the commodity at each possible price, over a period of time.

There are two basic types of market demand: primary and selective. Primary demand is the total demand for all of the brands that represent a given product or service, such as all phones or all high-end watches. Selective demand is the demand for one particular brand of product or service, such as the iPhone or a Rolex watch.

(b) Demand for firm's product and industry's products

The quantity of a firm's product that can be disposed of at a given price over a time period connotes the demand for the firm's product. The aggregate of demand for the product of all the firms of an industry is known as the market demand or demand for industry's product. This distinction between the two kinds of demand is not of much use in a highly competitive market—since it merely signifies the distinction between a sum and its parts.

However, where market structure is oligopolistic, a distinction between the demand for a firm's product and for the industry's product is useful from the managerial point of view. For, in such markets, products of each firm are so differentiated from the products of the rival firms that consumers treat each product as different from the other. This gives firms an opportunity to manoeuvre the price, capture a larger market share through advertisement and, thereby, to enhance their own profit. For instance, markets for motor cars, radios, TV sets, refrigerators, scooters, toilet soaps, toothpastes, etc., belong to this category of markets.

In the case of monopoly and perfect competition, the distinction between demand for a firm's product and that of the industry is not of much use from the managerial point of view. In case of monopoly, the industry is a one-firm industry and the demand for the firm's product is the same as that of the industry. In case of perfect competition, products of all firms of the industry are homogeneous; consumers do not distinguish between products of different firms; and the price for each firm is determined by the market forces (i.e., demand and supply for the industry as whole). Firms have only little opportunity to manoeuvre the prices permissible under local conditions and advertisement by a firm becomes effective for the whole industry. Therefore, conceptual distinction between demand for a firm's product and for that of the industry is not of much use in business decisions-making.

(c) Autonomous and derived demand

An *Autonomous demand* or *direct demand* for a commodity is one that arises on its own out of a natural desire to consume or possesses a commodity. An autonomous demand is independent of the demand for any other commodity. For example, consider the demand for commodities which arise directly from the biological or physical needs of human beings, e.g., demand for food, clothes and shelter. Demand for the these goods and the like is *autonomous demand.* Autonomous demand may also arise as a result of 'demonstration effect' of a rise in income, increase in population and advertisement of new products.

On the other hand, the demand for a commodity that arises because of the demand for some other commodity, called 'parent product', is called *derived demand.* For instance, demand for land, fertilizers and agricultural tools and implements is a derived demand because these goods are demanded because food is demanded. Similarly, demand for steel, bricks, cement, etc. is a derived demand–derived from the demand for house and other buildings. In general, the demand for producer goods or industrial inputs is a derived one. Also the demand for complementary goods (which complement the use of other goods) or for supplementary goods (which supplement or provide additional utility from the use of other goods) is a derived demand. For instance, petrol is a complementary good for automobiles and a chair is a complement to a table. Consider some examples of *supplementary goods*. Butter is a supplement to bread; mattress is a supplement to cot; and sugar is a supplement to tea–for some, it is a complement. Therefore, demand for petrol, chair and sugar would be considered as derived demand.

The conceptual distinction between autonomous demand (i.e., demand for a 'parent product') and derived demand would be useful from a businessmen's point of view to the extent that the former can serve as an indicator of the latter.

(d) Demand for durable and non-durable goods

Demand is also often classified under demand for durable and non-durable goods. *Durable goods* are those whose total utility or usefulness is not exhausted in a single or short-run use. Such goods can be used repeatedly or continuously over a period of time. Durable goods may be consumer goods as well as producer goods. *Durable consumer goods* include clothes, shoes, houses, furniture, utensils, refrigerator scooters, cars, etc. The durable producer goods include mainly the

items under 'fixed assets', such as building, plant, machinery, office furniture and fixtures, etc. The durable goods, both consumer and producers goods, may be further classified as 'semi-durables' (e.g., clothes and furniture) and 'durables' (e.g., residential and factory building cars etc.).

Non-durable goods, on the other hand, are those which can be used or consumed only once (e.g., food items) and their total utility is exhausted in a single use. This category of goods too may be grouped under *non-durable consumer goods* and *non-durable producer goods*. All food items, drinks, soaps, cooking fuel, (gas, kerosene, coal etc.), lighting, cosmetics, etc. fall in the former category. In the latter, fall goods such as raw materials, fuel and power, finishing materials, packing items, etc. The demand for non-durable goods depends largely on their current prices, consumers' income and fashion and is subject to frequent change whereas the demand for the durable good is also influenced by their expected price, income and change in technology. The demand for durable goods changes over a relatively longer period.

There is another point of distinction between the demand for durable and non-durable goods. *Durable goods create replacement demand whereas nondurable goods do not. Also, the demand for nondurable goods increases (or decreases) lineally whereas the demand for durable goods increases (or decreases) exponentially due to an increase in stock of durable goods and hence accelerated depreciation.*

(e) Short-term and long-term demand

Short-term demand refers to the demand for goods that are demanded over a short period. In this category are found mostly the fashion consumer goods, goods of seasonal use, inferior substitutes during the scarcity period of superior goods, etc. For instance, the demand for fashion wear is short-term demand though the demand for generic goods (trousers, shoes, ties, etc.) continues to remain a long-term demand. Similarly, demand for umbrella, raincoats, gum boots, cold drinks, ice creams, etc. is of a seasonal nature. The demand for such goods lasts till the season lasts. Some goods of this category are demanded for a very short period (1–2 weeks), e.g., New Year Greeting cards, candles and crackers on the occasion of Diwali.

Although some goods are used only seasonally they are of durable nature, e.g., electric fans, woollen garments. The demand for such goods is of a durable nature but it is subject to seasonal fluctuation. Sometimes, demand for certain goods suddenly increases because of scarcity of their superior substitutes. For example, when supply of cooking gas suddenly decreases, demand for kerosene, cooking coal and charcoal increases. In such cases, additional temporal demand is of a short-term nature.

Long-term demand, on the other hand, refers to the demand which exists over a long period. The change in long-term demand is perceptible only after a long period. Most generic goods have long-term demand; for example, demand for consumer and producer goods, durable and nondurable goods is long-term demand, though their different varieties or brands may only have a short-term demand.

Short-term demand depends, by and large, on the price of commodities, price of their substitutes, current disposable income of the consumers, their ability to adjust their consumption pattern and their susceptibility to advertisement of a new product. The long-term demand depends on the long-term income trends, availability of better substitutes, sales promotion, consumer credit facility, etc.

The short-term and long-term concepts of demand are useful in designing new products for established producers and choice of products for the new entrepreneurs, in pricing policy, and in determining and phasing the advertisement expenditure.

Check Your Progress

4. What does a demand curve show?
5. List examples of complementary goods.
6. What effect does demonstration effect have on the demand?

1.4 DEMAND ANALYSIS: ELASTICITY OF DEMAND

The laws of demand states how demand for and supply of a product respond to change in its price and other determinants. This law does not bring out the extent of responsiveness of demand and supply to the change in price. For example, the law of demand does not tell how much will be the change in quantity demanded when price changes by a certain amount or percentage. From the decision-making point of view, however, the knowledge of only the nature of relationships is not sufficient. What is more important is the extent of relationship or the degree of responsiveness of demand to changes in its determinants. The responsiveness of demand for a good to the change in its determinants is called the *elasticity of demand.* The concept of elasticity of demand was introduced into the economic theory by Alfred Marshall. The elasticity concept plays an important role in various business decisions and government policies.

Let us discuss the kinds of elasticity in detail

- **Price elasticity**
- **Cross elasticity**
- **Income elasticity**
- **Price expectation elasticity**

1.4.1 Price Elasticity of Demand

The price elasticity of demand is defined as the degree of responsiveness or sensitiveness of demand for a commodity to the changes in its price. More precisely, elasticity of demand is the percentage change in the quantity demanded of a commodity as a result of a certain percentage change in its price. A formal definition of price elasticity of demand (e_p) is given below.

$$e_p = \frac{\text{Percentage change in quantity demanded}}{\text{Percentage change in price}}$$

The measure of price elasticity (e_p) is called coefficient of price elasticity. The measure of price elasticity is converted into a more general formula for calculating coefficient of price elasticity given as

$$e_p = -\frac{\Delta Q}{Q_0} \div \frac{\Delta P}{P_0} = -\frac{\Delta Q}{Q_0} \cdot \frac{P_0}{\Delta P}$$

$$= -\frac{\Delta Q}{\Delta P} \cdot \frac{P_0}{Q_0} \quad \ldots (1.9)$$

where Q_o = original quantity demanded, P_o = original price, ΔQ = change in quantity demanded, and ΔP = change in price.

Note also that a *minus* sign (–) is generally inserted in the formula before the fraction with a view to making elasticity coefficient a non-negative value.

Point and arc elasticities of demand

The elasticity of demand is conventionally measured either at a *finite point* or between any *two finite points,* on the demand curve. The elasticity measured on a finite point of a demand curve is called point elasticity and the elasticity measured between any two finite points is called arc elasticity. Let us now look into the methods of measuring point and arc elasticities and their relative usefulness.

(a) Measuring point elasticity

The *point elasticity* of demand is defined as the proportionate change in quantity demanded in response to a *very* small proportionate change in price. The concept of point elasticity is useful where change in price and the consequent change in quantity demanded are very small.

The point elasticity may be symbolically expressed as:

$$e_p = -\frac{\partial Q}{\partial P} \cdot \frac{P}{Q} \quad \ldots (1.10)$$

• Measuring point elasticity on a linear demand curve

To illustrate the measurement of point elasticity of a linear demand curve, let us suppose that a *linear* demand curve is given by *MN* in Figure 1.10 and that we want to measure elasticity at point *P*. Let us now substitute the values from Figure 1.10 in Equation 1.9. As is obvious from the figure, $P = PQ$, and $Q = OQ$. What we need now is to find the values for ΔQ and ΔP. These values can be obtained by assuming *a very small* decrease in the price. But it will be difficult to depict these changes in the figure as $\Delta P \rightarrow O$ and hence $\Delta Q \rightarrow O$. There is however an easier way to find out the value for $\Delta Q/\Delta P$. In fact, the derivative $\frac{\partial Q}{\partial P}$ gives the slope of the demand curve *MN*. The slope of a straight line demand curve *MN*, at point *P* is geometrically given by *QN/PQ*. That is,

$$\frac{\partial Q}{\partial P} = \frac{QN}{PQ}$$

Since at point P, $P = PQ$ and $Q = OQ$, substituting these values in Equation (1.10), ignoring the minus sign, we get

$$e_p = \frac{QN}{PQ} \cdot \frac{PQ}{OQ} = \frac{QN}{OQ} \qquad ...(1.11)$$

Geometrically, $\frac{QN}{OQ} = \frac{PN}{PM}$

This may be proved as follows. If we draw a horizontal line from P to the vertical axis, there will be three triangles ΔMON, ΔMRP and ΔPQN (Figure 1.10) in which ΔMON, ΔMRP and ΔPQN are right angles. Therefore, the other corresponding angles of the three triangles will always be equal and hence, ΔMON, ΔMRP and ΔPQN are similar triangles.

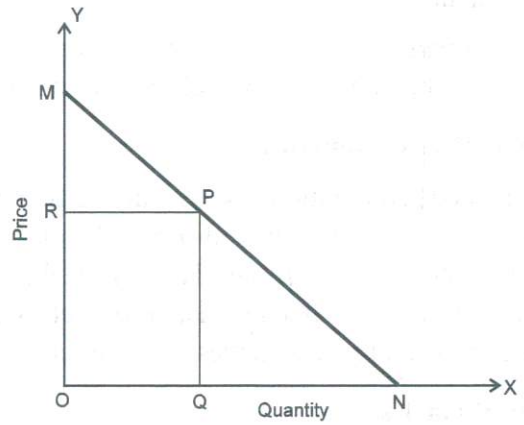


Fig. 1.10 *Point Elasticity of a Linear Demand Curve*

According to geometrical properties of similar triangles, the ratio of any two sides of a similar triangle is always equal to the ratio of corresponding sides of the other triangles. Therefore, in ΔPQN and ΔMRP,

$$\frac{QN}{PN} = \frac{RP}{PM} \qquad ...(1.12)$$

Since $RP = OQ$, by substituting OQ for RP in Equation (1.12), we get

$$\frac{QN}{PN} = \frac{OQ}{PM}$$

By proportionality rule, therefore,

$$\frac{QN}{OQ} = \frac{PN}{PM}$$

It may thus be concluded that price elasticity at point P (Figure 1.10) is given by

$$e_p = \frac{PN}{PM}$$

• Measuring point elasticity on a non-linear demand curve

Let us now explain the method of measuring point elasticity on a non-linear demand curve. Supposing we want to measure the elasticity of demand curve DD' at point P in Figure 1.11, let us draw a line (MN) tangent to the demand curve DD' at point P. Since demand curve DD' and the line MN pass through the same point (P) the

slope of demand curve and that of the line at this point is the same. Therefore, the elasticity of demand curve DD' at point P will be equal to the elasticity of demand line, MN, at point P. Elasticity of the line, MN, at point P can be measured (ignoring 'minus' sign) as

$$e_p = \frac{\partial Q}{\partial P} \cdot \frac{P}{Q}$$

$$\frac{QN}{PQ} \cdot \frac{PQ}{OQ} = \frac{QN}{OQ}$$

Geometrically, $QN/OQ = PN/PM$.

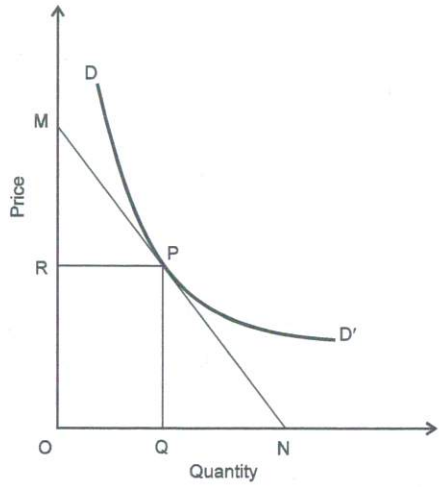


Fig. 1.11 *Point Elasticity on a Non-linear Demand Curve*

Given the graphical measurement of point elasticity, it is obvious that the elasticity at a point of a demand curve is the ratio between the lower and the upper segments of a linear demand curve from the point chosen for measuring point elasticity. That is,

$$e_p = \frac{\text{Lower segment}}{\text{Upper segment}}$$

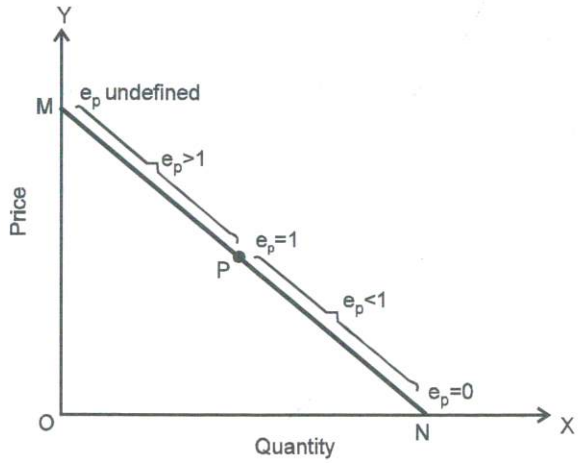


Fig. 1.12 *Point Elasticities of Demand*

It follows that at *mid-point* of a linear demand curve, $e_p = 1$, as shown at point P in Figure 1.12, because both lower and upper segments are equal (i.e., $PN =$

NOTES

PM). At any other point to the left of point P, $e_p > 1$, and at any point to the right of point P, $e_p < 1$.

- **Price elasticity at terminal points**

The price elasticity at terminal point N equals 0, i.e. at point N, $e_p = 0$. At terminal point M, however, price-elasticity is *undefined*, though most texts show that at terminal point M, $e_p = \infty$. According to William J. Baumol, a Nobel Prize winner, price elasticity at upper terminal point of the demand curve is undefined. It is undefined because measuring elasticity at terminal point (M) involves division of zero and division by zero is undefined. In his own words, 'Here the elasticity is not even defined because an attempt to evaluate the fraction p/x at that point forces us to commit the sign of dividing by zero. The reader who has forgotten why division by zero is immoral may recall that division is the reverse operation of multiplication. Hence, in seeking the quotient $c = a/b$ we look for a number, c, which when multiplied by b gives us the number a, i.e., for which $cb = a$. But if a is not zero, say $a = 5$, and b is zero, there is no such number because there is no c such that $c \times 0 = 5$.'

(b) Measuring arc elasticity

The concept of point elasticity is relevant where change in price and the resulting change in quantity is infinitesimally small. But where change in price and the consequent change in demand are substantial, the concept of *arc elasticity* is the relevant concept.

Arc elasticity is a measure of the average of responsiveness of the quantity demanded to a substantial change in the price. In other words, the measure of price elasticity of demand between two finite points on a demand curve is known as *arc elasticity.* For example, the measure of elasticity between points J and K (Figure 1.13) is the measure of arc elasticity. The movement from point J to K along the demand curve (D_x) shows a fall in price from ₹25 to ₹10 so that $\Delta P = 25 - 10 = 15$. The consequent increase in demand, $\Delta Q = 30 - 50 = -20$. The arc elasticity between point J and K and (moving from J to K) can be obtained by substituting these values in the elasticity formula.

$$e_p = -\frac{\Delta Q}{\Delta P} \cdot \frac{P}{Q} \qquad = -\frac{-20}{15} \cdot \frac{25}{30} = 1.11 \qquad \text{...(1.13)}$$

It means that a one per cent decrease in price of commodity X results in a 1.11 per cent increase in demand for it.

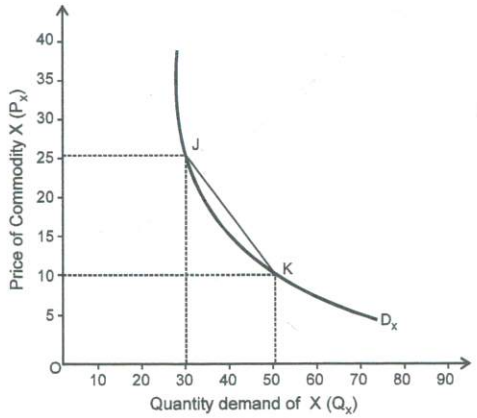


Fig. 1.13 Measuring Arc Elasticity

Problems in Using Arc Elasticity

The use of arc elasticity in economic analysis requires a good deal of carefulness because it is capable of being misinterpreted. Arc elasticity coefficients differ between the same two finite points on a demand curve if *direction* of change in price is reversed. That is, arc elasticity for a decrease in price will be different from that for the same increase in price between the same to points on a demand curve. For example, the price elasticity between points J and K—moving from J to K—is equal to 1.11. This is the elasticity for decrease in price from ₹25 to ₹10. But a reverse movement on the demand curve, i.e. from point K to J implies an increase in price from ₹10 to ₹25 which will give a different elasticity coefficient. In case of movement from point K to J, $P = 10$, $\Delta P = 10 - 25 = -15$, $Q = 50$, and $\Delta Q = 50 - 30 = 20$. Substituting these values in the elasticity formula, we get

$$e_p = -\frac{20}{-15} \cdot \frac{10}{50} = 0.26 \qquad \ldots (1.14)$$

The measure of arc elasticity coefficient in Eq. (1.14) for the reverse movement in price is obviously different from the one given in Eq. (1.13). Therefore, while measuring the arc elasticity, the direction of price change should be carefully noted, otherwise it may yield misleading conclusions.

A method suggested to resolve this problem is to use the average of upper and lower values of P and Q in fraction, P/Q, so that the formula is

$$e_p = \frac{\Delta Q}{\Delta P} \cdot \frac{(P_1 + P_2)/2}{(Q_1 + Q_2)/2}$$

$$= \frac{Q_1 - Q_2}{P_1 - P_2} \cdot \frac{(P_1 + P_2)/2}{(Q_1 + Q_2)/2} \qquad \ldots (1.15)$$

Substituting the values from the example, we get

$$e_p = -\frac{30 - 50}{25 - 10} \cdot \frac{(10 + 25)/2}{(30 + 50)/2} = 0.58$$

This method has its own drawbacks as the elasticity coefficient calculated through this formula, refers to the elasticity of demand at mid-point between points J and K (Figure 1.13). Elasticity coefficient (0.58) is not applicable for the whole range of price-quantity combinations at different points between J and K on the demand curve (Figure 1.12). It gives only mean of the elasticities between the two points. It is important to note that elasticity between the mid-point and the upper point J or lower point K will be different. Thus, this method does not give one measure of elasticity.

Nature of Demand Curves and Elasticities

Generally, elasticity of a demand curve throughout its length is not the *same* (Figure 1.12). It varies between O and ∞, or in other words,

$$0 \leq e_p \leq \infty$$

In some cases, however, the elasticity remains the same throughout the length of the demand curve. Such demand curves can be placed in the following categories: (i) perfectly inelastic ($e_p = 0$); (ii) unitary elastic ($e_p = 1$); and (iii) perfectly elastic ($e_p = \infty$). These three types of demand curves are illustrated in Figure 1.14 (a), (b) and (c), respectively.

However, from demand analysis point of view, demand is classified as elastic ***demand*** and inelastic demand. If price elasticity of demand for a product is greater than one, i.e. $e_p > 1$, the demand is called elastic demand. And, if $e_p < 1$, it is called inelastic demand. Also, a normal demand curve—linear or non-linear—has two segments—elastic and inelastic segments.

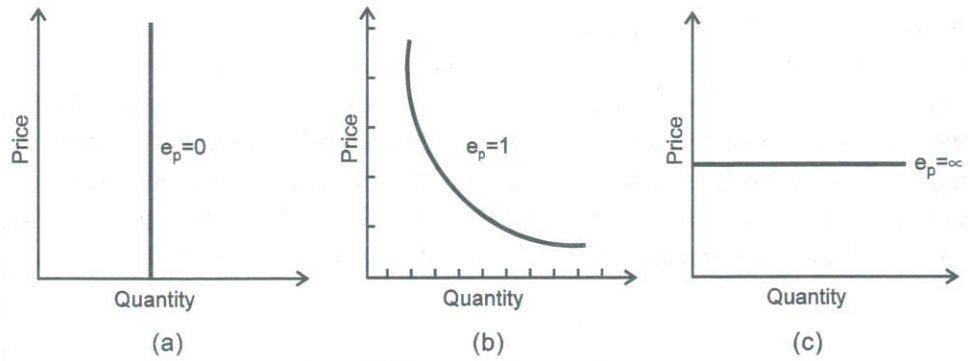


Fig. 1.14 *Types of Demand Elasticity*

Slope of the demand curve and price elasticity

The elasticity of a demand curve is often judged by its appearance: the flatter the demand curve, the greater the elasticity, and vice versa. But this conclusion is misleading because two demand curves with different slopes may have the same elasticity at a given price. In fact, what the appearance of a demand curve reveals is its *slope*, not the *elasticity*. The *slope* of the demand curve is the marginal relationship between change in price (D*P*) and change in quantity demanded (D*Q*). The slope of demand curve is expressed as D*P*/D*Q*. It is the reciprocal of the slope D*Q*/D*P* which appears in the elasticity formula, not the slope itself.

It has been shown that below (i) demand curves having different slopes may have the same elasticity at a given price, and (ii) demand curves having the same slope may have different elasticity at a given price.

(a) Elasticities of demand curves having different slopes

Let us first illustrate that two demand curves with different slops may have the same elasticity at a given price. In Figure 1.15, demand curves *AB* and *AD* have different slopes. This may be proved as follows:

Slope of the demand curve $AB = \frac{OA}{OB}$, and

Slope of the demand curve $AD = \frac{OA}{OD}$

Note that in these ratios, numerator *OA* is common to both the fractions, but in case of denominators *OB* < *OD*.

Therefore, $\frac{OA}{OB} < \frac{OA}{OD}$

Obviously, the slopes of the two demand curves are different.

Let us now show, that at a given price, both the demand curves have the same elasticity. As shown in Figure 1.15, at a given price *OP*, the relevant points for

measuring the elasticity are Q and R on the demand curves AB and AD, respectively. Recall that the elasticity at a point on a linear demand curve is obtained as

$$e_p = \frac{\text{Lower Segment}}{\text{Upper Segment}}$$

Thus, at point Q on demand curve AB, $e_p = QB/QA$, and at point R on AD, $e_p = RD/RA$.

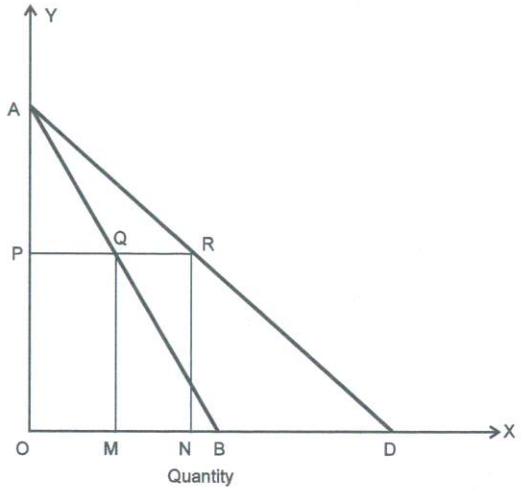


Fig. 1.15 *Demand Curves having Different Slopes*

It can be geometrically proved that the two elasticities are equal, i.e.

$$\frac{QB}{QA} = \frac{RD}{RA} \qquad ...(1.16)$$

Let us first consider ΔAOB. If we draw a horizontal line from point Q to intersect the vertical axis at point P, and an ordinate from Q to M at the horizontal axis, we have three triangles—ΔAOB, ΔAPQ and ΔQMB. Note that ΔAOB, ΔAPQ and ΔQMB are right-angled triangles. Therefore, all the three triangles are *right-angled triangles*. As noted above, the ratios of their two corresponding sides of similar right-angle triangles are always equal. Considering only the relevant triangles, ΔAPQ and ΔQMB, we have

$$\frac{BQ}{MQ} = \frac{AQ}{AP}$$

Since $MQ = OP$, by substituting OP for QM in ratio BQ/MQ, we get

$$\frac{BQ}{OP} = \frac{AQ}{AP}$$

Therefore, $\frac{BQ}{AQ} = \frac{OP}{AP}$ = Elasticity of demand curve AB, at point Q or at price OP.

We can similarly prove that

$$\frac{RD}{RA} = \frac{OP}{PA} = \text{Elasticity of demand curve } AD \text{ at point } R \text{ or at price } OP.$$

Thus, the Eq. (1.16) is proved, i.e., at price OP,

$$\frac{QB}{QA} = \frac{RD}{RA} = \frac{PO}{PA}$$

It is thus proved that two demand curves with different slopes have the same elasticity at a given price.

(b) Elasticities of parallel demand curves

Let us now show that two demand curves having the same slope have different elasticities. Consider the demand curves *JK* and *LM* in Figure 1.16. The demand curves *JK* and *LM* are parallel and therefore have the same slope. Point *R* on the line *JK* and point *Q* on the line *LM* show the quantities demanded at a given price, *OP*.

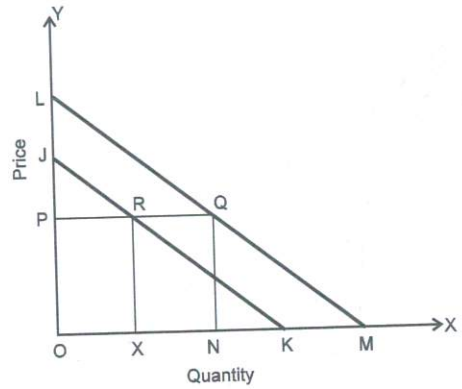


Fig. 1.16 *Demand Curves having the Same Slope*

The elasticity at point *R* on demand curve *JK* equals *RK/RJ* and the elasticity at *Q* on demand curve *LM* equals *QM/QL*. It can be easily proved that

$$\frac{RK}{RJ} \neq \frac{QM}{QL}$$

Following the logic of the preceding section, it can be proved that

$$\frac{RK}{RJ} = \frac{PO}{PJ}$$

and $$\frac{QM}{QL} = \frac{PO}{PL}$$

Note that, *PO* being the common numerator to both ratios, $PJ < PL$. Therefore,

$$\frac{PO}{PJ} > \frac{PO}{PL}$$

It is thus proved that $$\frac{RK}{RJ} > \frac{QM}{QL}$$

It may be concluded from the above that the demand curves having the same slope may have different elasticities, and the demand curves having different slopes may have the same elasticities, both at a given price.

Determinants of price elasticity of demand

Price elasticity of demand varies from commodity to commodity. While the demand for some commodities is highly elastic, the demand for others is highly inelastic. Let us discuss the main determinants of the price elasticity of demand.

1. **Availability of substitutes:** One of the most important determinants of elasticity of demand for a commodity is the availability of its substitutes.

The closer the substitute, the greater the elasticity of demand for the commodity. For instance, coffee and tea may be considered as close substitutes for one another. Therefore, if the price of one of these goods increases, its demand decrease more than proportionate increase in its price because consumers switch over to the relatively cheaper substitute. Besides, the wider the range of the substitutes, the greater the elasticity. For instance, soaps, toothpastes, cigarettes, etc. are available in different brands, each brand being a close substitute for the other. Therefore, the price-elasticity of demand for each brand will be much greater than the generic commodity. On the other hand, sugar and salt do not have close substitutes and hence their price elasticity is lower.

2. **Nature of commodity:** The nature of a commodity also affects the price elasticity of its demand. Commodities can be grouped as luxuries, comforts, and necessities, on the basis of their nature. Demand for luxury goods (e.g., luxury cars and decoration items) is more elastic than the demand for other kinds of goods because consumption of luxury goods can be dispensed with or postponed when their prices rise. On the other hand, consumption of necessary goods, (e.g., sugar, clothes and vegetables) cannot be postponed, and hence their demand is inelastic. Demand for comforts is generally more elastic than that for necessities and less elastic than the demand for luxuries. Commodities may also be classified as durable goods and perishable or non-durable goods. Demand for durable goods is more elastic than that for non-durable goods, because when the price of the former increases, people either get the old one repaired instead of replacing it or buy a 'second hand'.
3. **Proportion of income spent on a commodity:** Another factor that influences the elasticity of demand for a commodity is the proportion of income which consumers spend on a particular commodity. If proportion of income spent on a commodity is very small, its demand will be less elastic, and vice versa. Classic examples of such commodities are salt, matches, books and toothpastes, which claim a very small proportion of consumer's income. Demand for these goods is generally inelastic because increase in the price of such goods does not substantially affect consumer's consumption pattern and the total purchasing power. Therefore, people continue to purchase almost the same quantity even when their price increases.
4. **Time factor:** Price elasticity of demand depends also on the time consumers take to adjust to a new price: the longer the time taken, the greater the elasticity. For, over a period of time consumers are able to adjust their expenditure pattern to price changes. For instance, if price of cars decreases, demand may not increase immediately unless people possess excess purchasing power. Over time, however, people can adjust their expenditure pattern so that they can buy a car at (new) lower price.
5. **Range of alternative uses of a commodity:** The wider the range of alternative uses of a product, the higher the price elasticity of its demand for the decrease in price but less elastic for the rise in price. As the price of a multi-use commodity decreases, people extend their consumption to its other

uses. Therefore, the demand for such a commodity generally increases more than the proportionate decrease in its price. For instance, milk can be taken as it is, it may be converted into curd, cheese, ghee and buttermilk. The demand for milk will therefore be highly elastic for decrease in their price. Similarly, electricity can be used for lighting, cooking, heating, and for industrial purposes, therefore demand for electricity is highly price elastic for decrease in its price. For the same reason, however, demand for such goods is inelastic for the rise in their price.

6. **The proportion of market supplied:** Price elasticity of market demand depends also on the proportion of the market supplied at the ruling price. If less than half of the market is supplied at the ruling price, elasticity of demand will be higher than if more than half of the market is supplied. That is, demand curve is more elastic at the upper half than at the lower half.

Price elasticity and marginal revenue

Lwet us have a look at one of the most important uses of the price elasticity of demand, used especially in business decision-making. It pertain to the relationship between price elasticity and the marginal change in the total revenue of the firm planning to change the price of its product. The relationship between price elasticity and the marginal revenue (MR) can be derived as follows:

Let us suppose that a given output, Q, is being sold at a price P, so that the total revenue, TR, equals P times Q, i.e.,

$$TR = P \cdot Q \qquad \ldots (1.17)$$

Since P and Q in Eq. (1.17) are inversely related, a question arises, whether a change in P will increase or decrease or leave the TR unaffected. It depends on whether MR is greater than or less than or equal to zero, i.e. whether

$$MR > 0,\ MR < 0, \text{ or } MR = 0$$

The marginal revenue, (MR) can be obtained by differentiating $TR = PQ$ with respect to P as shown below. Since $TR = P.\ Q$.

$$MR = \frac{\partial (P.Q)}{\partial Q} = P\frac{\partial Q}{\partial Q} + Q\frac{\partial P}{\partial Q} = P + Q\frac{\partial P}{\partial Q}$$

$$MR = P\left(1 + \frac{Q}{P} \cdot \frac{\partial P}{\partial Q}\right) \qquad \ldots (1.18)$$

Note that $\frac{Q}{P} \cdot \frac{\partial P}{\partial Q}$ is the reciprocal of the elasticity which equals

$$-\frac{P}{Q} \cdot \frac{\partial Q}{\partial P}$$

Therefore, $$\frac{Q}{P} \cdot \frac{\partial P}{\partial Q} = -\frac{1}{e}$$

By substituting $-\frac{1}{e}$ for $\frac{Q}{P} \cdot \frac{\partial P}{\partial Q}$ in Eq. (1.18), we get

$$MR = P\left(1 - \frac{1}{e}\right) \qquad \ldots (1.19)$$

Equation (1.19) gives the relationship between price-elasticity (e) and MR. The relationship between price elasticity and marginal revenue is shown graphically in Figure 1.17.

1. Price elasticity and total revenue

Given the relationship between marginal revenue and price elasticity of demand in Eq. (1.19), the decision-makers can easily know whether or not it is advantageous to change the price. Given the Eq. (1.19), if $e = 1$, $MR = 0$. Therefore, change in price will not cause any change in TR. If $e < 1$, $MR < 0$ and, therefore, TR decreases when price decreases and TR increases when price increases. And, if $e > 1$, $MR > 0$, then TR increases if price decreases and TR increases when price increases.

The TR function is graphed in panel (a) and the demand and MR functions are presented in panel (b) of Figure 1.17. As the figure shows, at point P on the demand curve, $e = 1$ where output, $Q = 50$. Below point P, i.e., at price ₹10, $e <$ 1 and above point P, $e > 1$. It can be seen in panel (a) of Figure 1.17 that TR increases over the range of demand curve having $e > 1$; TR reaches its maximum level where $e = 1$; and it decreases when $e < 1$.

The relationship between price elasticity and TR is summed up in Table 1.4. As the table shows, when demand is *perfectly inelastic* (i.e., $e_p = 0$ as in the case of a vertical demand line) there is no decrease in quantity demanded when price is raised and vice versa. Therefore, a rise in price increases the total revenue and vice versa.

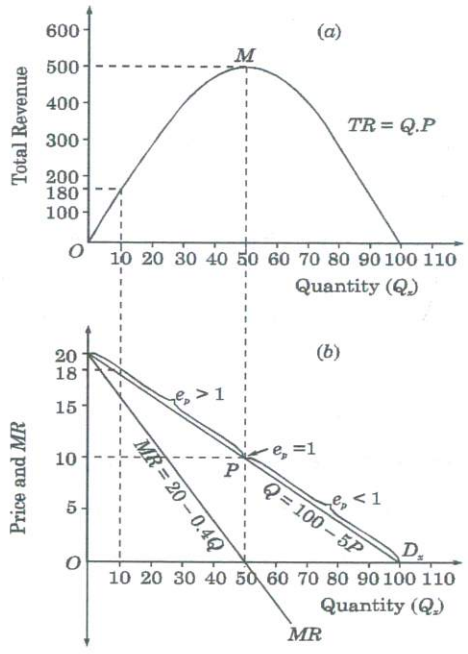


Fig. 1.17 *Price Elasticity and Total Revenue*

In the case of an *inelastic demand* (i.e., $e_p < 1$), quantity demanded increases by less than the proportionate decrease in price and hence the total revenue falls when price falls. The total revenue increases when price increases because quantity demanded decreases by less than the proportionate increase in price.

If demand for a product is *unit elastic* ($e_p = 1$) the quantity demanded increases (or decreases) in the proportion of decrease (or increase) in the price. Therefore, total revenue remains unaffected.

If demand for a commodity has $e_p > 1$, change in quantity demanded is greater than the proportionate change in price. Therefore, the total revenue increases when price falls and vice versa.

The case of *infinitely elastic* demand represented by a horizontal straight line is rare. Such a demand line implies that a consumer has the opportunity to buy any quantity of a commodity and the seller can sell any quantity of a commodity, at a given price. It is the case of a commodity being bought and sold in a perfectly competitive market. A seller, therefore, cannot charge a higher or a lower price.

Table 1.4 *Elasticities, Price-Change and Change in TR*

Elasticity coefficient	***Nature of Demand***	***Change in Price***	***Change in TR***
$e_p = 0$	Perfectly inelastic	Increase Decrease	Increases Decreases
$e_p < 1$	Inelastic	Increase Decrease	Increases Decreases
$e_p = 1$ Elastic	Unitary Decrease	Increase in *TR*	No change
$e_p > 1$	Elastic	Increase Decrease	Decrease Increases
$e_p = ¥$	Infinitely Elastic	Increase Decreases	Decreases to zero infinitely increase*

* Subject to the size of the market.

2. Price elasticity and marginal revenue

The relationship between price-elasticity and the total revenue (*TR*) can be known more precisely by finding the relationship between price-elasticity and marginal revenue (*MR*). *MR* is the first derivative of *TR*-function and $TR = P.Q$ (where P = price, and Q = quantity sold). The relationship between price elasticity, *MR* and *TR* is shown below.

$$TR = P.Q.$$

Promotional or advertisement elasticity of sales

The expenditure on advertisement and on other sales-promotion activities does help in promoting sales, but *not in the same degree at all levels of the total sales*. The concept of advertisement elasticity is useful in determining the optimum level of advertisement expenditure. The concept of *advertisement elasticity* assumes greater significance in deciding on advertisement expenditure, particularly when the government imposes restriction on advertisement cost or there is competitive advertising by the rival firms. Advertisement elasticity (e_A) of sales may be defined as

$$e_A = \frac{\Delta S/S}{\Delta A/A} = \frac{\Delta S}{\Delta A} \cdot \frac{A}{S}$$

where S = sales; ΔS = increase in sales; A = initial advertisement cost, and ΔA = additional expenditure on advertisement.

• **Interpretation of advertisement elasticity**

The advertisement elasticity of sales varies between $e_A = 0$ and $e_A = \infty$. Interpretation of some measures of advertising elasticity is given below.

Table 1.5 Interpretations of Measures of Advertising Elasticity

Elasticities	*Interpretation*
$e_A = 0$	Sales do not respond to the advertisement expenditure.
$e_A > 0$ but < 1	Increase in total sales is less than proportionate to the increase in advertisement expenditure.
$e_A = 1$	Sales increase in proportion to the increase in expenditure on advertisement.
$e_A >$	Sales increase at a higher rate than the rate of increase of advertisement expenditure.

• **Determinants of advertisement elasticity**

Some of the important factors which determine advertisement elasticity are the following:

(*i*) **Level of total sales:** In the initial stages of sale of a product, particularly of one which is newly introduced in the market, the advertisement elasticity is greater than unity. As sales increase, the elasticity decreases. For instance, after the potential market is supplied, the function of advertisement is to create additional demand by attracting more consumers to the product, particularly those who are slow in adjusting their consumption expenditure to provide for new commodities. Therefore, demand increases at a rate lower than the rate of increase in advertisement expenditure.

(*ii*) **Advertisement by rival firms:** In a highly competitive market, the effectiveness of advertisement by a firm is also determined by the relative effectiveness of advertisement by the rival firms.

(*iii*) **Cumulative effect of past advertisement:** In case expenditure incurred on advertisement in the initial stages is not adequate enough to be effective, elasticity may be very low. But over time, additional doses of advertisement expenditure may have a cumulative effect on the promotion of sales and advertising elasticity may increase considerably.

(*iv*) **Other factors:** Advertisement elasticity is also affected by other factors affecting the demand for a product, e.g., change in products' price, consumers' income and growth of substitutes and their prices.

1.4.2 Cross Elasticity of Demand

The cross elasticity is the measure of responsiveness of demand for a commodity to the changes in the price of its substitutes and complementary goods. For instance, cross-elasticity of demand for tea (T) is the percentage change in its quantity demanded with respect to the change in the price of its substitute, coffee (C). The formula for measuring cross elasticity of demand for tea ($e_{t,c}$) with respect to price of coffee (P_c) is

NOTES

$$e_{t,c} = \frac{\text{Proportionate change in demand for tea } (Q_t)}{\text{Proportionate change in price of coffee } (P_c)}$$

$$= \frac{P_c}{Q_t} \cdot \frac{\Delta Q_t}{\Delta P_c} \quad \ldots (1.20)$$

The cross elasticity of demand for coffee (Q_c) with respect to price of tea (P_t) is

$$e_{c,t} = \frac{P_t}{Q_c} \cdot \frac{\Delta Q_c}{\Delta P_t}$$

For example, suppose that price of coffee (P_c) increases from ₹10 to ₹15, per 10 grams, and as a result demand for tea increases from 20 tonnes to 30 tonnes per week, price of tea remaining constant. By substituting these value in Eq. (1.20), we get cross elasticity of demand for tea with respect to price of coffee, as

$$e_{t,c} = \frac{10}{20} \cdot \frac{20-30}{10-15}$$

$$= \frac{10}{20} \cdot \frac{-10}{-5}$$

$$= 1.0$$

It is important to note that cross-elasticity between any two substitute goods is always positive.

The same formula is used to measure the cross-elasticity of demand for a good in response to change in the price of its complementary goods. Electricity to electrical gadgets, patrol to automobile, butter to bread, sugar and milk to tea and coffee, are examples of complementary goods. Note that the demand for complementary goods has negative cross elasticity, for increase in the price of a good decreases the demand for its complementary goods.

An important aspect of cross-elasticity is that if cross-elasticities between two goods are positive, the two may be considered as substitutes for each other. Also the greater the cross elasticity, the closer the substitute. Similarly, if cross-elasticity of demand for two related goods is negative, the two may be considered as complementary of each other: the higher the negative cross elasticity, the higher the degree of complementarity.

1.4.3 Income Elasticity of Demand

Apart from price of a product and its substitutes, an important determinant of demand for a product is consumer's income. As noted earlier, the relationship between demand for normal and luxury goods and consumer's income is of positive nature, unlike the negative price-demand relationship. That is, the demand for normal goods and services increases with increase in consumer's income and vice versa. The responsiveness of demand to the change in consumer's income is known as income elasticity of demand.

Income elasticity of demand for a product, say X (i.e., e_y) is defined as

$$e_y = \frac{\Delta X_q / X_q}{\Delta Y / Y} = \frac{Y}{X_q} \cdot \frac{\Delta X_q}{\Delta Y} \quad \ldots (1.21)$$

where X_q = quantity of X demanded; Y = disposable income; DX_q = change in quantity demanded of X; and DY = change in income.

Unlike price elasticity of demand (which is negative except in case of Giffen goods), income elasticity of demand is positive because of a positive relationship between income and demand for a product. There is an exception to this rule. Income elasticity of demand for an *inferior good* is negative, because of negative income-effect. The demand for inferior goods decreases with increase in consumer's income and vice versa. When income increases, consumers switch over to the consumption of superior commodities. That is, they substitute superior goods for inferior ones. For instance, when income rises, people prefer to buy more of rice and wheat and less of inferior foodgrains like *bajra*, and ragi, and use more of taxi and less of bus service, and so on.

Nature of commodity and income elasticity

For all normal goods, income elasticity is positive though the degree of elasticity varies in accordance with the nature of commodities. Consumer goods are generally grouped under three categories, viz. necessities (essential consumer goods), comforts, and luxuries. The general pattern of income elasticities for goods of different categories for increase in income and their impact on sales are given in Table 1.6.

The income elasticity of demand for different categories of goods may however vary from household to household and from time to time, depending on choice, taste and preference of the consumers, levels of their consumption and income, and their susceptibility to 'demonstration effect'. The other factor which may cause deviation from the general pattern of income elasticities is the frequency of increase in income. If income increases regularly and frequently, income-elasticities as given in Table 1.5 will conform to the general pattern.

Some important uses of income elasticity are as follows:

First, the concept of income elasticity can be used to estimate the future demand provided the rate of increase in income and income elasticity of demand for the products are known. The knowledge of income elasticity can thus be useful in forecasting demand, when changes in personal incomes are expected, other things remaining the same.

Table 1.6 *Income Elasticity of Different Consumer Goods*

Commodities	*Coefficient of income elasticity*	*Impact on expenditure*
1. Necessities	Less than unity ($e_y < 1$)	Less than proportionate change in income
2. Comforts	Almost equal to unity ($e_y \cong 1$)	Almost proportionate change in income
3. Luxuries	Greater than unity ($e_y > 1$)	More than proportionate increase in income

Second, the concept of income elasticity may also be used to define the 'normal' and 'inferior' goods. The goods whose income elasticity is positive for all levels of income are termed as 'normal goods'. On the other hand, the goods for which income elasticities are negative, beyond a certain level of income, are termed as 'inferior goods'.

(iv) Price expectations elasticity

Sometimes, mainly during the period of price fluctuation, consumer's price expectations play a much more important role in determining demand for a commodity than any other factor. The concept of price expectation elasticity refers to the expected change in price in future as a result of change in current prices of a product. The elasticity of price expectation is defined and measured through the following formula:

$$e_e = \frac{\Delta P_f / P_f}{\Delta P_c / P_c}$$

$$= \frac{\Delta P_f}{\Delta P_c} \cdot \frac{P_c}{P_f} \qquad ...(1.22)$$

The coefficient, e_e, gives the measure of expected percentage change in future price (P_f) as a result of 1 per cent change in current price (P_c). If $e_e > 1$, it indicates that future change in price will be greater than the present changes in price, and vice versa. If $e_e = 1$, it indicates that the future change in price will be equal to the change in the current price. If $e_e = 0$, it indicates no change in future price as a result of change in current price. The concept of elasticity of price-expectation is very useful in formulating future pricing policy. For example, if $e_e > 1$, it indicates sellers will be able to sell more in the future at higher prices. Thus, businessmen may accordingly determine their future pricing policy.

1.4.4 The Uses of Elasticity

Although Samuelson condemned the concept of elasticity as an 'essentially arbitrary' and a more or less useless concept, it has many important uses in both economic analysis and formulation of economic policies. Some important uses of elasticity concept are as follows:

First, the concept of elasticity of demand plays a crucial role in business decisions regarding manoeuvering of prices with a view to making larger profits. For instance, when cost of production is increasing, the firm would like to pass incremental cost on to the consumer by raising the price. Firms may decide to change the price even without change in cost of production. But, whether this action (raising the price following the rise in cost or otherwise) will prove beneficial or not depends on (a) the price-elasticity of demand for the products; and (b) its cross elasticity because when the price of a product increases, its substitutes become automatically cheaper even if their prices remain unchanged. Raising price will be beneficial only if (i) demand of a product has an elasticity less than one, and (ii) demand for its substitute is much less elastic. Similarly, a firm not finding it feasible to increase the price during a period of growth in consumer's income would like to increase the production. The additional production can be determined only through the income elasticity of demand for the product, other factors remaining the same. Although most businessmen intuitively are aware of the elasticity of demand of the goods, they do make use of precise estimates of elasticity of demand as it adds precision to the business decisions.

Second, the elasticity concept can be used also in formulating government policies, particularly in its taxation policy meant to raise revenue or to control prices; in granting subsidies to the industries; in determining prices for public utilities; in fixing the prices of essential goods; and in determining export and import duties and the rate of currency devaluation. To consider an example, suppose government wants to impose sales tax on a particular commodity with the sole objective of raising revenue. Whether adequate revenue can be raised or not will depend on the price elasticity of that commodity. If demand is highly elastic, the revenue yield will be much less than expected. The sales tax will rather cause price distortion and affect production adversely.

Third, the concept of elasticity is useful in economic analysis, at least for specifying the relationship between the dependent and independent variables. Besides, the elasticity concept is used in specifying and estimating demand functions. The most commonly used form of demand function in applied research is the 'constant elasticity demand function' of the form.

$$Q_X = A\, P^{B}{}_X, Y^C, P^D, E^{FT}$$

in which P_X, Y, P^D and E^{FT} represent, respectively, price of X, consumer's income, price of other goods and a trend factor of 'taste', and if arc elasticity coefficients, and A is a constant.

To sum up, elasticity concept is undoubtedly a useful concept and has a wide application to economic analysis and policy.

Check Your Progress

7. List the kinds of demand elasticity.
8. What is point elasticity?
9. What is cross elasticity of demand?

1.5 DEMAND FORECASTING

From the firm's point of view, the knowledge of demand theory alone is not enough for planning and scheduling production of goods and services. The knowledge of future demand for its product, at least an approximate estimate of future demand is an essential aspect of many business decisions. There are various methods of demand forecasting—varying from simple survey methods to sophisticated method. The choice of methods depends on the purpose, experience and expertise of the forecaster, availability of necessary data and the quality and consistency of primary and secondary data.

Need for demand forecasting

Demand forecasting is predicting future demand for a product. The information regarding future demand is essential for planning and scheduling production, purchase of raw materials, acquisition of finance and advertising. It is much more important where large-scale production is being planned and production involves a long gestation period. The information regarding future demand is also essential for the existing firms to be able to avoiding under or over-production. Most firms

are, in fact, very often confronted with the question as to what would be the future demand for their product because they will have to acquire inputs and plan their production accordingly. The firms are hence required to estimate the future demand for their product. Otherwise, their functioning will be shrouded in uncertainty and their objective may be defeated.

This problem may not be of a serious nature for small firms which supply a very small fraction of the total demand, and whose product caters to the short-term, seasonal demand or to demand of a routine nature. Their past experience and business skills may suffice for their purpose in planning and production. But, firms working on a large scale find it extremely difficult to obtain fairly accurate information regarding the future market demand. In some situations, it is very difficult to obtain information needed to make even short-term demand forecasts and extremely difficult to make long-term forecasts or to determine how changes in specific demand variables like price, advertisement expenditure, credit terms, prices of competing products, etc. will affect demand. It is nevertheless indispensable for the large firms to have at least a rough estimate of the demand prospects. Demand forecast plays an important role in planning to acquire inputs, both men and material (raw material and capital goods), organizing production, advertising the product, and organizing sales channels. These functions can hardly be performed satisfactorily in an atmosphere of uncertainty regarding demand prospects for the product. The prior knowledge of market size, therefore, becomes an extremely important element of decision making by the large-scale firms.

Methods of Demand Forecasting

As mentioned above, there are various methods of demand forecasting and the choice of the method depends on such factors as purpose and perspective of demand forecasting, experience and expertise of the forecaster, availability of required data and the quality and consistency of both primary and secondary data. The various methods of demand forecasting are presented in the chart given below. However, we also provide a list of various methods of demand forecasting. Demand forecasting methods are classified on the basis of (i) the nature of data used—primary or secondary data; and (ii) applicability and usefulness of the secondary data. In case secondary data are not available or availing secondary data involves long time and cost, *survey methods* are generally used.

I. Survey Methods

1. Consumer Survey: Direct interviews
   - Complete enumeration
   - Sample survey
   - End-use method
2. Opinion Poll Methods
   - Expert opinion survey
     - (*a*) simple method
     - (*b*) Delphi method
   - Market studies and experiments
     - (*a*) Market tests
     - (*b*) Laboratory tests

II. Statistical Methods

1. Trend Projection Methods
   - Graphical method
   - Box-Jenkins method
   - Least square method
2. Barometric Methods
   - Lead lag indicators
   - Diffusion indices
3. Econometric Methods
   - Regression method
     (*a*) Bi variate regression method (*b*) Multi variate method
   - Simultaneous equation methods

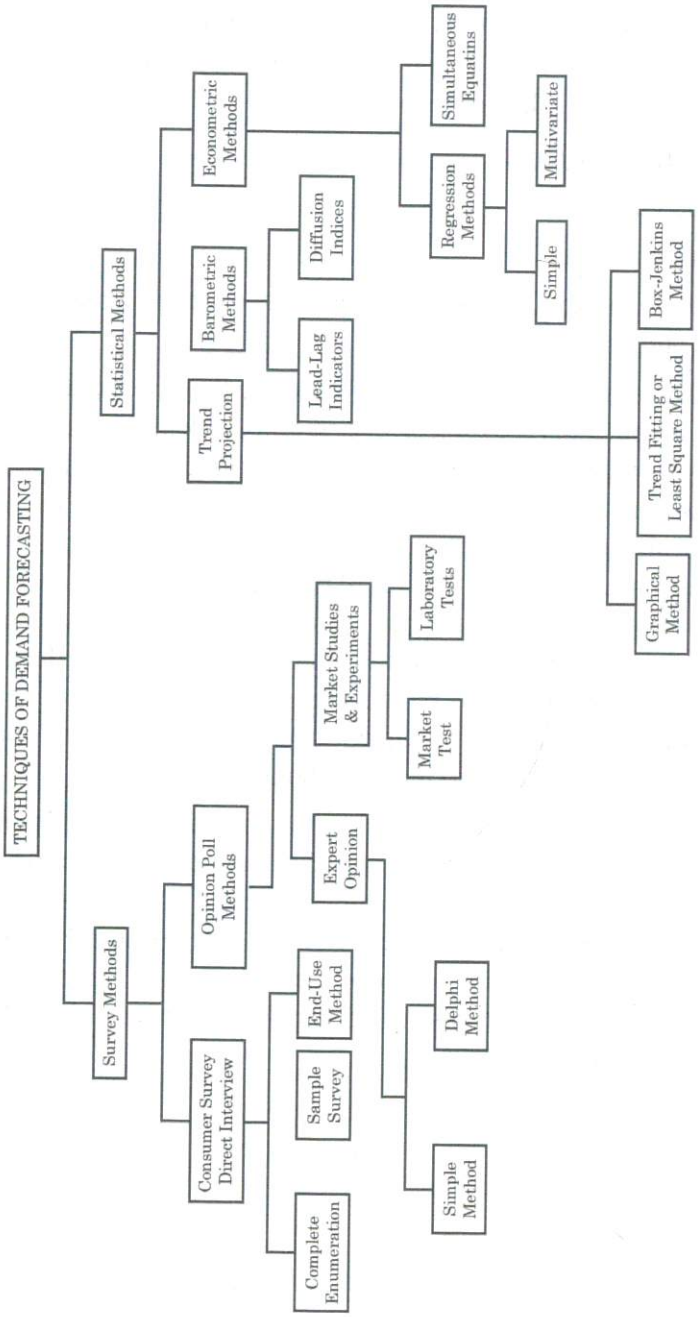


Fig. 1.18 *Techniques of Demand Forecasting*

1.5.1 Survey Methods

Survey methods are generally used where the purpose is to make short-run forecast of demand. Under this method, consumer surveys are conducted to collect information about their intentions and future purchase plans. This method includes:

(*i*) survey of potential consumers to elicit information on their intentions and plan;

(*ii*) opinion polling of experts, i.e., opinion survey of market experts and sales representatives and through market studies and experiments.

The following techniques are used to conduct the survey of consumers and experts.

1. Consumer Survey Methods: Direct interviews

The consumer survey method of demand forecasting involves direct interview of the potential consumers. It may be in the form of

(*a*) complete enumeration,

(*b*) sample survey, or

(*c*) end-use method.

These consumer survey methods are used under different conditions and for different purposes.

The most direct and simple way of assessing future demand for a product is to interview the potential consumers or users and to ask them what quantity of the product they would be willing to buy at different prices over a given period, say, one year. This method is known as *direct interview method*. This method may cover almost all the potential consumers or only selected groups of consumers from different cities or parts of the area of consumer concentration. When all the consumers are interviewed, the method is known as *complete enumeration survey or comprehensive interview method* and when only a few selected representative consumers are interviewed, it is known as *sample survey method*. In the case of industrial inputs, interviews or postal inquiry of only *end-users* of a product may be required. Let us now describe these methods in detail.

(a) **Complete enumeration method:** In this method, almost all potential users of the product are contacted and asked about their future plan of purchasing the product in question. The quantities indicated by the consumers are added together to obtain the probable demand for the product. For example, if a majority of n out of m number of households in a city report the quantity (q) they are willing to purchase of a commodity, then total probable demand (D_p) may be calculated as

$$D_p = q_1 + q_2 + q_3 + \ldots + q_n$$

$$= \sum_{i=1}^{n} q_i$$

where q_1, q_2, q_3, etc. denote demand by the individual household 1, 2, 3, etc.

This method has certain limitations. It can be used successfully only in the case of those products whose consumers are concentrated in a certain region or locality. In the case of a widely dispersed market, this method may not be physically possible or may prove very costly in terms of both money and time. Besides, the demand forecast through this method may not be reliable for many reasons: (*i*) consumers themselves may not know their actual demand in future and hence may be unable or unwilling to answer the query; (*ii*) even if they answer, their answer to hypothetical questions may be only hypothetical and not real; (*iii*) consumers' response may be biased according to their own expectations about the market conditions; and (*iv*) their plans may change with a change in the factors not included in the questionnaire.

(b) **Sample survey method:** Under this method, only a few potential consumers and users selected from the relevant market through a sampling method are surveyed. Method of survey may be direct interview or mailed questionnaire to the sample-consumers. On the basis of the information obtained, the probable demand may be estimated through the following formula:

$$D_p = \frac{H_R}{H_S} (H \,.\, A_D)$$

where D_p = probable demand forecast; H = census number of households from the relevant market; H_s = number of households surveyed or sample households; H_R = number of households reporting demand for the product; A_D = average expected consumption by the reporting households (= total quantity reported to be consumed by the reporting households , numbers of households).

This method is simpler, less costly, and less time-consuming than the comprehensive survey method. This method is generally used to estimate short-term demand from business firms, government departments and agencies and also by the households who plan their future purchases. Business firms, government departments and other such organizations budget their expenditure at least one year in advance. It is, therefore, possible for them to supply a fairly reliable estimate of their future purchases. Even the households making annual or periodic budget of their expenditure can provide reliable information about their purchases.

The sample survey method is widely used to forecast demand. This method, however, has some limitations similar to those of *complete enumerations* or exhaustive survey method. The forecaster, therefore, should not attribute more reliability to the forecast than is warranted. Besides, the sample survey method can be used to verify the demand forecast made by using quantitative or statistical methods. Although some authors suggest that this method should be used to supplement the quantitative method for forecasting rather than to replace it, this method can be gainfully used where the market is localized. Sample survey method can be of greater use in forecasting where quantification of variables (e.g., feelings, opinion and expectations) is not possible and where consumer's behaviour is subject to frequent changes.

(c) **The end-use method:** The end-use method of demand forecasting has a considerable theoretical and practical value, especially in *forecasting demand for inputs*. Making forecasts by this method requires building up a schedule of probable aggregate future demand for inputs by consuming industries and various other sectors. In this method, technological, structural and other changes which might influence the demand, are taken into account in the very process of estimation. This aspect of the end-use approach is of particular importance.

Stages in the end-use method

The end-use method of demand forecasting consists of four distinct stages of estimation. In the ***first stage***, it is necessary to identify and list all the possible users of the product in question. This is, of course, a difficult process, but it is fundamental to this method of forecasting. Difficulty arises because published data on the end-users is rarely available. Despatch records of the manufacturers, even if available, need not necessarily enumerate all the final users. Records of the sales pattern by individual firms or establishments are difficult to be assembled. In several cases, the distribution of the products covers such a wide network and there are so many wholesale and retail agencies in the chain, that it would be virtually impossible to organize and collect data from all these sources so as to know all the final end-uses of the product.

Where relevant and adequate data is not available, the managers need to have a through knowledge of the product and its uses. Such knowledge and experience need to be supplemented by consultations and discussions with manufacturers or their association, traders, users, etc. Preparation of an exhaustive list of all possible end-users is, in any case, a necessary step. Despite every effort made to trace all the end–users, it is quite likely that some of the current uses of the product are overlooked. In order to account for such lapses, it may be necessary at the final stage of estimation to provide some margin for error. A margin or allowance is also necessary to provide for possible new applications of the product in the future.

The ***second stage*** of this method involves fixing suitable technical 'norms' of consumption of the product under study. Norms have to be established for each and every end-use. Norms are usually expressed in physical terms either per unit of production of the complete product or in, some cases, per unit of investment or per capita use. Sometimes, the norms may have to be on value basis. But value-based norms should be avoided as far as possible because it might be rather difficult to specify later the types and sizes of the product in question if value norms are used.

The establishment of norms is also a difficult process mainly due to lack of data. For collecting necessary data, the questionnaire method is generally employed. The preparation of a suitable questionnaire is of vital importance in the end-use method, as the entire subsequent analysis has to be based on and conclusions to flow mainly from the information collected through the questionnaires. Where estimating future demand is called for in great detail, such as by types and sizes of the concerned product, framing of the questionnaire requires a good knowledge

of all the variations of the product. For a reliable forecast, it is necessary that response is total; if not, then as high as possible.

Having established the technical norms of consumption for the different industries and other end-uses of the product, the ***third stage*** is the application of the norms. For this purpose, it is necessary to know the desired or targeted levels of output of the individual industries for the reference year and also the likely development in other economic activities which use the product and the likely output targets.

The ***fourth stage*** in the end-use method is to aggregate the product-wise or use-wise content of the item for which the demand is to be forecast. This aggregate result gives the estimate of demand for the product as a whole for the terminal year in question. By the very nature of the process of estimation described here, it is obvious that the end-use approach results in what may be termed as a 'derived' demand.

Advantages of the end-use method

The end-use method has two exclusive advantages. ***First,*** it is possible to work out the future demand for an industrial product in considerable details by types and size. In other methods, the future demand can be estimated only at aggregate levels. This is because past data are seldom available in such details as to provide the types and sizes of the product demanded by the economy. Hence, projections made by using the past data, either by the trend method, regression techniques or by historical analogies produce only aggregate figures for the product in question. On the other hand, by probing into the present use-pattern of consumption of the product, the end-use approach provides every opportunity to determine the types, categories and sizes likely to be demanded in future.

Second, in forecasting demand by the end-use approach, it is possible to trace and pinpoint at any time in future as to where and why the actual consumption has deviated from the estimated demand. Besides, suitable revisions can also be made from time to time based on such examination. If projections are based on other methods and if actual consumption falls below or rises above the estimated demand, all that one can say is that the economy has or has not picked up as anticipated. One cannot say exactly which use of the product has not picked up and why. In the case of end-use method, however, one can.

2. Opinion poll methods

The opinion poll methods aim at collecting opinions of those who are supposed to possess knowledge of the market, e.g., sales representatives, sales executives, professional marketing experts and consultants. The opinion poll methods include:

(*a*) Expert-opinion method,

(*b*) Market studies and experiments.

(a) **Expert-opinion method:** Firms having a good network of sales representatives can put them to the work of assessing the demand for the product in the areas, regions or cities that they represent. Sales representatives, being in close touch with the consumers or users of goods,

are supposed to know the future purchase plans of their customers, their reaction to the market changes their response to the introduction of a new product, and the demand for competing products. They are, therefore, in a position to provide at least an approximate, if not accurate, estimate of likely demand for their firm's product in their region or area. The estimates of demand thus obtained from different regions are added up to get the overall probable demand for a product. Firms not having this facility, gather similar information about the demand for their products through the professional market experts or consultants, who can, through their experience and expertise, predict the future demand. This method is also known as *opinion poll method.*

Although this method too is simple and inexpensive, it has its own limitations. ***First,*** estimates provided by the sales representatives or professional experts are reliable only to an extent depending on their skill to analyse the market and their experience. ***Second,*** demand estimates may involve the subjective judgement of the assessor which may lead to over or under-estimation. ***Finally,*** the assessment of market demand is usually based on inadequate information available to the sales representatives who have only a narrow view of the market. The factors of wider implication, such as change in GNP, availability of credit, future prospects of the industry, etc., fall outside their purview. These methods include simple and delphi methods.

The delphi method: The Delphi method of demand forecasting is an extension of the *simple expert opinion poll* method. This method is used to consolidate the divergent expert opinions and to arrive at a compromise estimate of future demand. The process is simple.

Under the Delphi method, the experts are provided information on estimates of forecasts of other experts along with the underlying assumptions. The experts may revise their own estimates in the light of forecasts made by other experts. The consensus of experts about the forecasts constitutes the final forecast. It may be noted that the empirical studies conducted in the USA have shown that unstructured opinions of the experts is the most widely used technique of forecast. This may appear a bit unusual in as much as this gives the impression that sophisticated techniques, e.g., simultaneous equations model and statistical methods, are not the techniques which are used most often. However, the unstructured opinions of the experts may conceal the fact that information used by experts in expressing their forecasts may be based on sophisticated techniques. The Delphi technique can be used for cross-checking information on forecasts.

(b) **Market studies and experiments:** An alternative method of collecting necessary information regarding demand is to carry out market studies and experiments on consumer's behaviour under actual, though controlled, market conditions. This method is known in common parlance as *market experiment method.* Under this method, firms first select some areas of the representative markets—three or four cities having similar features, viz. population, income levels, cultural and social background, occupational distribution, choices and preferences of consumers. Then, they carry out

market experiments by changing prices, advertisement expenditure and other controllable variables in the demand function under the assumption that other things remain the same. The controlled variables may be changed over time either simultaneously in all the markets or in the selected markets. After such changes are introduced in the market, the consequent changes in the demand over a period of time (a week, a fortnight, or month) are recorded. On the basis of data collected, elasticity coefficients are computed. These coefficients are then used along with the variables of the demand function to assess the demand for the product.

Alternatively, market experiments can be replaced by *consumer clinics* or *controlled laboratory experiment*. Under this method, consumers are given some money to buy in a stipulated store goods with varying prices, packages, displays, etc. The experiment reveals the consumers' responsiveness to the changes made in prices, packages and displays, etc. Thus, the laboratory experiments also yield the same information as the market experiments. But the former has an advantage over the latter because of greater control over extraneous factors and its somewhat lower cost.

Limitations

The market experiment methods have the following serious limitations and disadvantages which reduce the usability and reliability of this method:

First, a very important limitation of the experimental methods is that they are *very expensive*. Therefore, experimental methods cannot be afforded by small firms.

Second, being a costly affair, experiment are usually carried out on a scale too small to permit generalization with a high degree of reliability.

Third, experimental methods are based on short-term and controlled conditions which may not exist in an uncontrolled market. Hence the results may not be applicable to the uncontrolled long-term conditions of the market.

Fourth, changes in socio-economic conditions during the *field experiments*, such as *local strikes or lay-offs, advertising programme* by competitors, political changes, natural calamities may invalidate the results.

Finally, a big disadvantage of experimental methods is that 'tinkering with price increases may cause a permanent loss of customers to competitive brands that might have been tried.'

Despite these limitations, however, the *market experiment method* is often used to provide an alternative estimate of demand and also 'as a check on results obtained from statistical studies.' Besides, this method generates elasticity co-efficients which are necessary for statistical analysis of demand relationships. For example, an experiment of this kind was conducted by Simmons Mattress Company (US). It put on sale two types of identical mattress—one with Simmons label and the other with an unknown name at the same price and then at different prices for determining the cross-elasticity. It was found that at the equal price, Simmons mattress sold 15 to 1; and at a price higher by 5 dollars it sold 8 to 1, and at a price higher by 25 per cent, it sold almost 1 to 1.

1.5.2 Statistical Methods

In this section, we will explain statistical methods which utilize historical (time-series) and cross-section data for estimating long-term demand. Statistical methods are considered to be superior techniques of demand estimation for the following reasons:

(*i*) In the statistical methods, the element of subjectivity is minimum.

(*ii*) The method of estimation is scientific, as it is based on the theoretical relationship between the dependent and independent variables.

(*iii*) Estimates are relatively more reliable.

(*iv*) Estimation involves smaller cost.

Statistical methods of demand projection include the following techniques:

(1) Trend projection methods

(2) Barometric methods

(3) Econometric method

1. **Trend projection methods:** Trend projection method is a 'classical method' of business forecasting. This method is essentially concerned with the study of movement of variables through time. The use of this method requires a long and reliable time-series data. The trend projection method is used under the assumption that the factors responsible for the past trends in the variable to be projected (e.g., sales and demand) will continue to play their part in future in the same manner and to the same extent as they did in the past in determining the magnitude and direction of the variable. This assumption may be quite justified in many cases.

However, since cause-and-effect relationship is not revealed by this method, the projections made on the trend basis are considered by many as a mechanical or a 'naïve' approach. Nevertheless, 'There is nothing uncomplimentary in the adoption of such an approach. It merely represents one of the several means to obtain an insight of what the future may possibly be and whether or not the projections made using these means are to be considered as most appropriate will depend very much on the reliability of past data and on the judgement that is to be exercised in the ultimate analysis.'

In projecting demand for a product, the trend method is applied to time-series data on sales. Firms of a long standing may obtain time-series data on sales from their own sales department. New firms can obtain necessary data from the older firms belonging to the same industry.

There are following three techniques of trend projection based on time-series data:

(*a*) Graphical method

(*b*) Fitting trend equation or least square method

(*c*) Box–Jenkins method

In order to explain these methods, let us suppose that a local bread manufacturing company wants to assess the demand for its product for the

years 2006, 2007 and 2008. For this purpose, it uses time-series data of its total sales over the past 10 years. Suppose its time-series sales data is given as in Table 1.7.

Table 1.7 Time-Series Data on Sale of Bread

Year	*Sales of Bread ('000 tonnes)*
1996	100
1997	120
1998	110
1999	150
2000	180
2001	140
2002	200
2003	180
2004	210
2005	250

Let us first use the graphical method and project demand for only one year, 2008.

(a) **Graphical method:** Under this method, annual sales data is plotted on a graph paper and a line is drawn through the plotted points. Then a free hand line is so drawn that the total distance between the line and the points is minimum. This is illustrated in Figure 1.19 by the dotted lines. The dotted line *M* is drawn through the mid-values of variations and line *S* is a straight trend line. The solid, fluctuating line shows the actual trend, while the dotted lines show the secular trend. By extending the trend lines (market *M* and *S*), we can forecast an approximate sale of 260,000 tonnes in 2004. Although this method is very simple and least expensive, the projections made through this method are not very reliable. The reason is that the extension of the trend line involves subjectivity and personal bias of the analyst. For example, an optimist may take a short-run view, say since 2003, and extend the trend line beyond point *P* towards *O*, and predict a sale of 300,000 tonnes of bread in 2008. On the other hand, a conservative analyst may consider the fluctuating nature of sales data and expect the total sale in 2008 to remain the same as in 2005 as indicated by the line *PC*. One may even predict a fall in the sale to 250,000 tonnes, if one over-emphasizes the fluctuating nature of sales in one's judgement. This is indicated by the line *PN*.

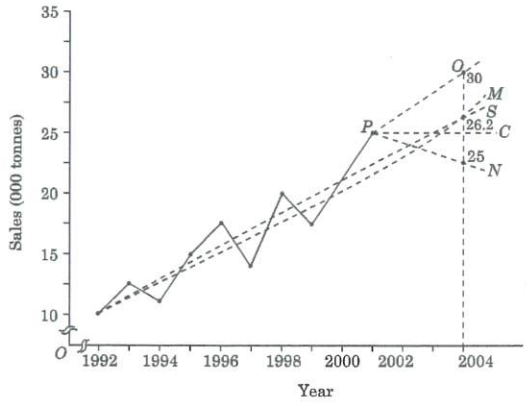


Fig. 1.19 Trend Projection

(b) **Fitting trend equation: Least square method:** Fitting trend equation is a formal technique of projecting the trend in demand. Under this method, a trend line (or curve) is fitted to the time-series sales data with the aid of statistical techniques. The form of the trend equation that can be fitted to the time-series data is determined either by plotting the sales data (as shown in Figure 1.19) or by trying different forms of trend equations for the best fit.

When plotted, a time-series data may show various trends. Let us discuss here only the most common types of trend equations, viz. (*i*) linear, and (*ii*) exponential trends.

(i) **Linear trend:** When a time-series data reveals a rising trend in sales, then a straight-line trend equation of the following form is fitted:

$$S = a + bT \quad \ldots (1.23)$$

where S = annual sales, T = time (years), a and b are constants. The parameter b gives the measure of annual increase in sales.

The coefficients a and b are estimated by solving the following two equations based on the principle of least square.

$$\Sigma S = na + b\Sigma T \quad \ldots (i)$$

$$\Sigma ST = a\Sigma T + b\Sigma T^2 \quad \ldots (ii)$$

The terms included in equations (*i*) and (*ii*) are calculated using sales data given in Table 1.7 and presented in Table.

By substituting numerical values in equations (*i*) and (*ii*), we get

$$1640 = 10\,a + 55\,b \quad \ldots (iii)$$

$$10240 = 55\,a + 385\,b \quad \ldots (iv)$$

By solving equations (*iii*) and (*iv*), we get the trend equation as

$$S = 8.26 + 1.48T$$

Table 1.8 *Estimation of Trend Equation*

Year	*Sales*	*T* (*'000 tonnes*)	T^2	*ST*
1996	100	1	1	100
1997	120	2	4	240
1998	110	3	9	330
1999	150	4	16	600
2000	180	5	25	900
2001	140	6	36	840
2002	200	7	49	1400
2003	180	8	64	1440
2004	210	9	81	1890
2005	250	10	100	2500
$n = 10$	$\Sigma S = 1640$	$\Sigma T = 55$	$\Sigma T^2 = 385$	$\Sigma ST = 10240$

Having estimated the trend equation, it is quite easy to project the sales for 2006, 2007 and 2008, i.e., for the 11th, 12th and 13th year, respectively. The calculation procedure is given follow.

2006 $S_2 = 8.26 + 1.480\ (11) = 171{,}000$ tonnes

2007 $S_3 = 8.26 + 1.480\ (12) = 186{,}000$ tonnes

2008 $S_4 = 8.26 + 1.480\ (13) = 261{,}000$ tonnes

NOTES

Treatment of the abnormal years

Time series data on sales may, more often than not, reveal abnormal years. An abnormal year is one in which sales are abnormally low or high. Such years create a problem in fitting the trend equation and lead to under or over-statement of the projected sales. Abnormal years should, therefore, be carefully analysed and data be suitably adjusted. The abnormal years may be dealt with (*i*) by excluding the year from time-series data, (*ii*) by adjusting the sales figures of the year to the sales figures of the preceding and succeeding years, or (*iii*) by using a 'dummy' variable.

(ii) **Exponential trend:** When sales (or any dependent variable) have increased over the past years at an increasing rate or at a constant percentage rate, then the appropriate trend equation to be used is an exponential trend equation of any of the following forms:

(1) $Y = ae^{bT}$ ... (1.24)

or its semi-logarithmic form

$\log Y = \log a + bT$ ... (1.25)

This form of trend equation is used when growth rate is constant.

(2) Double-log trend of the form

$Y = aT^b$ ... (1.26)

or its double logarithmic form

$\log Y = \log a + b \log T$ ... (1.27)

This form of trend equation is used when growth rate is increasing.

(3) Polynomial trend of the form

$Y = a + bT + cT^2$ ... (1.28)

In these equations a, b and c are constants, Y is sales, T is time and $e = 2.718$. Once the parameters of the equations are estimated, it becomes quite easy to forecast demand for the years to come.

The trend method is quite popular in business forecasting because of its simplicity. It is simple to apply because only time-series data on sales are required. The analyst is supposed to possess only a working knowledge of statistics. Since data requirement of this method is limited, it is also inexpensive. Besides, the trend method yields fairly reliable estimates of the future course of demand.

Limitations

The ***first*** limitation of this method arises out of its assumption that the past rate of change in the dependent variable will persist in the future too. Therefore, the forecast based on this method may be considered to be reliable only for the period during which this assumption holds.

Second, this method cannot be used for short-term estimates. Also, it cannot be used where trend is cyclical, with sharp turning points of troughs and peaks.

Third, this method, unlike regression analysis, does not bring out the measure of relationship between dependent and independent variables. Hence, it does not yield the necessary information (e.g., price and income elasticies) which can be used for future policy formulations. These limitations need to be borne in mind while making the use of this method.

(c) **Box-Jenkins method:** The Box-Jenkins method of forecasting is used only for short-term predictions. Besides, this method is suitable for forecasting demand with only stationary time-series sales data. Stationary time-series data is one which does not reveal a long-term trend. In other words, the Box-Jenkins technique can be used only in those cases in which time-series analysis depicts monthly or seasonal variation recurring with some degree of regularity. When sales data of various commodities are plotted, many commodities will show a seasonal or temporal variation in sales. For example, sale of woollen clothes will show a hump during months of winter in all the years under reference. The sale of New Year Greeting Cards will be particularly very high in the last week of December every year. Similarly, the sale of desert coolers is very high during the summers each year. This is called seasonal variation. The Box-Jenkins technique is used for predicting demand where time-series sales data reveal this kind of seasonal variation.

According to the Box-Jenkins approach, any stationary time-series data can be analysed by the following three models:

(*i*) Auto-regression model

(*ii*) Moving average model

(*iii*) Autoregressive-moving average model

The three models are, in fact, the three stages of the Box-Jenkins method. The auto regressive-moving average model is the final form of the Box-Jenkins model. The purpose of the three models of the Box-Jenkins method is to explain movements in the stationary series with minimized error term, i.e., the unexplained components of stationary series.

The steps and models of the Box-Jenkins approach are described here briefly with the purpose of acquainting the reader with this approach rather than providing the entire methodology.

Steps in box-jenkins approach

As mentioned above, Box-Jenkins method can be applied only to stationary time-series data. Therefore, the ***first step*** in the Box-Jenkins approach is to eliminate trend from the time-series data. Trend is eliminated by taking first differences of time-series data, *i.e.*, subtracting observed value of one period from the observed value of the preceding year. After trend is eliminated, a stationary time-series is created.

The ***second step*** in the Box-Jenkins approach is to make sure that there is seasonality in the stationary time-series. If a certain pattern is found to repeat over time, there is seasonality in the stationary time-series.

The ***third step*** involves use of models to predict the sales in the intended period. We give here a brief description of the Box-Jenkins models which are used in the same sequence.

(*i*) **Auto-regressive model:** In a general autoregressive model, the behaviour of a variable in a period is linked to the behaviour of the variable in future periods. The general form of the autoregressive model is given below.

$$Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \ldots + a_n Y_{t-n} + e_t \quad \ldots (1.29)$$

This model states that the value of Y in period t depends on the values of Y in periods $t-1, t-2, \ldots t-n$. The term e_t is the random portion of Y_t that is not explained by the model. If estimated value of one or some of the coefficients $a_1, a_2, \ldots a_n$ are different from zero, it reveals seasonality in data. This completes the second step.

The model, however, does not specify the relationship between the value of Y_t and residuals (e_t) of previous periods. The Box-Jenkins method uses moving average method to specify the relationship between Y_t and e_t, values of residuals in previous years. This is the third step. Let us now look at the moving average model of the Box-Jenkins method.

(*ii*) **Moving average model:** The moving average model estimates Y_t in relation to residuals (e_t) of the previous years. The general form of moving average model is given below.

$$Y_t = m + b_1 e_{t-1} + b_2 e_{t-2} \ldots + b_p e_{t-p} + e_t \quad \ldots (1.30)$$

where m is mean of the stationary time-series and $e_{t-1}, e_{t-2}, \ldots e_{t-p}$ are the residuals, the random components of Y in $t-1$, $t-2$, … $t-p$ periods, respectively.

(*iii*) **Auto-regressive-moving average model:** After moving average model is estimated, it is combined with autoregressive model to form the final form of the Box-Jenkins model, called autoregressive-moving average model, given below.

$$Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \ldots + a_n Y_{t-n} + b_1 e_{t-1} + b_2 e_{t-2} + \ldots + b_p e_{t-p} + e_t \quad \ldots (1.31)$$

Clearly, the Box-Jenkins method of forecasting demand is a sophisticated and complicated method. Without the aid of computers it is rather an impracticable method.

Moving average method: An alternative technique

As noted above, the moving average model of the Box-Jenkins method is a part of a complicated technique of forecasting demand in period t on the basis of its past values. There is a simple, or rather a naïve, yet useful method of using moving average to forecast demand. This simple method assumes that demand in a future year equals the average of demand in the past years. The formula of simple moving average method is expressed as

$$D_t = \frac{1}{N} (X_{t-1} + X_{t-2} + \ldots + X_{t-n})$$

where D_t = demand in period t; $X_{t-1, t-2 \ldots t-n}$ = demand or sales in previous years; N = number of preceding years.

According to this method, the likely demand for a product in period t equals the average of demand (sales) in several preceding years. For example, suppose that the number of refrigerators sold in the past 7 years in a city is given as Table 1.9 and we want to forecast demand for refrigerators for the year 2002.

Table 1.9 *Sale of Refrigerators: 2000–06*

Year	2000	2001	2002	2003	2004	2005	2006
Sales ('000)	11	12	12	13	13	15	15

Given this sales data, demand for 2007 will be computed as follows:

$$D_{2007} = \frac{1}{7}(15 + 15 + 13 + 13 + 12 + 12 + 11) = 13$$

Thus, the demand for refrigerators for 2007 is forecast at 13,000 units. Now suppose that the actual sales of refrigerators in the city in 2007 turns out to be 15,000 refrigerators against the forecast figure of 13,000. Given the actual sales figure for 2007, the demand for 2008 can be forecast as

$$D_{2008} = \frac{1}{7}(15 + 15 + 15 + 13 + 13 + 12 + 12) = 13.57$$

Note that, in the moving average method, the sale of 2007 is added and the sale of 2000 (the last of the preceding years) is excluded from the formula.

This moving average method is simple and can be used to make only short term forecasts. This method has a serious limitation, which has to be borne in mind while using this method. In the case of rising trend in sales, this method yields an underestimate of future demand, as can be seen in the above example. And, in case of declining trend in sales, it may yield an overestimate of future demand. One way of reducing the margin of over and under-estimate is to take the average of fluctuations and add it to the moving average forecasts. This method is, in fact, more suitable where sales fluctuate frequently within a limited range.

2. Barometric method of forecasting

The barometric method of forecasting follows the method meteorologists use in weather forecasting. Meteorologists use the barometer to forecast weather conditions on the basis of movements of mercury in the barometer. Following the logic of this method, many economists use economic indicators as a barometer to forecast trends in business activities. This method was first developed and used in the 1920s by the Harvard Economic Service. This technique was, however, abandoned as it had failed to predict the Great Depression of the 1930s. The barometric technique was, however, revived, refined and developed further in the late 1930s by the National Bureau of Economic Research (NBER) of the US. It has since then been used often to forecast business cycles in the US.

It may be noted at the outset that the barometric technique was developed to forecast the general trend in overall economic activities. This method can nevertheless be used to forecast demand prospects for a product, not the actual quantity expected to be demanded. For example, development and allotment of land by the Delhi Development Authority (DDA) to the Group Housing Societies

(a lead indicator) indicates higher demand prospects for cement, steel, bricks and other construction materials.

The basic approach of barometric technique is to construct an index of relevant *economic indicators* and to forecast future trends on the basis of movements in the index of economic indicators. The indicators used in this method are classified as follows:

(*a*) Leading indicators

(*b*) Coincidental indicators

(*c*) Lagging indicators

A time-series of various indicators is prepared to read the future economic trend. The *leading series* consists of indicators which move up or down ahead of some other series. Some examples of the leading series are: (*i*) index of net business (capital) formation; (*ii*) new orders for durable goods; (*iii*) new building permits; (*iv*) change in the value of inventories; (*v*) index of the prices of the materials; (*vi*) corporate profits after tax.

The *coincidental series*, on the other hand, are the ones that move up or down simultaneously with the level of economic activity. Some examples of the coincidental series are: (*i*) number of employees in the non-agricultural sector; (*ii*) rate of unemployment; (*iii*) gross national product at constant prices; (*iv*) sales recorded by the manufacturing, trading and the retail sectors.

The *lagging series,* consists of those indicators which follow a change after some time-lag. Some of the indices that have been identified as lagging series by the NBER are: (*i*) labour cost per unit of manufactured output, (*ii*) outstanding loans, and (*iii*) lending rate for short-term loans.

The time-series of various indicators are selected on the basis of the following criteria:

(*i*) Economic significance of the indicator: the greater the significance, the greater the score of the indicator

(*ii*) Statistical adequacy of time-series indicators: a higher score is given to an indicator provided with adequate statistics

(*iii*) Conformity with overall movement in economic activities

(*iv*) Consistency of series to the turning points in overall economic activity

(*v*) Immediate availability of the series

(*vi*) Smoothness of the series

The problem of choice may arise because some of the indicators appear in more than one class of indicators. Furthermore, it is not advisable to rely on just one of the indicators. This leads to the usage of what is referred to as the 'diffusion index'. A diffusion index copes with the problem of differing signals given by the indicators. A diffusion index is the percentage of rising indicators. In calculating a diffusion index, for a group of indicators, scores allotted are 1 to rising series, ½ to constant series and zero to falling series. The diffusion index is obtained by the ratio of the number of indicators, in a particular class, moving up or down to the

total number of indicators in that group. Thus, if three out of six indicators in the lagging series are moving up, the index shall be 50 per cent. It may be noted that the most important is the diffusion index of the leading series. However, there are problems of identifying the leading indicator for the variable under study. Also, lead time is not of an invariable nature.

Leading indicators can be used as inputs for forecasting aggregate economic variables, GNP, aggregate consumers' expenditure, aggregate capital expenditure, etc. The only advantage of this method is that it overcomes the problem of forecasting the value of independent variable under the regression method. The major limitations of this method are: (*i*) it can be used only for short-term forecasting, and (*ii*) a leading indicator of the variable to be forecast is not always easily available.

3. Econometric methods of forecasting

The econometric methods combine statistical tools with economic theories to estimate economic variables and to forecast the intended economic variables. The forecasts made through econometric methods are much more reliable than those made through any other method. The econometric methods are, therefore, most widely used to forecast demand for a product, for a group of products and for the economy as a whole. Our concern here is primarily to explain econometric methods used for forecasting demand for a product.

An econometric model may be a single-equation regression model or it may consist of a system of simultaneous equations. Single-equation regression serves the purpose of demand forecasting in the case of most commodities. But, where relationships between economic variables are complex and variables are so interrelated that unless one is determined, the other cannot be determined, a single-equation regression model does not serve the purpose. In that case, a system of simultaneous equations is used to estimate and forecast the target variable.

The econometric methods are briefly described here under two basic methods.

(a) Regression method

(b) Simultaneous equations model

(a) **Regression Method:** Regression analysis is the most popular method of demand estimation. This method combines economic theory and statistical techniques of estimation. Economic theory is employed to specify the determinants of demand and to determine the nature of the relationship between the demand for a product and its determinants. Economic theory thus helps in determining the general form of demand function. Statistical techniques are employed to estimate the values of parameters in the estimated equation. In regression technique of demand forecasting, the analysts estimate the demand function for a product. In the demand function, the quantity to be forecast is a 'dependent variable' and the variables that affect or determine the demand (the dependent variable) are called 'independent' or 'explanatory' variables. For example, demand for cold drinks in a city may be said to depend largely on 'per capita income' of the city and its population. Here demand for cold drinks is a 'dependent variable' and 'per capita income' and 'population' are the 'explanatory' variables.'

While specifying the demand functions for various commodities, the analyst may come across many commodities whose demand depends, by and large, on a single independent variable. For example, suppose in a city demand for such items as salt and sugar is found to depend largely on the population of the city. If this is so, then demand functions for salt and sugar are *single-variable demand functions*. On the other hand, the analyst may find that demand for sweets, fruits and vegetables, etc. depends on a number of variables like commodity's own price, price of its substitutes, household income, population, etc. Such demand functions are called *multi-variable demand functions*. For single-variable demand functions, simple regression equation is used and for multiple variable functions, multi-variable equation is used for estimating demand function. The single-variable and multi-variable regressions are explained below.

(i) **Simple or bivariate regression technique:** In simple regression technique, a single independent variable is used to estimate a statistical value of the 'dependent variable', that is, the variable to be forecast. The technique is similar to trend fitting. An important difference between the two is that in trend fitting the independent variable is 'time' (t) whereas in a regression equation, the chosen independent variable is the single most important determinant of demand. Besides, the regression method is less mechanical than the trend fitting method of projection.

Suppose we have to forecast demand for sugar for 2007 on the basis of 7-year data given in Table 1.10. When this data is graphed, it will produce a continuously rising trend in demand for sugar with rising population. This show a linear trend. Now, the demand for sugar in 2007 can be obtained by estimating a regression equation of the form.

$$Y = a + bX \qquad \ldots (1.32)$$

where Y is sugar consumed, X is population, and a and b are constants.

For an illustration, consider the hypothetical data on quarterly consumption of sugar given in Table 1.10.

Table 1.10 *Quarterly Consumption of Sugar*

Year	*Population (millions)*	*Sugar Consumed ('000) tonnes*
2000	10	40
2001	12	50
2002	15	60
2003	20	70
2004	25	80
2005	30	90
2006	40	100

Equation (1.32) can be estimated by using the 'least square' method. The procedure is the same as shown in Table 1.8. That is, the parameters a and b can be estimated by solving the following two linear equations:

$$\Sigma Y_i = na + b\Sigma X_i \qquad \ldots (i)$$

$$\Sigma X_i Y_i = \Sigma X_i\, a + bX^2{}_i \qquad \ldots (ii)$$

NOTES

The procedure of calculating the terms in equations (*i*) and (*ii*) above is presented in Table 1.11.

Table 1.11 *Calculation of Terms of the Linear Equations*

(Figures in million)

Year	*Population (X)*	*Sugar consumed (Y)*	X^2	*XY*
2000	10	40	100	400
2001	12	50	144	600
2002	15	60	225	900
2003	20	70	400	1400
2004	25	80	625	2000
2005	30	90	900	2700
2006	40	100	1600	4000
$\Sigma n = 7$	$\Sigma X_i = 152$	$\Sigma Y_i = 490$	$\Sigma X_i^2 = 3994$	$\Sigma X_i Y_i = 12000$

By substituting the values from Table 1.11 into equations (*i*) and (*ii*), we get

$$490 = 7a + 152b \quad \ldots (iii)$$

$$12{,}000 = 152a + 3994b \quad \ldots (iv)$$

By solving equations (*iii*) and (*iv*), we get

$$a = 27.44$$

and $$b = 1.96$$

By substituting values for *a* and *b* in Eq. (1.32), we get the estimated regression equation as

$$Y = 27.44 + 1.96\,X \quad \ldots (1.33)$$

Given the regression Equation (1.33), the demand for sugar for 2007 can be easily projected if population for 2007 is known. Supposing population for 2007 is projected to be 70 million, the demand for sugar in 2007 may be estimated as $Y = 27.44 + 1.96(70) = 137$ million tonnes.

The simple regression technique is based on the assumptions (*i*) that independent variable will continue to grow at its past growth rate, and (*ii*) that the relationship between the dependent and independent variables will continue to remain the same in the future as in the past. (For further details and on the reliability of estimates consult a standard book on Statistics.)

(ii) **Multi-variate regression:** The multi-variate regression equation is used where demand for a commodity is deemed to be the function of many variables or in cases in which the number of explanatory variables is greater than one. The procedure of multiple regression analysis may be briefly described here. The first step in multiple regression analysis is to specify the variables that are supposed to explain the variations in demand for the product under reference. The explanatory variables are generally chosen form the determinants of demand, viz., price of the product, price of its substitute, consumers' income and their taste and preference. For estimating the demand for durable consumer goods, (e.g., TV sets,

refrigerators and houses), the other explanatory variables which are considered are availability of credit and rate of interest. For estimating demand for capital goods (e.g., machinery and equipment), the relevant variables are additional corporate investment, rate of depreciation, cost of capital goods, cost of other inputs (e.g., labour and raw materials), market rate of interest, etc.

Once the explanatory or independent variables are specified, the second step is to collect time-series data on the independent variables. After necessary data is collected, the next step is to specify the form of equation which can appropriately describe the nature and extent of relationship between the dependent and independent variables. The **final** step is to estimate the parameters in the chosen equations with the help of statistical techniques. The multi-variate equation cannot be easily estimated manually. They have to be estimated with the help of computer.

Specifying the form of equation

The reliability of the demand forecast depends to a large extent on the form of equation and the degree of consistency of the explanatory variables in the estimated demand function. The greater the degree of consistency, the higher the reliability of the estimated demand and vice versa. Adequate precaution should, therefore, be taken in specifying the equation to be estimated. Some common forms of multi-variate demand functions are given below.

- **Linear function**

Where the relationship between demand and its determinants is given by a straight line, the most common form of equation for estimating demand is

$$Q_x = a - bP_x + cY + dP_y + jA \quad \ldots (1.34)$$

where Q_x = quantity demanded of commodity X; P_x = price of commodity X; Y = consumers' income, P_y = price of the substitute; A = advertisement expenditure; a is a constant (the intercept), and b, c, d and j are the parameters expressing the relationship between demand and P_x, Y, P_y and A, respectively.

In a linear demand function, quantity demanded is assumed to change at a constant rate with a change in independent variables P_x, Y, P_y and A. The parameters (regression coefficients) are estimated by using the least square method. After parameters are estimated, the demand can be easily forecast if data on independent variables for the reference period is available. Suppose, the estimated equation for sugar takes the following form:

$$Q_s = 50 - 0.75P_s + 0.1Y + 1.25\,P_y + 0.05A \quad \ldots (1.35)$$

The numerical values in this equation express the quantitative relationship between demand for sugar and the variables with which they are associated. More precisely, regression coefficients give the change in demand for sugar as a result of unit change in the explanatory variables. For instance, it reveals that a change of one rupee in the sugar price results in a 0.75 unit (say, tonne) change in sugar demand and a change of one rupee in income leads to a 0.1 unit (tonne) change in sugar demand, and so on.

• **Power function**

It may be noted that in linear equation (1.34) the marginal effect of independent variables on demand is assumed to be constant and independent of change in other variables. For example, it assumes that the marginal effect of change in price is independent of change in income or other independent variables, and so on. But there may be cases in which it is intuitively or theoretically found that the marginal effect of the independent variables on demand is neither constant nor independent of the value of all other variables included in the demand function. For example, the effect of rise in sugar price may be neutralized by a rise in consumers income. In such cases, a multiplicative form of equation which is considered to be 'the most logical form of demand function' is used for estimating demand for a product. The multiplicative form of demand function or power function is given as:

$$Q_x = a P_x^{-b} Y^c P_y^d A^j \quad \ldots (1.36)$$

The algebraic form of multiplicative demand function can be transformed into a log-linear form for convenience in estimation, as given below.

$$\log Q_x = \log a - b \log P_x + c \log Y + d \log P_y + j \log A \quad \ldots (1.37)$$

The log-linear demand function can be estimated by the least square regression technique. The estimated function yields the intercept a and the values of the regression coefficients. After regression coefficients are estimated and data on the independent variables for the years to come are obtained, forecasting demand becomes an easy task.

Reliability of estimates

As mentioned earlier, statistical methods are scientific, devoid of subjectivity, and they yield fairly reliable estimates. But the reliability of forecast depends also on a number of other factors. A very important factor in this regard is the choice of the right kind of variables and data. Only those independent variables which have a causal relationship between the dependent and independent variables should be included in the demand function. The relationship between the dependent and independent variables should be clearly defined. Besides, the reliability of estimates also depends on the form of demand function used. The forecaster should, therefore, bear in mind that there is no hard and fast rule and an *a priori* basis of determining the most appropriate form of demand function. The demand function to be estimated is generally determined by testing different forms of functions. Whether a particular form of functions is a good fit is judged by the coefficient of determination, i.e., the value of R^2. The value of R^2 gives the proportion of the total variation in the dependent variable explained by the variation in the independent variables. The higher the value of R^2, the greater the explanatory power of the independent variables. Another test is the expected sign of coefficients of independent variables. What is more important, therefore, is to carefully ascertain the theoretical relationship between the dependent and the independent variables. Some case studies based on multiple regression analysis will be presented at the end of this chapter.

(b) **Simultaneous equation methods:** The regression technique of demand forecasting consists of a single equation. In contrast, the simultaneous equations model of forecasting involves estimating several simultaneous equations. These equations are, generally, behavioural equations, mathematical identities and market-clearing equations. Furthermore, regression technique assumes one-way causation, i.e., only the independent variables cause variations in the dependent variable, not vice versa. In simple words, regression technique assumes that a dependent variable affects in no way the independent variables. For example, in demand function $D = a - bP$ used in the regression method, it is assumed that price affects demand, but demand does not affect price. This is an unrealistic assumption. On the contrary, forecasting through econometric models of simultaneous equations enables the forecaster to take into account the simultaneous interaction between dependent and independent variables.

The *simultaneous equations method* is a complete and systematic approach to forecasting. This technique uses sophisticated mathematical and statistical tools which are beyond the scope of this book. Let us discuss basic features of this method of forecasting.

The ***first step*** in this technique is to develop a complete model and specify the behavioural assumptions regarding the variables included in the model. The variables that are included in the model are called (*i*) endogenous variables, and (*ii*) exogenous variables.

(i) **Endogenous variables:** The variables that are determined within the model are called endogenous variables. Endogenous variables are included in the model as dependent variables, i.e., the variables that are to be explained by the model. These are also called 'controlled' variables. *It is important to note that the number of equations included in the model must equal the number of endogenous variables.*

(ii) **Exogenous variables:** Exogenous variables are those that are determined outside the model. Exogenous variables are inputs of the model. Whether a variable is treated as endogenous or exogenous depends on the purpose of the model. The examples of exogenous variables are 'money supply; 'tax rates', 'government spending', 'time', and 'weather', etc. The exogenous variables are also known as 'uncontrolled' variables.

The ***second step*** in this method is to collect the necessary data on both endogenous and exogenous variables. More often than not, data is not available in the required form. Sometimes data is not available at all. In such cases, data has to be adjusted or corrected to suit the model and, in some cases, data has even to be generated from the available primary or secondary sources.

After the model is developed and necessary data are collected, the next step is to estimate the model through some appropriate method. Generally, a

two-stage least square method is used to predict the values of exogenous variables.

Finally, the model is solved for each endogenous variable in terms of exogenous variables. Then by plugging the values of exogenous variables into the equations, the objective value is calculated and prediction is made.

This method is theoretically superior to the regression method. The main advantage of this method is that it is capable of capturing the influence of interdependence of the variables. But, its limitations are similar to those of the regression method. The use of this method is sometimes hampered by non-availability of adequate data.

For example, consider a simple macroeconomic model, given below:

$$Y_t = C_t + I_t + G_t + X_t \quad \ldots (1.38)$$

where

Y_t = Gross national product,

C_t = Total consumption expenditure,

I_t = Gross private investment,

G_t = Government expenditure,

X_t = Net exports ($X - M$) where M = imports

and subscript t represents a given time unit.

Equation (1.38) is an identity, which may be explained with a system of simultaneous equations. Suppose in Eq. (1.38)

$$C_t = a + bY_t \quad \ldots (1.39)$$

$$I_t = 20 \quad \ldots (1.40)$$

$$G_t = 10 \quad \ldots (1.41)$$

$$X = 5 \quad \ldots (1.42)$$

In the above system of equations, Y_t and C_t are endogenous variables and I_t, G_t and X_t are exogenous variables. Eq. (1.39) is a regression equation that has to be estimated. Equations (1.40), (1.41) and (1.42) show the values of exogenous variables determined outside the model.

Suppose we want to predict the value of Y_t and C_t simultaneously. Suppose also that when we estimate equation (1.39), we get

$$C_t = 100 + 0.75\ Y_t \quad \ldots (1.43)$$

Now, using this equation system, we may determine the value of Y_t as

$$Y_t = C_t + 20 + 10 + 5 = C_t + 35$$

Since $C_t = 100 + 0.75\ Y_t$, by substitution, we get

$$Y_t = 100 + 0.75\ Y_t + 35$$

then

$$Y_t - 0.75\ Y_t = 100 + 35$$

$$0.25\ Y_t = 135$$

and

$$Y_t = 135/0.25 = 540$$

We may now easily calculate the value of C_t (using $Y_t = 540$). Since

$$C_t = 100 + 0.75\ Y_t$$
$$= 100 + 0.75\ (540)$$
$$= 505$$

Thus, the predicted values are

$$Y_t = 540$$

and $C_t = 505$

Thus, $Y_t = 505 + 20 + 10 + 5$

$$= 540$$

It is important to note here that the example of the econometric model given above is an extremely simplified model. The econometric models used in actual practice are generally very complex. They include scores of simultaneous equations.

Check Your Progress

10. What are the forms consumer survey method?
11. What is the delphi method of demand forecasting?
12. Name the 'classical method' of business forecasting.

1.6 ANSWERS TO 'CHECK YOUR PROGRESS'

1. Capital budgeting involves various methods such as net present value (NPV), internal rate of return (IRR), equivalent annuity, profitability index, and modified internal rate of return (MIRR).
2. The theory of the firm explains the relationship between inputs and output.
3. The time gap between investment and return is called time lag.
4. The demand curve shows an inverse relationship between price and quantity demanded. This inverse relationship between price and quantity demanded makes the demand curve slope downward to the right.
5. Examples of complement goods: petrol is a complement of motor vehicles, butter and jam are complements of bread, milk and sugar are complements of tea and coffee.
6. Purchases made on account of demonstration effect have a positive effect on demand.
7. The four kinds of demand elasticity are price elasticity, cross elasticity, income elasticity and price expectation elasticity of demand.
8. The elasticity measured on a finite point of a demand curve is called point elasticity.
9. The cross elasticity is the measure of responsiveness of demand for a commodity to the changes in the prices of its substitutes and complementary goods.

10. The consumer survey method of demand forecasting involves direct interview of the potential consumers. It may be in the form of:
    (a) Complete enumeration
    (b) Sample survey
    (c) End-use method
11. The Delphi method of demand forecasting is an extension of the simple expert opinion poll method. This method is used to consolidate the divergent expert opinions and to arrive at a compromise estimate of future demand.
12. The trend projection method is a 'classical method' of business forecasting.

1.7 SUMMARY

- Managerial economics is the study of economic theories, concepts, logic and tools of analysis that are used in analysing business conditions with the objective of finding an appropriate solution to business problems.
- The areas of business issues to which economic theories can be directly applied may be broadly divided into two categories: (a) operational or internal issues, and (b) environment or external issues.
- The time perspective refers to the duration of time period extending from the relevant past and foreseeable future taken in view while taking a business decision.
- Opportunity cost may be defined as income expected from the second-best use of the resource which is sacrificed for the best use of the resource.
- Externalities refer to the costs that people have to bear without consuming goods and services produced by the firms and benefits which people reap without paying for them.
- Trade-off refers to a situation in which two opposite variables moving in the same or opposite direction are in balance.
- The term 'demand' refers to the quantity demanded of a commodity per unit of time at a given price
- Giffen goods do not refer to any specific commodity. They may be any inferior commodities much cheaper than their superior substitutes, consumed mostly by poor households as essential consumer goods.
- The market demand for a product is determined by a number of factors, viz. price of the product, price and availability of the substitutes, consumer's income, utility derived from the commodity, 'past levels of demand, past levels of income, government policy, etc.
- Consumers' taste and preferences play an important role in determining the demand for a product.
- When a commodity becomes a thing of common use, some people, mostly the rich, decrease or give up the consumption of such goods. This is known

as the 'snob effect'. It has a negative effect on the demand for the related goods.

- A demand function may include a single independent variable (price) or many other independent variables, depending on the purpose of analysis.
- A demand function is said to be linear when the slope of the demand curve remains constant throughout its length.
- A demand function is said to be non-linear or curvilinear when the slope of a demand curve changes all along the demand curve.
- The laws of demand, supply and market equilibrium state how demand for and supply of a product respond to change in its price and other determinants.
- The point elasticity of demand is defined as the proportionate change in quantity demanded in response to a very small proportionate change in price.
- Arc elasticity is a measure of the average of responsiveness of the quantity demanded to a substantial change in the price.
- The main determinants of the price elasticity of demand are Availability of substitutes, nature of commodity, proportion of income spent on a commodity, time factor, range of alternative uses of a commodity, and proportion of market supplied.
- The cross elasticity is the measure of the responsiveness of demand for a commodity to the changes in the price of its substitutes and complementary goods.
- The responsiveness of demand to the change in consumer's income is known as income elasticity of demand.
- From a firm's point of view, the knowledge of demand theory alone is not enough for planning and scheduling production of goods and services. The knowledge of future demand for its product, at least an approximate estimate of future demand is an essential aspect of many business decisions.
- Demand forecasting is predicting future demand for a product. There are various methods of demand forecasting and the choice of the method depends on such factors as purpose and perspective of demand forecasting, experience and expertise of the forecaster, availability of required data and the quality and consistency of both primary and secondary data.
- Demand forecasting methods are classified either as survey methods or statistical methods.
- Survey methods are classified as consumer survey and opinion poll methods. The consumer survey methods are categorized as complete enumeration, end-use method and sample survey. The opinion poll methods are divided into expert-opinion survey method and market studies and experiments.
- The statistical methods are divided as trend projection methods, barometric and econometric methods. Trend projection methods are further classified

as graphical method, least square method and Box-Jenkins method. The barometric methods are classified as lead lag indicators and diffusion indices. The econometric methods are classified as regression method and simultaneous equation methods.

1.8 KEY TERMS

- **Demand:** It refers to the quantity demand of a commodity per unit of time at a given price.
- **Law of demand:** It is the quantity of product demanded per unit of time increases when its price falls, and decreases when its prices increases, other factors remaining constant.
- **Elasticity of demand:** It refers to the responsiveness of demand for a good to the change in its determinants is called the elasticity of demand.
- **Demand Forecasting:** It involves predicting future demand for a product.

1.9 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. What is the role of time perspective in business decisions?
2. What are the determinants of market demand for a commodity?
3. Name the three major components of incremental cost.
4. How is the income effect negative in the case of inferior goods?
5. Differentiate between bivariate regression method and multivariate method.
6. What are the roles and responsibilities of a managerial economist?

Long-Answer Questions

1. Why does a demand curve slope downward to the right?
2. Analyse the major macroeconomic or environmental issues which figure in business decision-making.
3. Explain the uses of elasticity.
4. Discuss various consumer survey methods along with their advantages and disadvantages.

1.10 FURTHER READING

Dwivedi, D.N. 2015. *Managerial Economics,* Eighth Edition. New Delhi: Vikas Publishing.

Keat, Paul G. and Philip, K.Y. Young. 2005. *Managerial Economics: Economics Tools for Today's Decision Makers,* Fourth Edition. Singapore: Pearson Education.

Keating, B. and J.H. Wilson. 2005. *Managerial Economics: An Economic Foundation for Business Decisions,* Second Edition. New Delhi: Biztantra.

Peterson, H.C. and W.C. Lewis. 1999. *Managerial Economics*, Fourth Edition. Singapore: Pearson Education.

Salvantore, Dominick. 2001. *Managerial Economics in a Global Economy*, Fourth Edition. Australia: Thomson-South Western.

Thomas, Christopher R. and Maurice S. Charles. 2005. *Managerial Economics: Concepts and Applications*, Eight Edition. New Delhi: Tata McGraw-Hill.

NOTES

UNIT 2 PRODUCTION ANALYSIS AND COST ANALYSIS

NOTES

Structure

2.0 INTRODUCTION

The process of combining various material inputs and immaterial inputs (plans, know-how) to produce a commodity for its consumption is known as production. It is the act of creating an output, a good or service which has value and contributes to the utility of individuals. Production of a commodity depends on the availability of inputs (labour and capital) and technology and the supply of a product depends on its production. In Economics, production refers to the creation or addition of value. It simply transforms the inputs into output. This relationship between input and output is known as production analysis. Production may be at varying levels. The scale of production is affected by the cost of production. All manufacturers know that when a commodity is produced takes on a larger scale, the average cost of its production is low. Therefore, the entrepreneurs want to increase the scale of production of their goods. On the other hand, the cost analysis refers to the measure of the cost-output relationship, i.e., economists are concerned with determining the cost incurred in hiring the inputs and how well these inputs can be rearranged to increase the productivity (output) of the firm. This unit will explain certain production- and cost-related concepts. In addition, it will describe production and cost analysis. Also, it discusses various tools related to production and cost.

NOTES

2.1 OBJECTIVES

After going through this unit, you will be able to:

- Discuss the laws of production in the short and long run
- Explain the types of costs
- Describe the theories of cost

2.2 PRODUCTION ANALYSIS

In economics, the term 'production' means a process by which tangible and intangible inputs or factors of production (land, labour, capital, etc.) are converted or transformed into useful tangible or intangible output. In other words, production means transforming *inputs* (labour, machines and raw materials) into *output.* Production may take the form of transforming tangible inputs into tangible output. This kind of production is called 'manufacturing'. However, production process does not necessarily involve physical conversion of raw materials into tangible goods. In the process of production, an input may be intangible (service) and an output may be intangible too. For example, in the production of legal, medical, social and consultancy services both input and output are intangible; lawyers, doctors, social workers, consultants, hair-dressers, musicians, orchestra players are all engaged in productive activity.

In the economic sense, the *production process* may take a variety of forms. For example, transporting a commodity from one place to another where it can be used is called *production* of a commodity. A coal seller does virtually nothing more than transporting coal from coal mines to the marketplace. Similarly, a fisherman only catches and transports fish to the marketplace. Their activities are productive activities. Transporting men and materials from one place to another is in itself a productive activity; it produces *service. Storing* a commodity for future sale or consumption is production. Wholesaling, retailing, packaging and assembling are all productive activities. These activities are just as good examples of production as manufacturing.

2.2.1 Production-related Concepts

Let's discuss some of the fundamental concepts related to production.

(i) Input and output

An *input* is a good or service used in the process of production. In the words of Baumol, 'An input is simply anything which the firm buys for use in its production or other processes. An 'output' is any commodity which the firm produces or processes for sale.' An *output* is any good or service that comes out of the production process.

The term 'input' needs some more explanation. The production process requires a wide variety of inputs, depending on the nature of the product. However, economists have classified inputs as (*i*) labour, (*ii*) capital, (*iii*) land, (*iv*) entrepreneurship or managerial efficiency, (*v*) raw materials, and (*vi*) time. All

these variables are 'flow' variables in the sense that they are measured per unit of time.

(ii) Fixed and variable inputs

NOTES

For the sake of analytical convenience, inputs are further classified as (*i*) *fixed* inputs, and (*ii*) *variable* inputs. A *fixed input* is one whose quantity remains constant for a certain level of output, e.g., plant, building and machinery. The supply of fixed inputs remains inelastic in the short-run. A *variable input* is defined as one whose quantity changes with change in output. The supply of such inputs (as labour and raw materials) is elastic in the short-run.

(iii) Short-run and long-run

Corresponding to fixed and variable inputs, economists use two other terms, viz. *short-run* and *long-run.* Short-run refers to a period of time in which the supply of certain inputs (e.g., plant, building and machines) is fixed or inelastic. In the short-run, therefore, production of a commodity can be increased by increasing the use of variable inputs, like labour and raw materials.

It is worth noting that short-run does not refer to any fixed time period. While in some industries it may be a matter of weeks or a few months, in some others (e.g., electric and power industry), it may mean three or more years.

Long-run refers to a period of time in which the supply of all the inputs is elastic but not enough to permit a change in technology. That is, in the long-run, the supply of even fixed variables increases. Therefore, in the long-run, production of a commodity can be increased by employing more of both variable and fixed inputs.

The economists use another term, i.e., *very long period*, which refers to a period in which the technology of production is subject to change. In the very long-run, the production function also changes. The technological advances mean that a larger output can be created with a given quantity of inputs.

2.2.2 Issues in Production Theory and Production Function

Production theory seeks to analyse the input and output relations and answers the following theoretical issues:

1. If all the inputs are simultaneously increased (or decreased) at a certain rate, will the output increase (or decrease) in the same proportion? Or if, for example, the amount of each input is doubled (or halved), will the output be doubled (or halved) or will it change in a different proportion?
2. Supposing there is more than one process of producing a commodity, how will output change (or behave) in response to change in factor proportions? Or, how will output change if one input is substituted for another?
3. How can the least-cost combination of inputs be achieved? In other words, how is optimum technique of production chosen?

NOTES

Production functions

The tool of analysis which is used to explain the input–output relationships and give the probable answer to the above queries is *Production Function.* Let us now briefly describe the nature and forms of a production function.

The production function describes the technological relationship between inputs and outputs in physical terms. It specifies the maximum quantity of a commodity that can be produced per unit of time with given quantities of inputs and technology. Besides, the production function represents the technology of a firm, of an industry or of the economy as a whole in a relevant case.

A production function may take the form of a schedule or table, a graphed line or curve, an algebraic equation or a mathematical model. But each of these forms of a production function can be converted into the other forms.

Before we illustrate the various forms of a production function, let us note how a complex production function is simplified and how the number of inputs included in the production function, as independent variables, reduce to a manageable number.

A general empirical form of production function can be expressed as

$Q = f(L, K, LB, M, T, t, e \ldots)$

where Q = quantity, L = labour, K = capital, LB = land/building, M = materials, T = technology, t = time, and e = managerial efficiency.

All these variables enter the actual production function of a firm. The economists have, however, reduced the number of variables used in a production function to only two, viz., *capital* and *labour,* for the sake of convenience and simplicity in the analysis of input–output relations.

The reasons given for ignoring the other inputs are as follows: *Land/building* as an *input* is constant for the economy as a whole, and hence, it does not enter into the aggregate production function. However, land/building is not a constant variable for an individual firm or industry, at least in the long-run. In the case of individual firm and industries, however, land/building is lumped together with 'capital'. In case of 'raw materials', it has been observed that this input 'bears a constant relation to output at all levels of production'. For example, cloth bears a constant relation to the number of garments, similarly for a given size of a house, the quantity of bricks, cement, steel, etc., remains constant, irrespective of the number of houses constructed. This constancy of input–output relations leaves the method of production unaffected. That is why, in most production functions, only two inputs—labour and capital—are included. Technology, time and managerial efficiency are also assumed to be given in the short-run.

We will illustrate the tabular and graphic forms of a production function when we move on to explain the laws of production. Here, let us illustrate the algebraic or mathematical form of a production function, which is most commonly used in production analysis.

To illustrate the algebraic form of production function, let us suppose that a firm employs only two inputs—capital *(K)* and labour *(L)*—in production of a

commodity. As such, the general form of its production function may be algebraically expressed as

$$Q = f(K, L) \quad \ldots (2.1)$$

where Q = the quantity of output produced per unit time, K = capital, and L = labour.

NOTES

The production function (2.1) implies that Q is the maximum quantity of the product that can be produced, given the total volume of capital, K, and the total number of workers, L, employed to produce coal. Increasing production will require increase in K and L. Whether the firm can increase both K and L or only L depends on the time period it takes into account for increasing production, i.e., whether the firm considers the *short-run* or the *long-run.* As noted earlier, *short-run* refers to a period during which supply of certain factors of production (viz. capital and land) is supposed to be inelastic. On the other hand, *long-run* is a period of time during which supply of all the factors of production is assumed to be elastic, though not long enough to permit change in technology.

By definition, supply of capital is *inelastic* in the short-run and *elastic* in the long-run. In the short-run, therefore, the firm can increase its production by increasing only labour, since the supply of capital in the short-run is fixed.

In the long-run, however, the firm can employ more of both capital and labour. Accordingly, the firm would have two types of production functions: (i) *short-run production function;* and (ii) *long-run production function.* The short-run production function, or what may also be termed as *'single variable production function',* can be expressed as

$$Q = f(L) \quad \ldots (2.2)$$

In the long-run production function, both K and L are included and the function takes the form

$$Q = f(K, L) \quad \ldots (2.3)$$

2.2.3 Laws of Production: An Overview

The traditional theory of production states the nature of relationship between inputs and output. It is a universal fact that increasing quantity of the inputs (labour and capital) leads to an increase in the output and *vice versa.* In fact, the theory of production answers the questions: How does output change with the change in the quantity of inputs? Does the output change in proportion to change in inputs or more than proportionately or less than proportionately? Economists have attempted to answer question these under two different conditions, viz.,

(i) short-run production conditions, and

(ii) long-run production conditions.

The short-run production conditions are specified as follows:

(i) In the short-run, labour is the only variable input. Since the supply of labour continues to increase with increase in population, more labour can be used at any time to produce a commodity.

(ii) The supply of capital remains constant in the short-run. The reason is that supply of capital is market determined. Since market conditions are static in the short-run, supply of capital remains constant. Capital supply remaining constant, firms cannot use more and more of capital to increase production.

(iii) Production technology, i.e., labour–capital ratio, for a product is given — it does not change in the short-run.

On the other hand, the long-run production conditions are specified as follows:

NOTES

(i) The supply of both ***labour*** and ***capital*** is variable in the long-run. So more and more of both labour and capital can be used to produce a commodity.

(ii) In the long-run, the technology of production is also variable. However, for theoretical purpose, technology of production is assumed to remain constant.

Give the short-run and long-run production conditions, economists have developed two kinds of theories of production and have formulated two kinds of laws of production.

(i) Short-run laws of production, which are based on short-run production conditions, and

(ii) Long-run laws of production, which are based on long-run production conditions.

The two kinds of laws of production are discussed as follows:

2.2.4 The Short-run Laws of Production or the Laws of Variable Proportions: Production with One Variable Input

The short-run laws of production are also known as the laws of diminishing returns, the laws of variable proportions, and the laws of returns to variable input.

The laws of returns to variable proportions explain the relationship between the variable input and the output in the short term. As noted above, certain factors of production (viz., land and capital equipments such as plant and machinery) are available in short supply in the short-run. Such factors are known as *fixed factors*. On the other hand, the factors that are available in unlimited supply even in the short-run are known as *variable factors*. In short-run, therefore, the firms can employ only a limited or fixed quantity of fixed factors and an unlimited quantity of the variable factor. In other words, firms can employ in the short-run varying quantities of variable inputs against a given quantity of fixed factors. This kind of change in input combinations leads to variation in factor proportions. As mentioned above, the law that brings out the relationship between varying factor proportions and output is known as the law of variable proportions.

The law of diminishing returns states that if more and more units of a variable input are applied to a given quantity of fixed inputs, the total output may initially increase on increasing rate, but beyond a certain level, output increases at a diminishing rate. Precisely, marginal increase in total output eventually decreases when additional units of variables factors are applied to a given quantity of the fixed factors. The main reason behind the operation of this law is the decreasing labour–capital ratio. Given the quantity of fixed factor (capital), with increasing

variable input (labour), capital–labour ratio goes on decreasing. That is, each additional worker has less and less tools and equipments to work with. Consequently, the productivity of the marginal worker eventually decreases. As a result, the total output increases but at a diminishing rate beyond a point.

To illustrate the law of diminishing returns, let us assume that the firm possesses a set of machinery as its capital (K), which is fixed in the short-run, and that it can increase only the number of workers to increase its production. Thus, the short-run production function for the firm will be of the form given in Equation (2.2), i.e.,

$$Q = f(L)$$

Let us assume that the labour–output relationship is given by the following production function

$$Q = -L^3 + 10L^2 + 20L^1 \quad \ldots (2.4)$$

Given the production function (2.4), we may subtitute numerical values for L in the function and workout a series of Q, i.e., the quantity of output that can be produced with different number of workers. For example, if we substitute 5 for L, the production function (2.4) will read as

$$\begin{aligned} Q &= -5^3 + 10 \times 5^2 + 20 \times 5 \\ &= -125 + 250 + 100 \\ &= 225 \end{aligned}$$

A tabular array of output levels associated with different number of workers, from 1 to 10, in our hypothetical example is given in Table 2.1 (Cols. 1 and 2). From the table, we derive the marginal product (MP_L) and the average product (AP_L) schedules, as given in the table (Cols. 3 and 4).

Table 2.1 *Schedule Illustrating Law of Diminishing Returns*

No. of workers (N)	*Total product (TP)(tons)*	*Marginal product (MP_L) physical units*	*Average product (AP_L) (tons)*	*Stage of production (tons)*
1	*2*	*3*	*4*	*5*
1	29	29	29	
2	72	43	36	
3	133	51	41	I
4	176	53	44	
5	225	49	45	
6	264	39	44	
7	287	23	41	II
8	288	1	36	
9	261	(–) 27	29	
10	200	(–) 61	20	III

Note: $MP_L = TP_n - TP_{n-1}$ where n is the number of workers. For example, productivity of 7th worker may be known as $TP_n - TP_{n-1} = 287 - 264 = 39$ and $AP_L = TP/N$.

The labour-output data contained in Table 2.1 is presented graphically in Figure 2.1. The TP_L and MP_L schedules demonstrate the law of diminishing returns. As the curve TP_L indicates, the total output continues to increase at an increasing rate till the employment of the 4th worker—the rate of increase in the total output,

NOTES

i.e., marginal product (MP_L)—is increasing. Beyond the employment of 4th worker, although TP_L continues to increase (until 8th worker), the rate of increase in TP_L (i.e., marginal addition to TP_L) begins to fall and continues to fall, ultimately becoming negative.

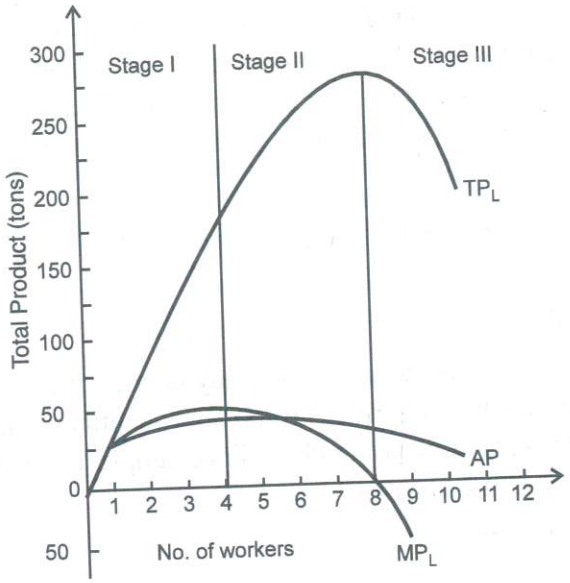


Fig. 2.1 *Total Average and Marginal Products*

Three Stages in Production

Table 2.1 and Figure 2.1 present the three usual stages in the application of the laws of diminishing returns. In Stage I, MP_L continues to increase, making TP_L increase at an increasing rate. In Stage II, MP_L starts falling so that TP_L increases at a decreasing rate until MP becomes negative when TP starts falling.

The reasons that underlie the application of the laws of returns to variable proportions in Stages I and II may be described as follows:

The law of increasing returns operates because of *underutilization of the fixed factor*, i.e., capital. Let us suppose that optimum capital–labour combination is 1:4. It implies that if less than 4 workers are employed, the plant or machine would remain underutilized. When more and more workers are added, utilization of machine increases and also the productivity of additional workers. Another reason for increase in labour productivity is that employment of additional workers lead to advantages of *division of labour,* until optimum capital–labour combination is reached.

Once the optimum capital–labour ratio is reached, employment of additional workers means substitution of capital with labour. But technically, one factor can substitute another only to a limited extent. In other words, there is a limit to which one input can be substituted for another. That is, the elasticity of subtitution between inputs is not infinite. Hence, to replace the same amount of capital and to achieve the labour productivity at the optimum level of capital–labour combination, more and more workers will have to be employed. Naturally, per worker marginal productivity decreases.

NOTES

Assumptions

The application of the law of diminishing returns is subject to the following assumptions:

(a) The state of technology is given—it does not change in the course of application of the law.

(b) Input prices remain unchanged.

(c) The variable factors are homogeneous.

Application of the law of diminishing returns

The law of diminishing returns is an empirical law, frequently observed in the various production activities. This law, however, may not apply universally to all kinds of productive activities since the law is not as perfect as the laws of physics. In some productive activities, it may operate quickly; in some, its operation may be delayed; and in some others, it may not appear at all. The law of diminishing returns has been found to appear in agricultural production more quickly than in industrial production, because in the former, a natural factor (land) plays a predominant role, while in the latter, man-made factors play the major role. Despite the limitations of the law, if increasing units of an input are applied to the fixed factors, the marginal returns to the variable input eventually decrease.

2.2.5 Production with Two Variable Inputs: Isoquant Curve Fundamentals

The long-run laws of production, i.e., input-output relationship under the condition that both *labour and capital* are variable factors and more of these factors can be used to increase production. This is a long-run phenomenon. The *long-run* refers to a period in which the supply of both inputs — labour and capital — is variable. That is, the supply of both labour and capital is subject to rise in the long-run. So the firms can use more of both labour and capital to increase the production of their product. The study of the input-output relationship under the long-run conditions gives the *long-run laws of production*. Since increase in both the inputs — labour and capital — increases the *scale of production*, the long-run laws of production are known also as the laws of returns to scale.

The long-run relationship between inputs and output is generally expressed through a long-run production function, e.g., Cobb-Douglas production function. However, there is a simpler, graphical tool of analysis used to explain and illustrate the *laws of returns to scale*. The tool of analysis is called isoquant – a tool of analysis similar to *indifference curve*.

Isoquant curves

The term 'isoquant' has been derived from *iso* meaning 'equal' and *quant* meaning 'quantity'. The 'isoquant curve' is therefore also known as 'equal product curve' or 'production indifference curve'. An isoquant curve is a locus of points representing the various combinations of two inputs—capital and labour—yielding the same output. This hypothetical proposition is based on the assumption that capital and labour can be substituted for one another but at a diminishing rate.

NOTES

To draw an isoquant curve, the following assumptions are made:

(i) that a producer uses only two inputs, labour (L) and capital (K), to produce a commodity X;

(ii) that L and K can be substituted for one another at a diminishing rate;

(iii) that the technology of production is given for the period under reference; and

(iv) that the production function of the firm is continuous, i.e., labour and capital are perfectly divisible and substitutable.

Given the assumptions, the production function takes the following form:

$Q = f(L, K)$

The production function being continuous, it cannot be conveniently presented in a tabular form. It can however be conveniently presented graphically. In Figure 2.2, the vertical and horizontal axes measure K and L, respectively. Given the production function and the technology, it is technically possible to form a number of labour–capital combinations that can produce a given quantity of a commodity, X. When such combinations are plotted, it gives the isoquant. Consider, for example, the isoquant labelled $Q_1 = 100$. Each point on this curve shows a different capital-labour combination that can produce 100 units of X per unit of time. For example, points A and B represent two alternative combinations of capital and labour capable of producing 100 units of X. That is, OK_2 units of capital and OL_1 units of labour (as indicated by point A) or OK_1 units of capital and OL_2 units of labour (as indicated by point B) can be alternatively used to produce 100 units of X per unit of time. All other capital–labour combinations represented by this curve can produce the same quantity of X.

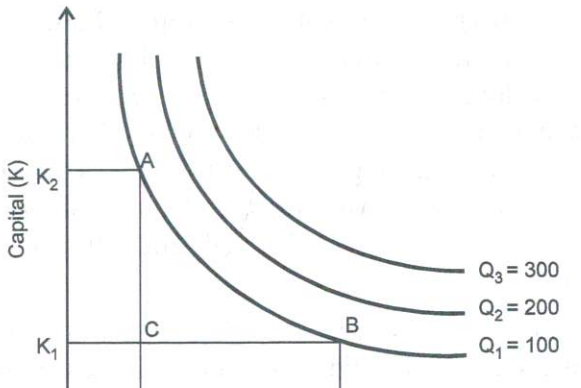


Fig. 2.2 Isoquants

Note that movement along an isoquant means substitution of one factor for another. For example, movement from point A to B means that L_1L_2 ($= CB$) units of labour is substituted for K_1K_2 ($= CA$) units of capital. It also implies that K_1K_2 units of capital can produce as much as L_1L_2 units of labour can. The rate at which one factor can substitute another is called *marginal rate of technical substitution*. This concept is of great significance in production analysis.

Marginal rate of technical substitution

The marginal rate of technical substitution (*MRTS*) is the rate at which one input can be substituted for another without changing the level of output. The rate at which one input can be substituted for another, holding the output constant, is given by the slope of the isoquant. Since the slope of an isoquant moving down the isoquant is given by – D*K*/D*L*,

$$MRTS = -\frac{\Delta K}{\Delta L} = \text{Slope of the isoquant}$$

The condition that the total output should remain constant implies that marginal product of *K* (i.e., MP_K) must equal marginal product of *L* (i.e., MP_L). That is,

$$(-DK \times MP_K) = (DL \times MP_L) \quad \ldots (2.5)$$

By rearranging Equation 2.5, we get

$$-\frac{\Delta K}{\Delta L} = \frac{MP_L}{MP_K}$$

Since $-\frac{\Delta K}{\Delta L} = MRTS$,

$$\frac{MP_L}{MP_K} = MRTS \quad \ldots (2.6)$$

Thus, *MRTS* of *L* for *K* is the ratio of the marginal product of labour (MP_L) to the marginal product of capital (MP_K).

To illustrate the *MRTS* numerically, let us suppose that a given production function may be presented in a tabular form, as given in Table 2.2. The table presents 5 alternative combinations of *K* and *L* that can be used to produce a *given quantity*, say 10 units, of a commodity.

Table 2.2 *Alternative Methods of Producing 10 Units of a Commodity*

K	*L*	Δ*K*	Δ*L*	*MRTS* = D*K*/D*L*
10	2			
		– 2	2	– 1.0
8	4			
		– 3	6	– 0.5
5	10			
		– 4	10	– 0.4
1	20			

Note that as we move down the table, the *MRTS* declines. This is an important factor in determining the shape of the isoquant. The downward movement on an isoquant indicates substitution of labour for capital. The amount of capital *decreases* while the number of workers *increases,* so that output remains constant. The units of labour which can substitute one unit of capital go on increasing. As a result, the *MRTS* (= – D*K*/D*L*) decreases. The reason is that both the factors are subject to the law of diminishing marginal return. As the number of labour increases, its marginal product decreases. On the other hand, with the decrease in the quantity of capital, its marginal productivity increases. Labour whose marginal productivity is decreasing substitutes capital whose marginal productivity is increasing. Therefore, to substitute each subsequent unit of capital, more and more units of labour are required to maintain the production at the same level. That is why the *MRTS* decreases.

NOTES

Properties of isoquant curves

Like indifference curves, isoquants have the following properties:

(a) **Isoquants have a negative slope:** An isoquant has a negative slope in the economic region or in the relevant range. Economic region is the region on the isoquant plane in which substitution between inputs is technically possible. The negative slope of the isoquant implies that if one of the inputs is reduced, the other input has to be so increased that the total output remains unaffected. For example, movement from A to B on the isoquant $Q_1 = 100$ (Figure 2.2) means that if K_1K_2 of units of capital are removed from the production process, L_1L_2 units of labour have to be employed to maintain the same level of output.

(b) **Isoquants are convex to the origin:** Convexity of isoquants implies not only the substitution of one factor for another but also diminishing marginal rate of technical substitution (*MRTS*). As mentioned above, the *MRTS* is the rate at which marginal unit of an input can be substituted for the other input so that the level of output remains the same. The *MRTS* decreases because, by assumption, no factor is a perfect substitute for another. For this reason, more and more units of an input are needed to replace each successive unit of other input (see Table 2.2).

(c) **Isoquants cannot intersect or be tangent to each other:** The intersection or tangency of two isoquants implies that a certain quantity of a commodity can be produced with a smaller input combination as well as with a larger input combination. This is not consistent with the theory of production so long as marginal productivity of an input is greater than zero. In Figure 2.3, two isoquants intersect each other at point M. Consider two other points—point J on isoquant marked 100 and point N on isoquant marked 200. One can easily infer that the quantity that can be produced with the combination of K and L at point M also can be produced with factor combination at J and N. On the isoquant 100, factor combinations at points M and J are equal in terms of productivity. On the isoquant 200, however factor combination at points M and N are equal in terms of productivity. Since point M is common to both the isoquants, it follows that input combinations at M and N are equal in terms of productivity. It means that in terms of output,

$$OL_2 + JL_2 = OL_2 + NL_2$$

Since OL_2 is common to both the sides, it means, that

$$JL_2 = NL_2$$

But it is not true because, as shown in Figure 2.3,

$$JL_2 < NL_2.$$

Intersection between isoquants would mean that $JL_2 = NL_2$, which is not consistent with theory of production. That is why isoquants will not intersect or be tangent to each other, unless marginal productivity of an input is zero.

NOTES

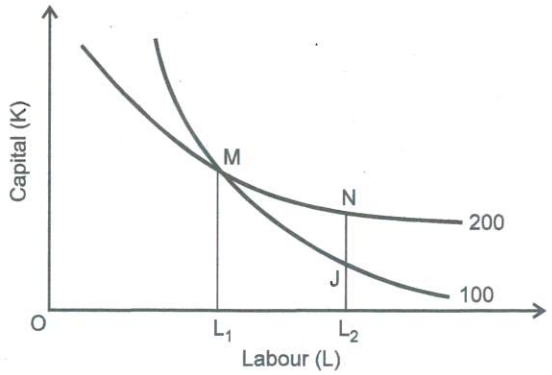


Fig. 2.3 Intersecting Isoquants

Isoquant map and economic region

One way to present a production function on a two-dimensional plane is to use its *isoquant map*. An *isoquant map* is a set of isoquants presented on a two-dimensional plane as shown by isoquants Q_1, Q_2, Q_3 and Q_4 in Figure 2.4. Each isoquant shows various combinations of two inputs that can be used to produce a given level of output. An upper isoquant is formed by a greater quantity of one or both of the inputs than that indicated by the lower isoquants. For example, isoquant Q_2 indicates a greater input combination than that shown by isoquant Q_1, and so on.

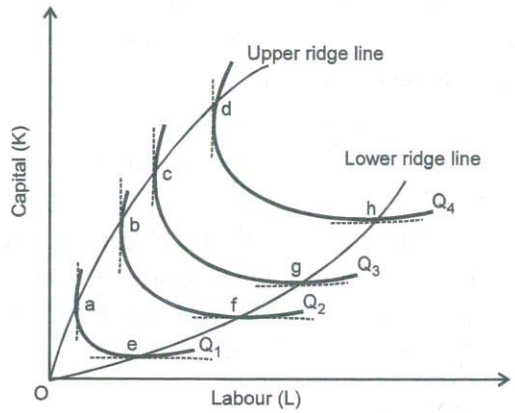


Fig. 2.4 Isoquant Map Economic Region

In addition, since upper isoquant indicates a larger input combination than the lower ones, each successive upper isoquant indicates a higher level of output than the lower ones. For example, if isoquant Q_1 represents an output equal to 100 units, isoquant Q_2 represents an output greater than 100 units. As one of the properties of isoquants, no two isoquants can intersect or be tangent to one another.

Conventional economic theory concentrates on the economically efficient range of output, i.e., till marginal productivity (*MP*) of an input is positive. Zero *MP* of an input is the limit of efficient range of output. This phenomenon can be presented through isoquants.

It is noteworthy that the whole isoquant map or production plane is neither technically efficient, nor is every point on isoquant technically efficient. The

NOTES

reason is that, on a convex isoquant, the *MRTS* decreases along the isoquant. The limit to which the *MRTS* can decrease is zero. The zero *MRTS* implies that there is a limit to which one input can substitute another. It also determines the minimum quantity of an input which must be used to produce a given output. Beyond this point, an additional employment of one input will necessitate employing additional units of the other input. Such a point on an isoquant may be obtained by drawing a tangent to the isoquant and parallel to the vertical and horizontal axes, as shown by dashed lines in Figure 2.4. By joining the resulting points *a*, *b*, *c* and *d*, we get a line called the *upper ridge line, Od.* Similarly, by joining the points *e, f, g* and *h*, we get the *lower ridge line*, *Oh. The ridge lines are locus of points on the isoquants where the marginal products of the inputs are equal to zero.* The upper ridge line implies that *MP* of capital is zero along the line, *Od.* The *lower ridge line* implies that *MP* of labour is zero along the line, *Oh.*

The area between the two ridge lines, *Od* and *Oh*, is called *economic region* or 'technically efficient region' of production. Any production technique, i.e., capital–labour combination, within the economic region is technically efficient to produce a given output, and any production technique outside this region is technically inefficient since it requires more of both inputs to produce the same quantity.

For example, suppose that the quantity represented by isoquant Q_2 is to be produced. We have two points *b* and *f*, on the isoquant Q_2, which fall on the ridge lines. Consider first the point *b*, i.e., the point of intersection between the isoquant Q_2 and the upper ridge line. Point *b* indicates that a minimum of capital is required to produce Q_2. Any smaller amount of capital, given the labour input at point *b*, would be insufficient to produce Q_2. Beyond point *b*, producing Q_2 would require more of both inputs, capital and labour, which is technically inefficient. It would mean uneconomic use of resources. It may be inferred from the above (i) that at point *b*, *MP* of capital is zero, and (ii) that further substitution of capital for labour is technically inefficient.

Optimum Combination of Inputs

A profit maximizing firm seeks to minimize its cost for a given output or to maximize the output from a given total costs. A necessary condition for accounting this objective is to find *the optimum combination of inputs* or *what is also called least-cost combination of inputs*, given the input prices. Given the technology, a given output can be produced with a large number of different input combinations. But only one of such input combinations conforms to the optimality rule of input combination. The problem here is how to find the optimum or the least-cost combination of inputs for a given output.

To look at this issue more closely, let us consider the information contained in Figure 2.5. As isoquant $I_1 = 100$ shows, 100 units of a commodity, *X*, can be produced with all the combinations of *K* and *L* that can be formed on the isoquant

I_1. For example, points A, B and C, represent three different combinations of K and L: (i) $OK_3 + OL_1$; (ii) $OK_2 + OL_2$; and (iii) $OK_1 + OL_3$. All these combinations of K and L can produce 100 units of X. Similarly, many other combinations of capital and labour can be formed on the isoquant I_1 that can produce 100 units of commodity X. But, given the input prices—interest and wages—the total cost of production varies from point to point on an isoquant and only one of the combinations gives the minimum cost, not necessarily any of A, B and C. The problem now is how to find an input combination that results in the minimum cost of production. The least-cost-input combination can be determined by combining firm's production and cost functions. We know that firm's production function is represented by *isoquants*. What we need here is to devise firm's cost function and draw the *isocosts*.

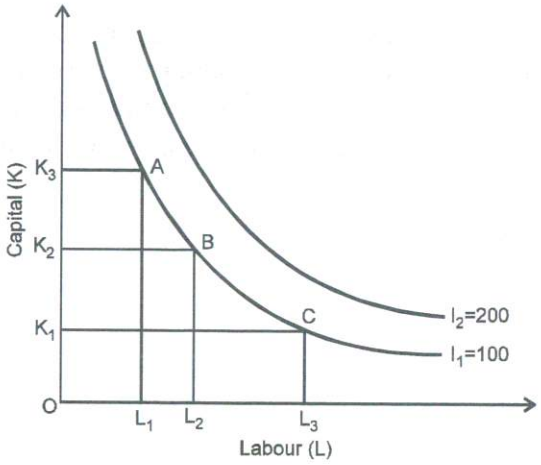


***Fig. 2.5** Input Combination*

Budgetary constraints and isocost lines

To construct the cost function, let us assume that a firm plans to incur a total cost, C, on both K and L and that P_k and P_l are the unit costs of K and L, respectively. Given these conditions, firm's cost function may be expressed as

$$C = K \cdot P_k + L \cdot P_l \qquad \ldots (2.7)$$

From Equation (3.7), the quantity of capital, K, that can be purchased out of the total cost, C, can be easily obtained as shown below:

$$K = \frac{C}{P_k} - \frac{P_l}{P_k} L \qquad \ldots (2.8)$$

Similary, $$L = \frac{C}{P_L} - \frac{P_k}{P_L} K \qquad \ldots (2.9)$$

Equation 2.8 yields a line like K_1L_1 in Figure 2.6 which gives the alternative combinations of K and L that can be purchased given the total cost C. This line is known as *isocost line*, also known as *budget line* or *budget constraint line*.

NOTES

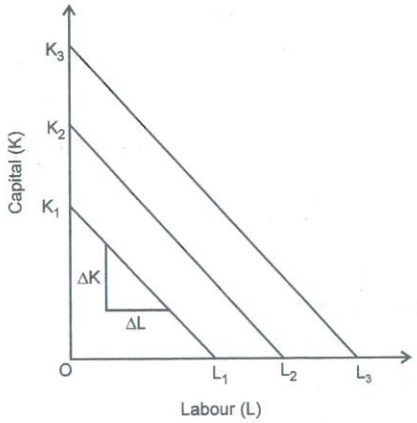


Fig. 2.6 Isocosts

Consider a family of three isocosts in Figure 2.6. They are drawn on the assumption that a firm has the option of spending its total cost, C, either on K or on L, or on both. If the resources are spent on K, or alternatively on L, the firms can buy either OK_1 units of K or OL_1 units of L, as shown below:

$$OK_1 = \frac{C}{P_k} - \frac{P_l}{P_k} L$$

where $L = 0$

and $$OL_1 = \frac{C}{P_1} - \frac{P_k}{P_1} K$$

where $K = 0$

The line connecting points K_1 and L_1 is termed isocost line. It shows the whole range of combinations of K and L that can be bought, given the total cost and factor prices. If the firm decides to spend more than C, its isocost line will shift upwards to K_2L_2 or to K_3L_3, and so on. The isocosts K_2L_2 and K_3L_3 show the upward movement of isocosts when the firm spends more than C, given the labour and capital prices.

It is important to note here that the slope of the isocosts i.e., DK/DL, gives the marginal rate of exchange (*MRE*) between K and L in value terms. Since factor prices are constant, marginal rate of exchange remains constant all along the line.

Criteria optimum combination of inputs

Having introduced the isocosts, now combine isoquants and isocosts to show the optimal input combination or least-cost-combination of inputs. The optimal input combination is determined by the 'least-cost criteria'.

(i) The least-cost criteria

The general criteria for least-cost input combination can be expressed in both physical and value terms. Given the two inputs K (capital) and L (labour) the criterion in *physical terms* is given by

$$\frac{-\Delta K}{\Delta L} = \frac{MP_l}{MP_k} \qquad \text{... (2.10)}$$

NOTES

where DK/DL is the exchange ratio between K and L, and MP_l/MP_k is the ratio of marginal productivities of L and K. This is an input combination at which factor exchange ratio (given factor prices) equals the marginal productivity ratios. This rule gives the *least-cost input combination.*

In terms of money value, the criterion for the least-cost or optimal input combination may be expressed as

$$\frac{MP_l}{MP_k} = \frac{P_l}{P_k}$$

or $$\frac{MP_l}{P_l} = \frac{MP_k}{P_k}$$

where P_l and P_k stand for prices of labour and capital, respectively.

In Equation (2.10), – DK/DL = slope of the isocost, and MP_k/MP_l = slope of the isoquants. It means that the least-cost combination exists at a point where isoquant is tangent to the isocost. The least-cost combination of K and L is shown in Figure 2.7. The isoquant $Q_2 = 200$ is tangent to isocost K_2L_2 at point P. At this point, the combination of K and L is OM of K plus ON of L. This combination of K and L is optimal since it satisfies the least-cost criterion, i.e.,

$$\frac{MP_k}{MP_l} = -\frac{\Delta L}{\Delta K}$$

Thus, the *necessary condition* of minimum cost is satisfied at the point of tangency between the isoquant and the isocost. At the points of tangency, the marginal exchange ratio of inputs is equal to the ratio of their marginal productivity, i.e., at this point, $P_l/P_k = MP_l/MP_k$.

There is, however, another condition, called *second order* condition, that must be satisfied simultaneously. Note that the least—cost condition is also satisfied on points A and D, the points of interaction between isoquant $Q_1 = 100$ and isocost K_2L_2. Note that points A, D and P are on the same isocost, but on different isoquants. While point P is associated with an output $Q_2 = 200$, points A and D being on a lower isoquant are associated with output of 100 units. It means that, given the total cost, a firm can produce 100 units as well as 200 units. Therefore, if the firm chooses input combinations at point A or D, it will produce only 100 units, whereas it could produce 200 units at the same cost at point P. This means that, only point P satisfies the second order condition of cost minimization.

Physical criterion can be converted into value terms by multiplying the factor exchange ratios with factor prices and *MRTS* with product price (P). In fact, factor price ratios are the same as the reciprocal of factor ratios, i.e., P_l/P_k = DK/DL, and

$$MRTS = \frac{MP_l}{MP_k} = \frac{(MP_l)P}{(MP_k)P} = \frac{MRP_l}{MRP_k} \quad \ldots (2.11)$$

NOTES

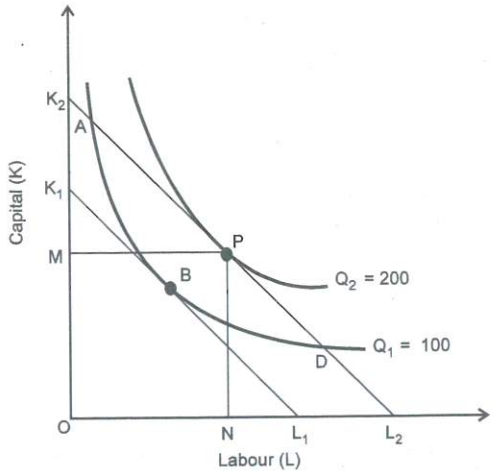


***Fig. 2.7** Least Cost Combination of Inputs*

where *MRP* = marginal revenue productivity of the factor and *P* = product price. Thus, the least-cost criterion given in Equation (2.11) can be converted in terms of input and output price as

$$\frac{P_l}{P_k} = \frac{MRP_l}{MRP_k}$$

or $$\frac{MRP_l}{P_l} = \frac{MRP_k}{P_k} \quad \text{... (2.12)}$$

In may be inferred from Equation (2.12) that the *least-cost or optimal input combination requires that the MRP ratios of inputs should be equal to their price ratios.*

(ii) Output maximization for a given cost

In an alternative situation, a firm faced with a *resource constraint* may seek to maximize the output. This is corollary of cost minimization hypothesis. Maximization of output at a given cost is illustrated in Figure 2.8. In the figure, the isocost of the firm is given as shown by the line *KL* in Figure 2.8. The firm has to maximize the output subject to the cost constraint.

The *first order condition* for maximizing output is that the slope of the isocost must be equal to the slope of the isoquant. That is, output is maximized where

$$\frac{P_l}{P_k} = \frac{MP_l}{MP_k}$$

The *second order condition* requires that the first order condition must be satisfied at the highest possible isocost. Both these conditions are fulfilled at point *P* on the isoquant Q_2. Thus, Q_2 is the maximum output attainable under the given cost condition. Although the *first order* condition is also fulfilled at points *Q* and *R* on isoquant Q_1, the output is not maximum at these points because the second order condition is not fulfilled. The points *Q* and *R* being on a lower isoquant (Q_1) denote an output lower than Q_2.

NOTES

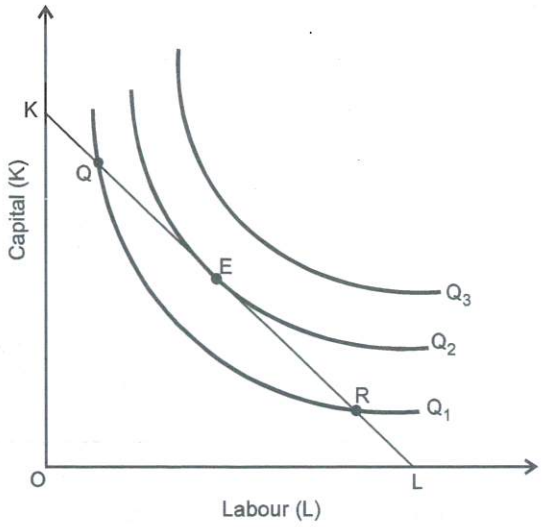


Fig. 2.8 Output Maximization for a Given Cost

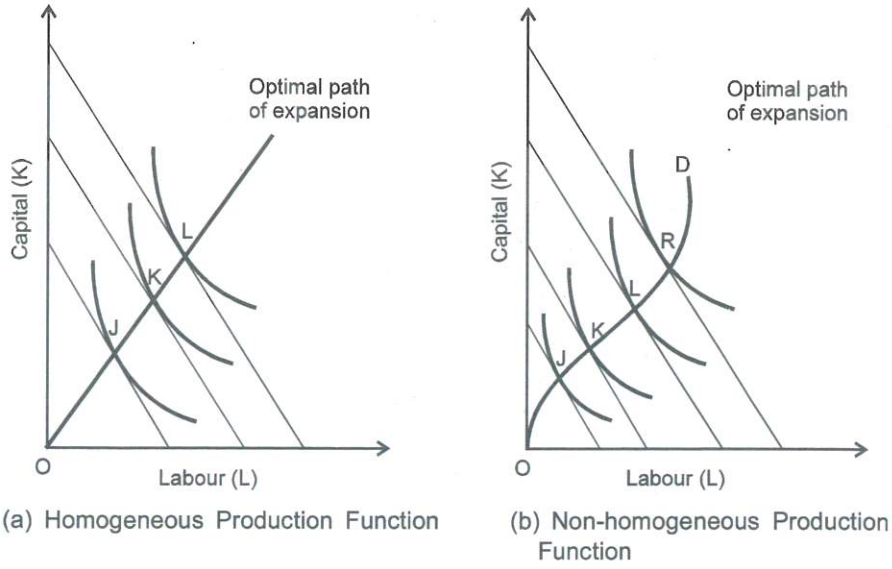
Choice of optimal expansion path

In the long-run, all inputs are variable. There is no constraint to the expansion of the output. The firms can employ capital and labour as much as they want in order to maximize their profit in the long-run. However, a profit-maximizing firm would employ capital and labour in the optimal proportion that is economically most efficient. Given the production function and input prices, the optimal factor proportion is determined by the point of tangency between the isocosts and isoquants. In other words, the optimality of factor proportion requires that at each successive employment of labour and capital, the factor-price ratio (P_k/P_l) equals the *MRTS*. The expansion of input and output through the points of optimal factor proportions gives the *optimal expansion path*.

The optimal paths of expansion are illustrated in Figure 2.9 under homogeneous and non-homogeneous production functions. If the production function is homogeneous of degree 1, the expansion path is a straight line (*OB*) from the origin, as shown in Figure 2.9 (a). The line *OB* is obtained by joining the tangential points *J*, *K* and *L*, each of which represents the optimal factor combination for a given level of output. That is, each point, *J*, *K* and *L* represents the equilibrium point at different levels of output. Note that all along the isocline, *MRTS* is constant.

If the production function is of non-homogeneous type, the expansion path is represented by the curve *OD* in Figure 2.9 (b). The expansion path represents the equilibrium path of output expansion.

NOTES



(a) Homogeneous Production Function (b) Non-homogeneous Production Function

Fig. 2.9 Optimal Paths of Expansion

Change in input prices and input combination

You have studied the least-cost combination of inputs assuming constant input prices. However, if input prices change, it will change the optimum input combination and also the level of output, given the total cost. It may be noted at the outset that if all input prices change in the same proportion, the relative prices of inputs remain unaffected. But, relative prices of the inputs will change when input prices change in different proportions and in the same direction or change unproportionately in the opposite direction, or price of only one input changes while prices of other inputs remain constant. A change in relative input prices changes both input combination and the level of output. The change in input combinations results from the *substitution effect* of change in relative prices of inputs. The change in relative prices of inputs implies that one input has become cheaper in relation to the other. The cost-minimizing firms, therefore substitute the relatively cheaper input for the costlier one. This is known as the *substitution effect* of change in relative input prices.

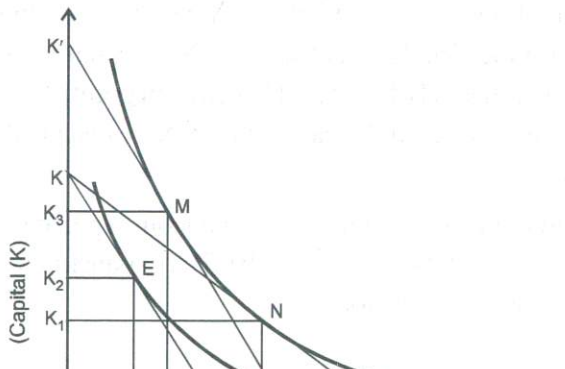


Fig. 2.10 Substitution Effect and Input Combination

NOTES

To explain the effect of change in factor prices on the input combination, let us make the following assumptions to begin with:

(i) P_k and P_l are given.

(ii) Total resources of the firm are given.

(iii) Firm's initial input combination is given.

Given the assumptions, the initial equilibrium conditions of the firm are depicted in Figure 2.10. The firm minimizes its cost of point E where the firm combines OK_2 of K and OL_1 of labour to produce output, represented by Q_1.

Given the initial conditions, let us suppose that P_l decreases while P_k remains constant so that the isocost KL shifts to KW. The isocost KW is tangent to isoquant Q_2 at point N. At this point, firm's new combination of inputs is $OK_1 + OL_3$. Thus, as a result of decrease in P_l, the firm reduces its K by K_1K_2 and increases L by L_1L_3. In other words, the firm substitutes L_1L_3 of labour for K_1K_2 of capital with the objective of maximizing its output. This change in input combination is the result of *price effect*.

The price effect is indicated by movement from point E to N. Note that after decrease in P_l, the firm reduces its K by K_1K_2 and adds L_1L_3 to its labour input. Given the slope of the isoquants, it is obvious that L_1L_3 of L is much greater than what K_1K_2 of K can substitute. It means that L_1L_3 is not the *substitution effect*.

To find the substitution effect, let us find how much additional labour the firm will employ if its resources increase so that the firm reaches the isoquant, Q_2, and the input prices remain the same. This can be found by drawing an isocost parallel to KL and tangent to Q_2, as shown by isocost $K'L'$. The isocost $K'L'$ is tangent to isoquant Q_2 at point M. It means that if P_k and P_l remain constant and firm's resources increase, it will settle itself at point M where its input combination will be OK_3 of K and OL_2 of L. This combination may be said to have resulted from the budget effect or resource effect, or the output effect. If we deduct the budget effect on labour from the price effect, we get the substitution effect, as given below.

Substitution effect = Price effect – Budget effect

Since price effect = L_1L_3,

and, budget effect = L_1L_2

Substitution effect = $L_1L_3 - L_1L_2 = L_2L_3$

Thus, we find that as a result of change in price of one input, input combination of the firm changes: the firm employs more of cheaper input and less of the costlier one. Besides, the level of output also changes. If price of an input decreases, the level of output increases, and vice versa. It is also noteworthy that the total effect of change in input price has two components: (i) substitution effect, and (ii) output effect. Thus, in our example, total price effect = $L_2L_3 + L_1L_2$.

Other Forms of Isoquants

A convex isoquant is the most widely used isoquant in traditional economic theory. The shape of an isoquant, in fact, depends on the assumption regarding the degree of substitutability between the factors in the production function. A *convex isoquant*, as shown in Figure 2.2, assumes continuous substitutability of capital and labour but at diminishing rates. Beyond a certain limit, however, substitution

NOTES

between factors is not possible, as shown is Figure 2.11. The economists have, however, assumed different degrees of substitutability between K and L and have demonstrated the existence of three other kinds of isoquants.

(i) Linear isoquants

A linear isoquant is presented by the line AB in Figure 2.11. A linear isoquant assumes perfect substitutability between the two inputs, K and L. The isoquant AB indicates that a given quantity of a product may be produced by using only capital or only labour or by using both. This is possible only when the two factors, K and L, are perfect substitutes for one another. A linear isoquant also implies that the $MRTS$ between K and L remains *constant* all along its length.

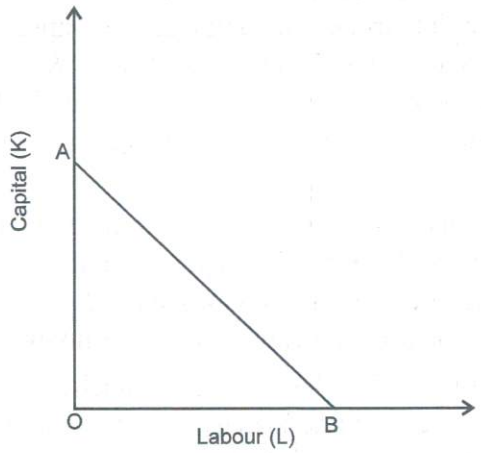


***Fig. 2.11** Linear Isoquant*

The particular mathematical form of the production function exhibiting perfect substitutability of factors is

$$Q = f(K, L)$$
$$= aK + bL \quad \ldots (2.13)$$

The production function (2.13) means that the total output, Q, is simply the weighted sum of K and L. The slope of the resulting isoquant from this production function is given by $-b/a$. This can be proved in the following way:

Given the production function (2.13),

$$MP_K = \frac{\partial Q}{\partial K} = a$$

and $$MP_L = \frac{\partial Q}{\partial L} = b$$

Since $$MRTS = \frac{MP_L}{MP_K} \quad \text{and} \quad \frac{MP_L}{MP_K} = \frac{-b}{a}$$

Therefore $$= \frac{-b}{a} = \text{slope of the isoquant}$$

The production function exhibiting perfect substitutability of factors is, however, unlikely to exist in the real-world production process.

NOTES

(ii) L-shaped isoquants

When production function assumes a technology with fixed proportion of K and L, the isoquant takes 'L' shape, as shown by Q_1 and Q_2 in Figure 2.12. Such an isoquant assumes zero substitutability between K and L. Instead, it assumes perfect complementarity between the factors. That is, K and L are treated as *perfect complements* to one another. The perfect complementarity assumption implies that a given quantity of a commodity can be produced by one and only one combination of K and L and that the *proportion* of the inputs is fixed. In other words, K and L are required in a fixed proportion to produce a given quantity of a commodity. It implies also that if quantity of an input is increased and the quantity of other input is held constant, there will be no change in the output. The output can be increased by increasing both the inputs proportionately.

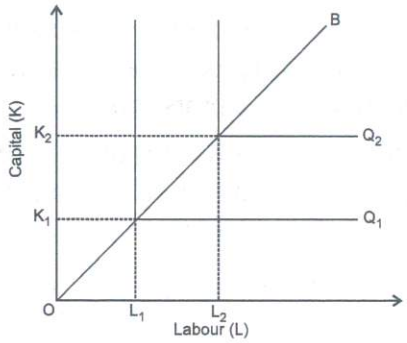


Fig. 2.12 *The L-shaped Isoquant*

As shown in Figure 2.12, to produce Q_1, OK_1 units of K and OL_1 units of L are required. It means that if OK_1 units of K are being used, OL_1 units of labour must be used to produce Q_1 units of a commodity. Similarly, if OL_1 units of labour are employed, OK_1 units of capital must be used to produce Q_1. In either case, if units of K or L are increased, output will not increase. For example, if OL_2 units of labour are used with OK_1 units of capital, then L_1L_2 $(= OL_2 - OL_1)$ units of labour would remain idle or redundant. Similarly, if OK_2 units of capital are used with OL_1 units of labour, then K_1K_2 $(= OK_2 - OK_1)$ units of K would be redundant.

One possible mathematical form of a fixed-proportion production function, frequently called a Leontief function, is

$$Q = f(K, L)$$
$$= min\ (aK, bL) \qquad \ldots (2.14)$$

where '*min*' means that Q equals the lower of the two terms, aK and bL. That is, if $aK = bL$, it would mean that both K and L are fully employed. Then the fixed capital-labour ratio will be $K/L = b/a$.

In contrast to linear production function, the fixed proportion production function has a wide range of application in the real world. One can easily find the techniques of production in which the proportion of labour and capital is fixed. To run a taxi or to operate a tractor needs only one worker—the driver. In these cases, the machine-labour proportion is fixed. Any extra labour would be redundant. Similarly, one can find cases in manufacturing industries where capital–labour proportions are fixed and an additional unit of capital would require a minimum number of workers.

(iii) Kinked or linear programming isoquant

The fixed-input-proportion production function (Figure 2.12) assumes that there is only one process of production and that capital and labour can be combined only in a fixed proportion. For example, doubling production would require doubling both the inputs, *K* and *L*. The line *OB* in Figure 2.12 represents the only production process available. In real life, however, the businessmen and the production engineers find in existence many, but not infinite, techniques of producing a given quantity of a commodity, each technique having a different fixed proportion of inputs. In fact, there is a wide range of machines available to produce a commodity. Each machine requires a fixed number of workers to work with. This number is different for each machine. For example, 40 persons can be transported from one place to another by two methods: (i) by hiring 10 four-seater taxis and 10 drivers, or (ii) by hiring a bus and 1 driver. Each of these methods is a different process of production and has a different fixed proportion of capital and labour. One can similarly find many such processes of production in manufacturing industries, each process having a different fixed-factor proportion.

Let us suppose that for producing 10 units of a commodity, *X*, there are four different techniques of production available. Each technique has a different fixed factor-proportion, as given in Table 2.3.

Table 2.3 *Alternative Techniques of Producing 100 Units of X*

Technique	*Capital*	*Labour*	*Capital/labour ratio*
I : *OA*	10	2	10:2
II : *OB*	6	3	6:3
III : *OC*	4	6	4:6
IV : *OD*	3	10	3:1

Note: *OA*, *OB*, *OC*, etc. are rays in Figure 2.13, each indicating a different technique of production.

The four hypothetical production techniques, as presented in Table 3.3, are presented graphically in Figure 3.13. The ray *OA* represents a production process having a fixed factor proportion of 10:2. Similarly, the three other production processes having fixed capital-labour ratios 6:3, 4:6, and 3:10 are shown by the rays *OB*, *OC* and *OD*, respectively. By joining the points *A*, *B*, *C* and *D*, we get a *Kinked Isoquant*, *ABCD*.

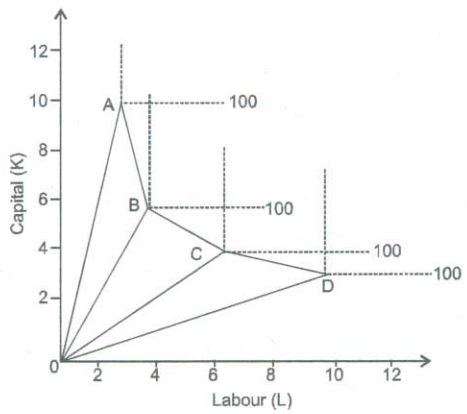


Fig. 2.13 *Fixed Factor Proportion and Isoquant*

Each of the points *A*, *B*, *C* and *D* on the kinked isoquant represents a combination of capital and labour capable of producing 100 units of commodity *X*. If there are other techniques of production, many other rays would be passing through different points between *A* and *B*, *B* and *C*, and *C* and *D*, increasing the number of kinks on the isoquant *ABCD*. The resulting isoquant would then resemble the typical isoquant. However, there is a difference: each point on a typical isoquant is technically feasible, but on a kinked isoquant, only kinks are technically feasible points.

The kinked isoquant assumes a limited substitutability of *K* and *L*, i.e., only at the points of kinks. Since this form of isoquant is used basically in linear programming, it is also called *linear programming isoquant* or *activity analysis isoquant*.

Elasticity of Technical Substitution

We have discussed above the concept and measure of marginal rate of technical substitution (*MRTS*). The *MRTS* refers only to the slope of the isoquant, i.e., to the ratio of only marginal changes in the inputs. It does not reveal how 'difficult' or 'easy' it is to substitute one input for another. Besides, the measurement of the *MRTS* depends on the units of the measurement of the factors, which does not tell much about the substitutability of factors.

Economists have devised a method of measuring the degree of substitutability of factors, called the *Elasticity of Technical Substitution*. The elasticity of substitution (s) is formally defined as *the percentage change in the capital-labour ratio (K/L) divided by the percentage change in the marginal rate of technical* substitution (*MRTS*), i.e.,

$$s = \frac{\text{Percentage change in } K/L}{\text{Percentage change in } MRTS}$$

or $$s = \frac{\partial(K/L)/(K/L)}{\partial(MRTS)/MRTS)}$$

Since along an isoquant, *K/L* and *MRTS* move in the same direction, the value of s is always positive. Besides, the elasticity of substitution (s) is 'a pure number independent of the units of the measurement of *K* and *L*, since both the numerator and the denominator are measurement of *K* and *L*, and since both the numerator and the denominator are measured in the same units.'

The concept of elasticity of substitution is graphically presented is Figure 2.14. The movement from point *A* to point *B* on the isoquant *Q* gives the ratio of change in the *MRTS*. The rays *OA* and *OB* represent two techniques of production with different factor intensities, i.e., *K/L*; while process *OA* is capital intensive, the process *OB* is labour intensive. The shift of *OA* to *OB* gives the change in factor intensity. The ratio between the two changes measures the *elasticity of technical substitution.*

NOTES

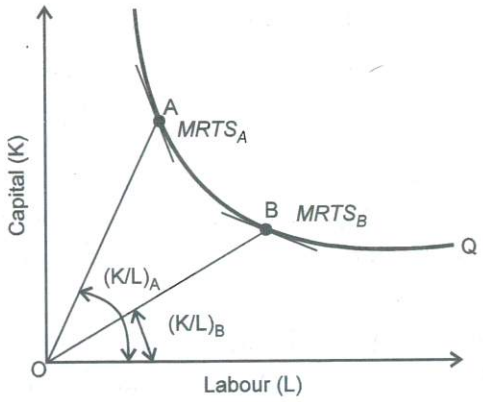


Fig. 2.14 Graphic Derivation of Elasticity of Substitution

The value of elasticity depends on the curvature of the isoquants. It varies between O and ∞, depending on the nature of the production function, which determines the curvature of the various kinds of isoquants. For example, in case of a fixed-proportion production function yielding an L-shaped isoquant (see Figure 2.12), $\sigma = 0$. If production function is such that resulting isoquant is linear (Figure 2.11), $\sigma = \infty$. In case of a homogeneous production function of degree 1 of the Cobb-Douglas type, $\sigma = 1$.

2.2.6 Laws of Returns to Scale: Isoquant and Production Function

The laws of returns to scale explain the behaviour of the total output in response to changes in the scale of the firm, i.e., in response to a simultaneous and proportional increase in all the inputs it uses. More precisely, the laws of returns to scale explain how a simultaneous and proportionate increase in both labour (L) and capital (K) affects the total output at various levels of input combination.

When a firm increases all its inputs proportionately, technically, there are three possibilities, i.e., the total output may increase proportionately, more than proportionately, or less than proportionately. If increase in the total output is proportional to the increase in inputs, it means *constant returns to scale*. For example, if inputs are doubled, then output is also doubled. If increase in the output is greater than the proportional increase in the inputs, it means *increasing returns to scale*. If increase in output is less than proportional to the increase in inputs, it means *diminishing returns to scale*.

(i) Laws of Returns to Scale through Isoquants

You have studied the laws of returns to scale with the help of homogeneous production function. The law of returns to scale may also be presented graphically, assuming a production function of Cobb-Douglas type. Let us first illustrate the *increasing returns to scale* which is first in the sequence of laws of returns to scale.

(a) Increasing returns to scale

If a proportionate change in both the inputs, K and L, leads to more than proportionate change in output, it exhibits *increasing returns to scale*. For example,

if quantities of both the inputs, K and L, are doubled and the output is more than doubled, the returns to scale is said to be increasing. The increasing returns to scale is illustrated in Figure 2.15. The movement from point *a* to *b* on the product line *OB* means doubling the inputs.

As Figure 2.15 shows, quantities of inputs increase from $K+L$ to $2K+2L$. As a result of doubling the inputs, output is more than doubled as it increases from 10 units to 25 units. Similarly, the movement from point *a* to *c* indicates a trebling of inputs as a result of which the output is more than trebled—it increases four times. This kind of relationship between the inputs and outputs shows *increasing returns to scale.*

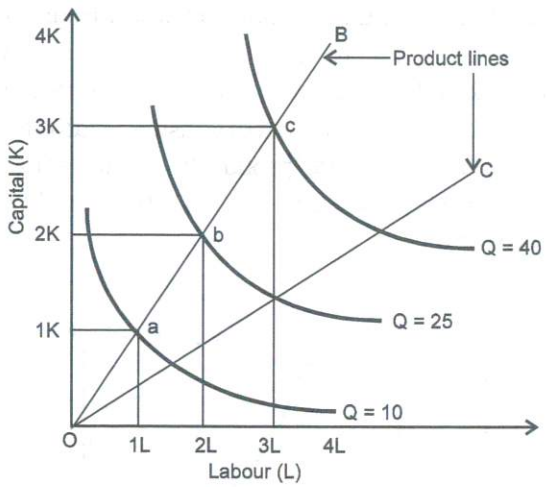


Fig. 2.15 *Increasing Returns to Scale*

Causes of increasing returns to scale

There are at least three plausible reasons for increasing returns to scale, which are as follows:

- **Technical and managerial indivisibilities:** Certain inputs, particularly capital equipments and managerial skills, used in the process of production are available in a given minimum size. Such inputs cannot be divided into a smaller size to suit a smaller scale of production. For example, half a turbine cannot be used; a quarter or a part of a locomotive engine cannot be used; one-third or a part of composite harvester and earth-mowers cannot be used. Similarly, half of a production manager cannot be employed, if part-time employment is not acceptable to the manager. Because of indivisibility of such factors, they have to be employed in a minimum quantity even if scale of production is relatively small. Therefore, when scale of production is increased by increasing all inputs, the productivity of indivisible factors increases exponentially. This results in increasing returns to scale.
- **Higher degree of specialization:** Another factor causing increasing returns to scale is higher degree of specialization of both labour and machinery, which becomes possible with increase in scale of production. The use of specialized labour and machinery increases productivity per unit of inputs. Their cumulative effects contribute to the increasing returns

to scale. Besides, managerial specialization contributes a great deal to increasing returns to scale.

- **Dimensional relations:** Increasing returns to scale is also a matter of dimensional relations. For example, when the size of a room (15′ × 10′ = 150 sq ft) is doubled to 30′ × 20′, the area of the room is more than doubled, i.e., 30′ × 20′ = 600 sq ft. When diameter of a pipe is doubled, the flow of water is more than doubled. Following this dimensional relationship, when the labour and capital are doubled, the output is more than doubled.

(b) Constant returns to scale

When a proportional change in output equals the proportional change in inputs, it exhibits *constant returns to scale*. In other words, if quantities of both the inputs, K and L, are doubled and output is also doubled, the returns to scale is said to be constant (Equation 2.18). The phenomenon of constant returns to scale is illustrated in Figure 2.16. The lines OA and OB are 'product lines' indicating three hypothetical technique of production. The isoquants, $Q = 10$, $Q = 20$ and $Q = 30$ indicate the three different levels of output.

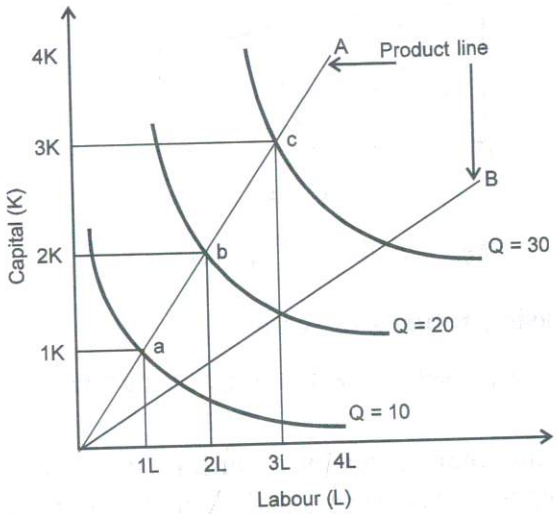


Fig. 2.16 *Constant Returns to Scale*

In the figure, the movement from point a to b indicates doubling both the inputs. That is, K increases to $2K$ and L increases to $2L$. By doubling inputs, output is also doubled from 10 to 20. Likewise, the movement from a to c indicates trebling the inputs, as K increases to $3K$ and L to $3L$. This also indicates trebling the output, as it increases from 10 to 30. This relationship between the change in inputs and the proportional change in outputs may be summed up as follows:

$1K + 1L = \quad Q = 10$

$2K + 2L = \quad Q = 20$

$3K + 2L = \quad Q = 30$

This relationship between input and output exhibits the *constant returns to scale.*

Causes of constant returns to scale

The constant returns to scale are attributable to the limits of the economies of scale. With the expansion in the scale of production, economies arise from such factors as indivisibility of certain factors, greater possibility of specialization of capital and labour, use of labour saving techniques of production, etc. But there is a limit to the economies of scale. When economies of scale disappear and diseconomies are yet to begin, the returns to scale become constant. The diseconomies arise mainly because of decreasing efficiency of management and scarcity of certain inputs.

The constant returns to scale are said to occur in the field where factors of production are perfectly divisible. When the factors of production are perfectly divisible, the production function is homogeneous of degree 1 like the Cobb-Douglas production function.

(c) Decreasing returns to scale

The firms are faced with *decreasing returns to scale* when an increase in inputs, K and L, leads to a less than proportionate increase in the output. That is, when inputs are doubled, output is less than doubled and so on. The law of decreasing returns to scale is illustrated in Figure 2.17. As the figure shows, when inputs, K and L, are doubled, i.e., increased from $K+L$ to $2K+2L$, output increases from 10 to 18 units, which is less than the proportionate increase. The movement from point b to c indicates a 50 per cent increase in the inputs. However, the output increases by only 33.3 per cent. This exhibits the *decreasing returns to scale.*

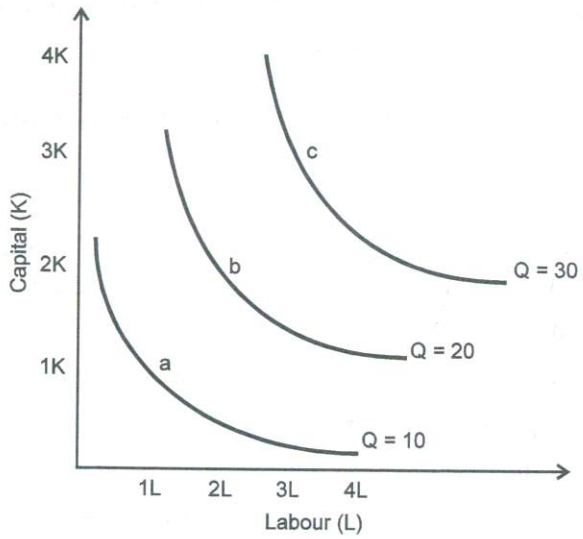


Fig. 2.17 *Decreasing Returns to Scale*

Causes of diminishing returns to scale

The decreasing returns to scale are attributed to the *diseconomies of scale.* The most important factor causing diminishing returns to scale is the diminishing returns to management, i.e., managerial diseconomies. As the size of the firm expands, managerial efficiency decreases. Another factor responsible for diminishing returns to scale is the limitedness or exhaustibility of the natural resources. For example, doubling of coal-mining plants may not double the coal output because of limitedness

NOTES

of coal deposits or difficult accessibility to coal deposits. Similarly, doubling the fishing fleet may not double the fish output because availability of fish may decrease when fishing is carried out on an increasing scale.

(ii) Production function and returns to scale

The laws of returns to scale may be explained precisely through the production function. Let us assume a production model involving only two variable inputs, K and L, and one commodity, X. Given the assumption, the production function can be expressed in its usual form as

$$Q_x = f(K, L)$$

where O_x denotes the quantity of commodity X.

Let us also assume that production function is homogeneous of degree 1. *A production function is said to be homogeneous when all the inputs are increased in the same proportion and this proportion can be factored out.* And, if all the inputs are increased in a certain proportion (say k) and the output increases in the same proportion, the production function is said to be homogeneous of degree 1. This is also known as 'linear homogeneous production function'. A production function of this type may be expressed as

$$kQ_x = f(kK, kL) \quad \text{... (2.15)}$$
$$= k(K, L)$$

A homogeneous production function implies constant returns to scale. As Equation (2.15) shows, if K and L are increased by factor k, then O_x also increases by factor k. This need not always be true. It is quite likely that if all the inputs are increased in a certain proportion, the total output may not increase in the same proportion. For example, if all the inputs are doubled, the output increases by less than or more than double. The production function in such cases may be expressed as

$$hQ_x = f(kK, kL) \quad \text{... (2.16)}$$

where h denotes the h-time increase in O_x as a result of k-time increase in the inputs, K and L. The proportion h may be greater than, equal to, or less than k. Accordingly, it brings out the three laws of returns to scale.

(i) If $h = k$, the production function reveals the *constant returns to scale.*

(ii) If $h > k$, the production function reveals *increasing returns to scale.*

(iii) If $h < k$, production function reveals *decreasing returns to scale.*

For example, consider the following production function.

$$O_{x1} = 5K + 10L \quad \text{... (2.17)}$$

If we assume that initially $K = 1$, and $L = 2$, the total output may be obtained by substituting 1 for K and 2 for L in Equation (2.17). Thus,

$$O_{x1} = 5(1) + 10(2)$$
$$25 = 5 + 20$$

In inputs are doubled (i.e., $K = 2$ and $L = 4$),

$$O_{x2} = 5(2) + 10(4)$$
$$50 = 10 + 40$$

Thus, we find that when inputs are doubled, the output is also doubled. Here $k = 2$ and since $Q_{x2}/O_{x1} = 50/25 = 2$, $h = 2$. Thus, $k = h$. This exhibits constant returns to scale. If $h < k$ or $h > k$, the production function exhibits the decreasing increasing returns to scale, and if $h > k$, it shows increasing returns to scale.

We will now use two famous production functions, viz., (a) Cobb-Douglas production function and (b) CES production function to explain the laws of returns to scale.

(a) Cobb-Douglas production function and returns to scale

Cobb-Douglas Production Function is homogeneous of degree 1. In case of a homogeneous production function of degree 1, the factor k has an exponent equal to 1, i.e., $k = k^1$. In other words, if factor k has exponent equal to 1, the production function is homogeneous of degree 1. But, as mentioned earlier, all the production functions need not be homogeneous of degree 1. A production function may be homogeneous of a degree less than 1 or greater than 1, i.e., the exponent of k may be less than 1 or greater than 1. Let us assume that exponent of k is r. A production function is of degree r if all the inputs are multiplied by constant k and output increases by a multiple, k^r. That is, if

$$f(kK, kL) = k^r(K, L) = k^r Q \qquad \ldots (2.18)$$

then the production function (2.18) is homogeneous of degree r.

From the function (2.18), we can derive the laws of returns to scale. If $k > 1$, and

(i) if $r < 1$, the production function gives decreasing returns to scale;

(ii) if $r > 1$, the function gives increasing returns to scale; and

(iii) if $r = 1$, it gives constant returns to scale

For example, consider a Cobb-Douglas type of production function.

$$Q = K^{0.25}L^{0.50} \qquad \ldots (2.19)$$

If K and L are multiplied by k, then

$$hQ = (kK)^{0.25}\ (kL)^{0.50}$$

By factoring out k, we get

$$hQ = k^{(0.25 + 0.50)}\ [K^{0.25}\ L^{0.50}]$$
$$= k^{0.75}\ [K^{0.25}\ L^{0.50}]$$

Here $h = k^{0.75}$. Note that $h < k$. The function (2.19) shows, therefore, *decreasing returns to scale.*

Consider now another production function given as

$$Q = K^{0.75}\ L^{1.25} \qquad \ldots (2.20)$$

If K and L, are multiplied by k, then

$$hQ = (kK)^{0.75}\ (kL)^{1.25}$$

By factoring out, we get

$$hQ = k^{(0.75 + 1.25)}\ [K^{0.75}\ L^{1.25}]$$
$$hQ = k^{(0.75 + 1.25)}\ [K^{0.75}\ L^{1.25}]$$
$$= k^{2.0}\ [K^{0.75}\ L^{1.25}]$$

NOTES

Here $h = k^{2.0}$, so $h > k$ and $r > 1$. Therefore, the function (2.20) gives *increasing returns to scale.* Similarly, if in a production function, $h = k$ or $r = 1$, the function shows *constant returns to scale.*

(b) CES production function and returns to scale

The Constant Elasticity of Substitution (CES) production function is another class of widely used production function. The CES production function is expressed as

$$Q = B[aL^{-r} + (1 - a)\, K^{-r}]^{-1/r} \quad \ldots (2.21)$$

where B, a and r are the three parameters, and K and L are the two inputs, and $B < O$, $O < a < 1$, and $r > -1$.

The CES production function is homogeneous of degree 1 and shows constant returns to scale. Its homogeneity of degree 1 can be proved as follows:

Let us increase inputs K and L by a constant factor m, so that Equation (2.21) may be written as

$$\begin{aligned} Q' &= B[a(mL)^{-r} + (1 - a)\, mK^{-r}]^{-1/r} \\ &= B[m^{-r}\, \{aL^{-r} + (1 - a)\, K^{-r}\}]^{-1/r} \\ &= (m^{-r})^{-1/r}\, B[aL^{-r} + (1 - a)\, K^{-r}]^{-1/r} \\ &= mQ \end{aligned}$$

Thus, the CES production function is homogeneous of degree 1. It possesses the properties of a linearly homogeneous production function. It gives constant returns to scale.

Let us study both the emperical production functions in detail.

(a) Cobb-Douglas Production Function

Cobb-Douglas production function of the following form is the most widely used production function.

$$Q = AK^aL^b \quad \ldots (2.22)$$

where A is a positive constant, a and b are positive fractions, and $b = 1 - a$. The Cobb-Douglas production function is often used in its from

$$Q = AK^aL^{1-a} \quad \ldots (2.23)$$

The Cobb-Douglas production function, which is homogeneous of degree 1, is one of the most widely used production functions to illustrate the 'Constant Returns to Scale'. It may be shown as follows.

Given the production function in Equation 2.22, let us suppose that the initial levels of Q, K and L are given as

$$Q_1 = AK_1^{\,a}L_1^{\,b} \quad \ldots (2.24)$$

When we double inputs, K and L, then Equation (2.24) can be written as

$$\begin{aligned} Q_2 &= A(2K_1)^a\, (2L_1)^b \\ &= 2^{a+b}\, AK_1^{\,a}L_1^{\,b} \quad \ldots (2.25) \end{aligned}$$

Since, $AK_1^{\,a}L_1^{\,b} = Q_1$

by substituting Q_1 in Equation (2.25), we get

$$Q_2 = 2^{a+b}Q_1$$

NOTES

But, since in Cobb-Douglas production function, $a + b = 1$,

$Q_2 = 2^1 Q_1 = 2\,Q_1$

Obviously, doubling inputs, K and L, has led to doubling of output. Thus the function exhibits constant returns to scale. If the assumption that $a + b = 1$ is dropped, it can be shown (i) that if $a + b > 1$, the Cobb-Douglas production function exhibits increasing returns to scale, and (ii) that if $a + b < 1$, the function exhibits diminishing returns to scale.

Properties of Cobb-Douglas production function

The Cobb-Douglas production function has certain properties which are of great importance in empirical studies.

First, the multiplicative form of this function (2.16) can be changed into a log-linear form

$\log Q = \log A + a \log K + b \log L$

In logarithm's form, the function becomes simple to handle and can easily be estimated using linear regression analysis.

Second, this function is a homogeneous function and the degree of its homogeneity is given by the sum of the exponents, a and b. If $a + b = 1$, the function is homogeneous of degree 1 which implies constant returns to scale.

Third, a and b represent the elasticity coefficients of output for inputs, K and L, respectively. The output elasticity coefficient (ε) in respect of capital may be defined as a proportional change in output as a result of a given change in K (L remaining constant). Thus,

$$\varepsilon_k = \frac{\partial Q}{Q} \Big/ \frac{\partial K}{K} = \frac{\partial Q}{\partial K} \cdot \frac{K}{Q} \qquad \ldots(2.26)$$

Be differentiating the production function,

$Q = AK^a L^b$

with respect to K, we get

$\partial Q / \partial K = aAK^{a-1} L^b$

By substituting the values for Q and $\partial Q / \partial K$ in Equation (2.26), we get

$$\varepsilon_k = aAK^{a-1} L^b \left(\frac{K}{AK^a L^b} \right) = a$$

Thus, output elasticity coefficient for K is 'a'. The same procedure may be adopted to show that b is elasticity coefficient of output for L.

Fourth, a and b represent the relative distributive share of inputs K and L. The share of K in Q is given by

$$\frac{\partial Q}{\partial K} \cdot K$$

and the share of L by

$$\frac{\partial Q}{\partial K} \cdot L$$

NOTES

The relative share of K can be obtained as

$$\frac{\partial Q}{\partial K} \cdot K\left(\frac{1}{Q}\right) = \frac{aAK^{a-1}L^bK}{AK^aL^b} = a$$

Similarly, it can be shown that b represents the relative share of L in the total output.

Finally, Cobb-Douglas production function in its general form, $Q = K^aL^{1-a}$ indicates that at zero cost, there will be zero production.

Some input–output relationships

Some of the concepts used in production analysis can be easily derived from the Cobb-Douglas production function as shown below.

(*i*) Average Product (*AP*) of L and K:

$$AP_L = A\,(K/L)^{1-a}$$

$$AP_K = A\,(L/K)^1$$

(*ii*) Marginal Product of L and K

$$MP_L = a.A\,(K/L) - a = a\,(Q/L)$$

$$MP_K = (a-1)\,A\,(L/K)a = (1-a)\,Q/K$$

(*iii*) Marginal Rate of Technical Substitution

$$MRTS_{L,k} = \frac{MP_L}{MP_K} = \left[\frac{a}{1-a} \cdot \frac{K}{L}\right]$$

(b) CES production function

In addition to the Cobb-Douglas production function, there are other forms of production function, viz., 'constant elasticity substitution' production function, (*CES*), 'variable elasticity of substitution' (*VES*) production function, Leontief-type of production function, and linear-type of production functions. Of these forms of production function, the constant elasticity substitution (*CES*) production function* is more widely used, apart from Cobb-Douglas production function. We will, therefore, discuss the CES production function briefly.

The *CES* production function is expressed as

$$Q = A[\alpha 2K^{-}\beta + (1-\alpha)L^{-}\beta]^{-1/}\beta \qquad \ldots (2.27)$$

or $$Q = A[\alpha L^{-}\beta + (1-\alpha)K^{-}\beta]^{-1/}\beta$$

$$(A > 0,\ 0 < \alpha < 1,\ \text{and}\ \beta > -1)$$

where L = labour, K = capital, and A, α and β are the three parameters.

An important property of the *CES* production is that it is homogeneous of degree 1. This can be proved by increasing both the inputs, K and L, by a constant factor and finding the final outcome. Let us suppose that inputs K and L are increased by a constant factor m. Then the production function given in Equation (2.27) can be written as follows.

$$Q' = A[\alpha(mK)^{-}\beta + (1-\alpha)\,(mL)^{-}\beta]^{-1/}\beta \qquad \ldots (2.28)$$

$$= A[m^{-b}\,\{\alpha K^{-}\beta + (1-\alpha)L^{-}\beta\}]^{-1/}\beta$$

$$= (m^{-}\beta)^{-1/}\beta \times A[\alpha K^{-}\beta + (1-\alpha)\,L^{-}\beta]^{-1/}\beta$$

NOTES

Since the term $A[\alpha K^{-}\beta + (1 - \alpha) L^{-}\beta]^{-1/}\beta$ in Equation (2.28) = Q, by substitution, we get

$$Q' = mQ$$

Thus, the *CES* production function is homogeneous of degree 1.

Given the production function (2.27), the marginal product of capital (K) can be obtained as

$$\frac{\delta Q}{\delta K} = \frac{\alpha}{A^{\beta}} \cdot \left[\frac{Q}{K}\right]^{\beta+1}$$

and of labour (L) as

$$\frac{\delta Q}{\delta K} = \frac{1-\alpha}{A^{\beta}} \cdot \left[\frac{Q}{L}\right]^{\beta+1}$$

The rate of technical substitution (*RTS*) can be obtained as

$$RTS = \frac{\alpha}{1-\alpha}\left[\frac{L}{K}\right]^{\beta+1}$$

The advantages of the *CES* production function are: (*i*) it is a more general form of production function, (*ii*) it can be used to analyse all types of returns to scale, and (*iii*) it removes many of the problems involved in the Cobb-Douglas production function.

The *CES* production function has, however its own *limitations*. The claim that it is a general form of production function does not stand the empirical test. Also, it is difficult to fit this function to empirical data. Uzawa finds that it is difficult to generalize this function to n number of factors. Besides, in this production function, parameter b combines the effects of two factors, K and L. When there is technological change, given the scale of production, homogeneity parameter b may be affected by both the inputs. This production does not provide a measure to separate the effects on the productivity of inputs.

For example, consider the following production function.

$$Q_x = 5K + 10L \qquad \ldots (2.29)$$

If we assume that, initially, $K = 1, L = 2$, the total output may be obtained by substituting 1 for K and 2 for L in Equation (2.29) Thus,

$$Q_{x1} = 5(1) + 10(2)$$
$$25 = 5 + 20$$

When inputs are doubled (*i.e.*, $K = 2$ and $L = 4$), then

$$Q_{x2} = 5(1 \times 2) + 10(2 \times 2)$$
$$Q_{x2} = 5(2) + 10(4)$$
$$50 = 10 + 40$$

Thus, given the production function (2.16), when inputs are doubled, the output is also doubled. Here, $k = 2$, and since

$$Q_{x2}/Q_{x1} = 50/25 = 2$$

Therefore, $h = 2$

NOTES

In this case, $k = h$. This exhibits constant returns to scale. If the production function is such that $h < k$, then the production function exhibits decreasing returns to scale, and if $h > k$, it shows increasing returns to scale.

Check Your Progress

1. What do you understand by a fixed input?
2. Mention the three kinds of isoquants.
3. What does the Cobb-Douglas production function in its general form, $Q = K^a L^{1-a}$ indicate?

2.3 COST ANALYSIS

The term 'cost analysis' refers to the measurement of the cost-output relationship. It is a comparison of costs.

2.3.1 Cost-related Concepts

Some of the major concepts related to cost are as follows:

(i) **Actual cost**

Actual costs are those which are actually incurred by the firm in payment for labour, material, plant, building, machinery, equipments, travelling and transport, etc. The total money expenses recorded in the books of accounts are, for all practical purposes, the actual costs. Actual cost concept comes under the accounting cost concept. *Opportunity cost* is another *fundamental* cost concept used in business decisions. The concept of 'opportunity cost' is related to scarcity concept. The opportunity cost is the return expected from the second best use of the resources, which is foregone for availing the gains from the best use of the resources. For example, suppose a businessman with his limited resources can buy either a printing machine or a lathe. From the printing machine he expects an annual income of ₹ 20,000, and from the lathe he expects ₹ 15,000. The rational businessman will obviously invest his money in the printing machine and forego the expected income from the lathe. The opportunity cost of his income from printing machine is the expected income from the lathe, i.e., ₹ 15,000. The opportunity cost arises because of the foregone opportunity. In assessing the opportunity costs, both *explicit* and *implicit* costs are taken into account.

Associated with the concept of opportunity cost is the concept of *economic rent* or *economic profit.* In our example, economic rent of the printing machine is the excess of its earning over the income from the lathe. Given the returns from the printing machine and lathe, economic rent = ₹ 20,000 – ₹ 15,000 = ₹ 5,000. The business implication of this concept is that investing in printing machine is preferable so long as its economic rent is greater than zero. Also, if firms know the economic rent of the various alternative uses of their resources, the choice of the best investment avenue will be a problem.

(ii) Business cost and full cost

Business costs include all the expenses which are incurred in carrying out a business. The concept of business cost is similar to the actual or real cost. Business costs

'include all the payments and contractual obligations made by the firm together with the book cost of depreciation on plant and equipment.' Both these concepts are used in calculating the profits and losses in the business in filing returns for income tax and for other legal purposes.

NOTES

The concept of *full cost* includes two other costs: *opportunity cost* and *normal profit.* Opportunity cost, as noted above, includes the expected earning from the next best use of the resources or the market rate of interest on the total money capital, and also the value or entrepreneurs own services which are not charged in the current business. Normal profit is a necessary minimum earning, in addition to opportunity cost, which a firm must get to remain in its present occupation.

(iii) Explicit and implicit or imputed costs

Explicit costs are those which fall under actual or business costs entered in the books of accounts. The payments on account of wages, salaries, utilities, interest, rent, purchase of materials, licence fee, insurance premium and depreciation charges are the examples of explicit costs. These costs involve cash payment and are clearly reflected by the normal accounting practices. In contrast with these costs, there are certain other costs which do not take the form of cash outlays, nor do they appear in the accounting system. Such costs are known as *implicit* or *imputed* costs. Implicit costs may be defined as the earning of owner's resources employed in their best alternative uses. For example, suppose an entrepreneur does not utilize his services in his own business and works as a manager in some other firm on a salary basis. If he joins his own business, he foregoes his salary as manager. This loss of salary which is opportunity cost of his services utilized in his own firm becomes an implicit cost of his own business. It is implicit because the income foregone by the entrepreneur is not charged as the explicit cost of his own business. The implicit cost includes implicit wages, implicit rent, implicit interest, etc. Although implicit costs are not taken into account while calculating the loss or gains of the business, these costs do figure in business decisions.

(iv) Total, average and marginal costs

Total cost represents the value of the total resources used in the production of goods and services. It refers to the total outlays of money expenditure, both explicit and implicit on the resources used to produce a given output. For theoretical purpose, total cost includes payments for labour, capital, land and opportunity cost. The total cost for a given output is obtained from the cost function.

Average cost is of statistical nature. It is obtained simply by dividing the total cost (TC) by the total output (Q), i.e., TC/Q = average cost.

Marginal cost is the addition to the total cost on account of producing one additional unit of the product. Or, marginal cost is the cost of marginal unit produced. Marginal cost (MC) is also defined as DTC/DQ.

(v) Fixed and variable costs

Fixed costs are those which are fixed in volume for a certain given output. Fixed costs do not vary with the variation in the output between zero and a certain level of output. The costs that do not vary over a certain level of output are known as

NOTES

fixed cost. Fixed costs include cost of (i) managerial and administrative staff; (ii) depreciation of machinery, building and other fixed assets; and (iii) maintenance of land, etc. The *concept of fixed cost is associated with short-run.*

Variable costs are those which vary with the variation in the total output. Variable costs are the function of the output. Variable costs include cost of raw materials, running cost of fixed capital, such as fuel, ordinary repairs, routine maintenance expenditure, direct labour charges associated with the level of output, and the costs of all other inputs that vary with output.

These cost concepts are economic in nature and are used in economic analysis of costs behaviour in relation to output.

(vi) Short-run and long-run costs

Two other important cost concepts which are analogous with variable and fixed costs and often figure in economic analysis are short-run and long-run costs. Short-run costs can be defined as the costs which vary with the variation of output, the size of the firm remaining the same. In other words, short-run costs are the same as variable costs. Long-run costs, on the other hand, can be defined as the costs which are incurred on the fixed assets, like plant, building machinery, land, etc. Such costs have long-run implication in the sense that these costs are not used up in the single batch of production, and are used over time in the process of production. Long-run costs are, by implication, the same as fixed costs. In the long-run, however, even the fixed costs become variable costs as the size of the firm or scale of production increases. Broadly speaking, the short-run costs are those associated with variable costs in the utilization of fixed plant or other facilities, whereas long-run cost-behaviour encompasses changes in the size and kind of plant.

(vii) Private and social costs

There are certain other costs which arise due to functioning of the firm but do not normally figure in the business decisions nor are such costs explicitly paid by the firms. Some such costs are paid by the society. Thus, the total cost generated by the firm's decision may be divided into two categories: (a) those paid out or provided for by the firms; (b) those not paid by the firms, including use of resources freely available *plus* the disutility created in the process of production. The costs of the category (a) are known as *private costs,* and costs of category (b) are known as *external* or *social costs*. For instance, private firms, situated closer Yamuna river discharge their wastes into the Yamuna river causing water-pollution; mills and factories located in a city cause air pollution by emitting smoke; and plying cars, buses, and trucks cause both air and noise pollution. Such pollutions cause tremendous health hazards which involve cost to the society as a whole. Such costs do not figure in the cost structure of the firms and hence are termed *external costs* from the firm's point of view, and *social cost* from society's point of view.

Private costs are those which are actually incurred or provided for by an individual or a firm on the purchase of goods and services from the market. For a firm, all the actual costs, both explicit and implicit, are private costs. Private costs are internalized in the sense that 'the firm must compensate the resource owner in order to acquire the right to use the resource.' It is only the internalized cost that is incorporated in the firm's total cost of production.

Social cost, on the other hand, implies the cost which a society bears on account of production of a commodity. Social cost includes both *private cost*

and the *external cost*. External cost includes (a) the cost of 'resources for which the firm is not compelled to pay a price,' e.g. atmosphere, rivers, and lakes and also for the use of public utility services like roadways and drainage system and (b) the cost in the form of 'disutility' caused by air, water, and noise pollution, etc. The cost of category (*b*) is generally assumed to be equal to the total private and public expenditure incurred to safeguard the individual and public interest against the various kinds of health hazards created by the production system. But private and public expenditure serves only as an indicator of trends in 'public disutility'; it does not give the exact measure of the public disutility.

2.3.2 Traditional Theory of Cost

Cost function is a symbolic statement of the technological relationship between cost and output. In its general form, it is expressed by an equation. Cost function can be expressed also in the form of a schedule and a graph. In fact, tabular, graphical, and algebraic equation forms of cost function can be converted in the form of each other. Going by its general form, total cost (*TC*) function is expressed as follows.

$$TC = f(Q)$$

This form of cost function tells only that there is a relationship between *TC* and output (*Q*). But it does not tell the nature of relationship between *TC* and *Q*. Since there is a positive relationship between *TC* and *Q*, cost function must be written as

$$TC = f(Q), \quad \Delta TC/\Delta Q > 0$$

This cost function means that *TC* depends on *Q* and that increase in output (*Q*) causes increase in *TC*. The nature and extent of this relationship between *TC* and *Q* depends on the product and technology. For example, cost of production increases at a constant rate in case of clothes, furniture and building, given the technology. In case raw materials and labour become scarce as production increases, cost of production increases at increasing rate. In case of agricultural products, cost of production increases first at decreasing rate and then at increasing rate. When these three kinds of *TC* and *Q* relationships are estimated on the basis of actual production and cost data, three different kinds of cost functions emerge as given below.

Kinds of Cost Functions and Change in TC

Nature of Cost Function	*Cost Function*	*Change in TC*
Linear	$TC = a + bQ$	*TC* increases at constant rate
Quadratic	$TC = a + bQ + Q^2$	*TC* increases at increasing rate
Cubic	$TC = a + bQ - Q^2 + Q^3$	*TC* increases first at decreasing rate than at increasing rate

These cost functions are explained further and illustrated graphically.

Short-run Cost-output Relations

The theory of cost deals with the behaviour of cost in relation to a change in output. In other words, the cost theory deals with cost-output relations. The basic principle of the cost behaviour is that the *total cost increases with increase in*

output. This simple statement of an observed fact is of little theoretical and practical importance. What is of importance from a theoretical and managerial point of view is not the absolute increase in the total cost but the direction of change in the average cost (*AC*) and the marginal cost (*MC*). The direction of change in *AC* and *MC*—whether *AC* and *MC* decrease or increase or remain constant—depends on the nature of the cost function. The specific form of the cost function depends on whether the time framework chosen for cost analysis is short-run or long-run. It is important to recall here that some costs remain constant in the short-run while all costs are variable in the long-run. Thus, depending on whether cost analysis pertains to short-run or to long run, there are two kinds of cost functions:

(*i*) short-run cost functions, and (*ii*) long-run cost functions,

Accordingly, the cost output relations are analyzed in short-run and long-run framework. In this section, we will analyse the short-run cost-output relations by using cost function. The long-run cost-output relations are discussed in the following section.

Cost Concepts used in Cost Analysis

Before we discuss the cost-output relations, let us first look at the cost concepts and the components used to analyse the short-run cost-output relations.

The basic analytical cost concepts used in the analysis of cost behaviour are Total, Average and Marginal costs. The total cost *(TC)* is defined as the actual cost that must be incurred to produce a given quantity of output. The short-run *TC* is composed of two major elements: (*i*) *total fixed cost (TFC),* and (*ii*) *total variable cost (TVC)*. That is, in the short-run,

$$TC = TFC + TVC \qquad ...(2.30)$$

As mentioned earlier, *TFC* (i.e., the cost of plant, machinery building, etc.) remains fixed in the short-run, whereas *TVC* varies with the variation in the output.

For a given quantity of output *(Q)*, the average total cost *(AC)*, average fixed cost *(AFC)* and average variable cost *(AVC)* can be defined as follows:

$$AC = \frac{TC}{Q} = \frac{TFC + TVC}{Q}$$

$$AFC = \frac{TFC}{Q}$$

$$AVC = \frac{TVC}{Q}$$

and $$AC = AFC + AVC \qquad ...(2.31)$$

Marginal cost *(MC)* is defined as the change in the total cost divided by the change in the total output, i.e.,

$$MC = \frac{\Delta TC}{\Delta Q} \qquad ...(2.32)$$

or as the first derivative of cost function, i.e., $\frac{\partial TC}{\partial Q}$.

Note that since $\Delta TC = \Delta TFC + \Delta TVC$ and, in the short-run, $\Delta TFC = 0$, therefore, $\Delta TC = \Delta TVC$. Furthermore, under the marginality concept, where $\Delta Q = 1$, $MC = \Delta TVC$. Now we turn to cost function and derivation of cost curves.

NOTES

Short-Run Cost Functions and Cost Curves

The cost-output relations are determined by the cost function and are exhibited through cost curves. The shape of the cost curves depends on the nature of the cost function. Cost functions are derived from actual cost data of the firms. Given the cost data, estimated cost functions may take a variety of forms, yielding different kinds of cost curves. The cost curves produced by *linear, quadratic* and *cubic cost functions* are illustrated below.

1. Linear Cost Function. A linear cost function takes the following form.

$$TC = a + bQ \qquad \ldots(2.33)$$

(where TC = total cost, Q = quantity produced, $a = TFC$, and $b = \partial TC/\partial Q$).

Given the cost function (Eq. 2.33), AC and MC can be obtained as follows.

$$AC = \frac{TC}{Q} = \frac{a + bQ}{Q} = \frac{a}{Q} + b$$

and $$MC = \frac{\partial TC}{\partial Q} = b$$

Note that since 'b' is a constant factor, MC remains constant throughout in case of a linear cost function.

Assuming an actual cost function given as

$$TC = 60 + 10Q \qquad \ldots(2.34)$$

the cost curves (TC, TVC and TFC) are graphed in Fig. 2.18.

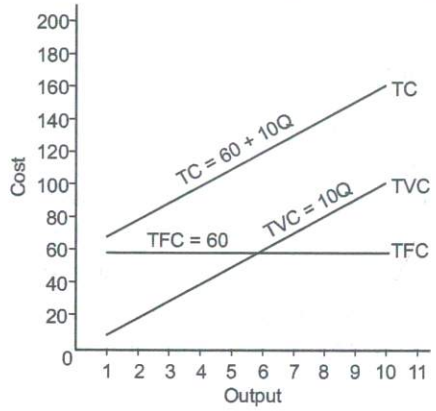


Fig. 2.18 *Linear Cost Functions*

Given the cost function (Eq. 2.34),

$$AC = \frac{60}{Q} + 10$$

and $$MC = 10$$

Fig. 2.19 shows the behaviour of TC, TVC and TFC. The straight horizontal line shows TFC and the line marked $TVC = 10Q$ shows the movement in TVC. The total cost function is shown by $TC = 60 + 10Q$.

NOTES

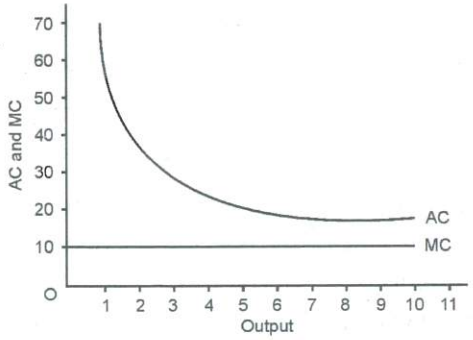


Fig. 2.19 AC and MC Curves Derived from Linear Cost Function

More important is to notice the behaviour of *AC* and *MC* curves in Fig. 2.19. Note that in case of a linear cost function *MC* remains constant, while *AC* continues to decline with the increase in output. This is so simply because of the logic of the linear cost function.

2. *Quadratic Cost Function.* A quadratic cost function is of the form

$$TC = a + bQ + Q^2 \qquad ....(2.35)$$

where a and b are constants.

Given the cost function (Eq. 2.35), *AC* and *MC* can be obtained as follows.

$$AC = \frac{TC}{Q} = \frac{a + bQ + Q^2}{Q} = \frac{a}{Q} + b + Q \qquad ...(2.36)$$

$$MC = \frac{\partial TC}{\partial Q} = b + 2Q \qquad ...(2.37)$$

Let us assume that the actual (or estimated) cost function is given as

$$TC = 50 + 5Q + Q^2 \qquad ...(2.38)$$

Given the cost function,

$$AC = \frac{50}{Q} + Q + 5 \quad \text{and} \quad MC = \frac{\partial C}{\partial Q} = 5 + 2Q$$

The cost curves that emerge from the cost function (2.38) are graphed in Fig. 2.20 (*a*) and (*b*). As shown in panel (*a*), while fixed cost remains constant at 50, *TVC* is increasing at an increasing rate. The rising *TVC* sets the trend in the total cost *(TC)*. Panel (*b*) shows the behaviour of *AC*, *MC* and *AVC* in a quadratic cost function. Note that *MC* and *AVC* are rising at a constant rate whereas *AC* first declines and then increases.

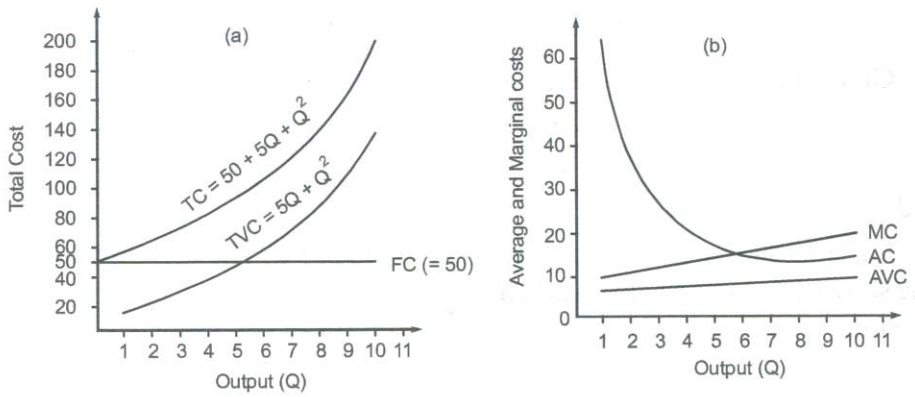


Fig. 2.20 Cost Curves Derived from a Quadratic Cost Function

3. *Cubic Cost Function.* A cubic cost function is of the form

$$TC = a + bQ - cQ^2 + Q^3 \quad \ldots(2.39)$$

where a, b and c are the parametric constants.

From the cost function (2.39), AC and MC can be derived as follows.

NOTES

$$AC = \frac{TC}{Q} = \frac{a + bQ - cQ^2 + Q^3}{Q} = \frac{a}{Q} + b - cQ + Q^2$$

and $$MC = \frac{\partial TC}{\partial Q} = b - 2cQ + 3Q^2$$

Let us suppose that the cost function is empirically estimated as

$$TC = 10 + 6Q - 0.9Q^2 + 0.05Q^3 \quad \ldots(2.40)$$

Given the cost function (2.40), the TVC function can be derived as

$$TVC = 6Q - 0.9Q^2 + 0.05Q^3 \quad \ldots(9.12)$$

The TC and TVC, based on Eqs. (2.40) and (2.41), respectively, have been calculated for $Q = 1$ to 16 and presented in Table 9.1. The TFC, TVC and TC have been graphically presented in Fig. 2.21. As the figure shows, TFC remains fixed for the whole range of output, and hence, takes the form of a horizontal line—TFC. The TVC curve shows that the total variable cost first increases at a decreasing rate and then at an increasing rate with the increase in the output. The rate of increase can be obtained from the slope of the TVC curve. The pattern of change in the TVC stems directly from the law of increasing and diminishing returns to the variable inputs. As output increases, larger quantities of variable inputs are required to produce the same quantity of output due to diminishing returns. This causes a subsequent increase in the variable cost for producing the same output.

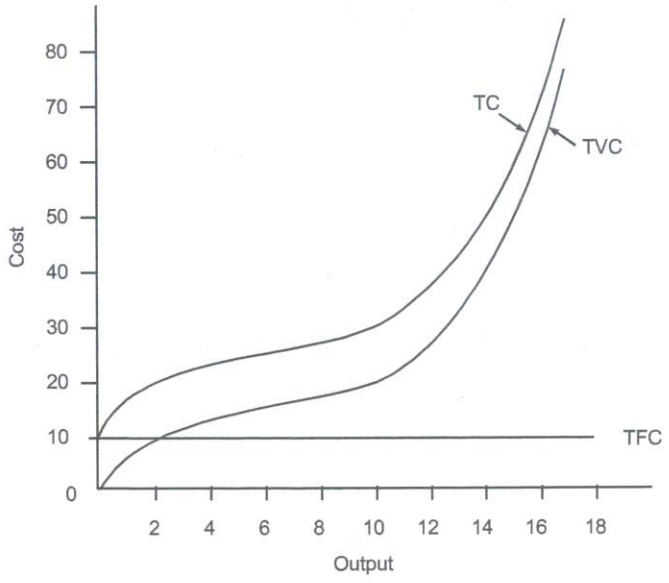


Fig. 2.21 *TC, TFC and TVC Curves*

NOTES

Table 2.4 *Cost-Output Relations*

Q	*FC*	*TVC*	*TC*	*AFC*	*AVC*	*AC*	*MC*
(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*	*(8)*
0	10	0.0	10.00	—	—	—	—
1	10	5.15	15.15	10.00	5.15	15.15	5.15
2	10	8.80	18.80	5.00	4.40	9.40	3.65
3	10	11.25	21.25	3.33	3.75	7.08	2.45
4	10	12.80	22.80	2.50	3.20	5.70	1.55
5	10	13.75	23.75	2.00	2.75	4.75	0.95
6	10	14.40	24.40	1.67	2.40	4.07	0.65
7	10	15.05	25.05	1.43	2.15	3.58	0.65
8	10	16.00	26.00	1.25	2.00	3.25	0.95
9	10	17.55	27.55	1.11	1.95	3.06	1.55
10	10	20.00	30.00	1.00	2.00	3.00	2.45
11	10	23.65	33.65	0.90	2.15	3.05	3.65
12	10	28.80	38.80	0.83	2.40	3.23	5.15
13	10	35.75	45.75	0.77	2.75	3.52	6.95
14	10	44.80	54.80	0.71	3.20	3.91	9.05
15	10	56.25	66.25	0.67	3.75	4.42	11.45
16	10	70.40	80.40	0.62	4.40	5.02	14.15

From Eqs. (2.40) and (2.41), we may derive the behavioural equations for *AFC*, *AVC* and *AC*. Let us first consider *AFC*.

Average Fixed Cost (AFC). As already mentioned, the costs that remain fixed for a certain level of output make the total fixed cost in the short-run. The fixed cost is represented by the constant term '*a*' in Eq. (2.39) and $a = 10$ as given in Eq. (2.40). We know that

$$AFC = \frac{TFC}{Q} \qquad \text{....(2.42)}$$

Substituting 10 for *TFC* in Eq. 2.37, we get

$$AFC = \frac{10}{Q} \qquad \text{....(2.43)}$$

Equation (2.43) expresses the behaviour of *AFC* in relation to change in *Q*. The behaviour of *AFC* for *Q* from 1 to 16 is given in Table 2.4 (col. 5) and presented graphically by the *AFC* curve in Fig. 2.22. The *AFC* curve is a rectangular hyperbola.

Average Variable Cost (AVC). As defined above, $AVC = TVC/Q$. Given the *TVC* function (Eq. 2.41), we may express *AVC* as follows.

$$AVC = \frac{6Q - 0.9Q^2 + 0.05Q^3}{Q}$$

$$= 6 - 0.9Q + 0.05Q^2 \qquad \text{...(2.44)}$$

Having derived the *AVC* function in Eq. (2.44), we may easily obtain the behaviour of *AVC* in response to change in *Q*. The behaviour of *AVC* for $Q =$ 1 to 16 is given in Table 2.4 (col. 6), and graphically presented in Fig. 2.19 by the *AVC* curve.

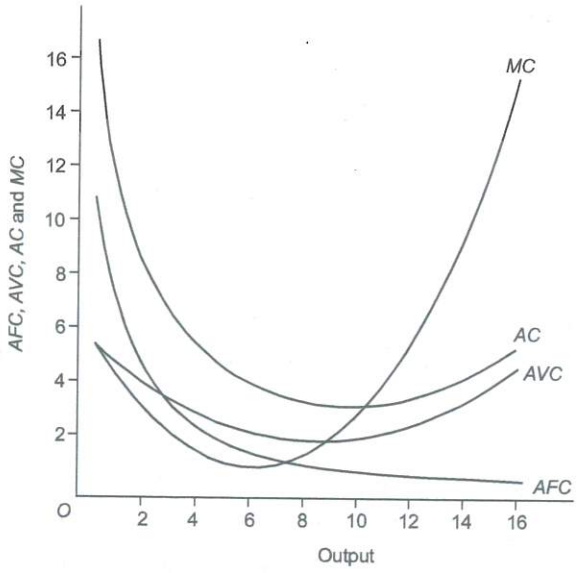


Fig. 2.22 Short-run AFC, AVC, AC and MC Curves

Critical Value of AVC. From Eq. (2.44), we may compute the critical value of Q in respect of *AVC*. *The critical value of Q (in respect of AVC) is one that minimizes AVC.* The *AVC* will be minimum when its rate of decrease equals zero. This can be accomplished by differentiating Eq. (2.44) and setting it equal to zero. Thus, critical value of Q can be obtained as

$$\text{Critical value of } Q = \frac{\partial AVC}{\partial Q} = -0.9 + 0.10\,Q = 0$$

$$0.10\,Q = 0.9$$

$$Q = 9$$

In our example, the critical value of $Q = 9$. This can be verified from Table 2.4. The *AVC* is minimum (1.95) at output 9.

Average Cost (AC). The average cost (AC) is defined as $AC = \frac{TC}{Q}$.

Substituting Eq. (2.40) for TC in the above equation, we get

$$AC = \frac{10 + 6Q - 0.9Q^2 + 0.05Q^3}{Q}$$

$$= \frac{10}{Q} + 6 - 0.9Q + 0.05Q^2 \qquad \ldots(2.45)$$

The Eq. (9.16) gives the behaviour of AC in response to change in Q. The behaviour of AC for $Q = 1$ to 16 is given in Col. 7 of Table 2.4 and graphically presented in Fig. 2.22 by the AC curve. Note that AC curve is U-shaped.

Minimization of AC. One objective of business firms is to minimize AC of their product or, which is the same as, to optimize the output. The level of output that minimizes AC can be obtained by differentiating Eq. (2.45) and setting it equal to zero. Thus, the optimum value of Q can be obtained as follows.

$$\frac{\partial AC}{\partial Q} = \frac{10}{Q^2} - 0.9 + 0.1Q = 0$$

NOTES

When simplified (multiplied by Q^2) this equation takes the quadratic form as

$$-10 - 0.9Q^2 + 0.1Q^3 = 0$$

or $$Q^3 - 9Q^2 - 100 = 0 \quad ...(2.46)$$

By solving equation (2.46) we get $Q = 10$.

Thus, the critical value of output in respect of AC is 10. That is, AC reaches its minimum at $Q = 10$. This can be verified from Table 2.4.

Marginal Cost (MC). The concept of marginal cost (MC) is useful particularly in economic analysis. MC is technically the first derivative of the TC function. Given the TC function in Eq. (2.40), the MC function can be obtained as

$$MC = \frac{\partial TC}{\partial Q} = 6 - 1.8Q + 0.15Q^2 \quad ...(2.47)$$

Eq. (2.47) represents the behaviour of MC. The behaviour of MC for Q = 1 to 16 computed as $MC = TC_n - TC_{n-1}$ is given in Table 2.4 (col. 8) and graphically presented by the MC curve in Fig. 2.22. The critical value of Q with respect to MC is 6 or 7. This can be seen from Table 2.4.

Cost Curves and the Law of Diminishing Returns

Now we return to the law of variable proportions and explain it through the cost curves. Figures 2.21 and 2.22 represent the cost curves conforming to the short-term law of production, i.e., the law of diminishing returns. Let us recall the law: it states that when more and more units of a variable input are applied, other inputs held constant, the returns from the marginal units of the variable input may initially increase but it decreases eventually. The same law can also be interpreted in terms of decreasing and increasing costs. The law can then be stated as, if more and more units of a variable input are applied to a given amount of a fixed input, the marginal cost initially decreases, but eventually increases. Both interpretations of the law yield the same information—one in terms of ***marginal productivity*** of the variable input, and the other in terms of the ***marginal cost***. The former is expressed through a production function and the latter through a cost function.

Figure 2.22 presents the short-run laws of return in terms of cost of production. As the figure shows, in the initial stage of production, both AFC and AVC are declining because of some internal economies. Since $AC = AFC + AVC$, AC is also declining. This shows the operation of the law of increasing returns to the variable input. But beyond a certain level of output (i.e., 9 units in our example), while AFC continues to fall, AVC starts increasing because of a faster increase in the TVC. Consequently, the rate of fall in AC decreases. The AC reaches its minimum when output increases to 10 units. Beyond this level of output, AC starts increasing which shows that the law of diminishing returns comes into operation. The MC curve represents the change in both the TVC and TC curves due to change in output. A downward trend in the MC shows increasing marginal productivity of the variable input due mainly to internal economy resulting from increase in production. Similarly, an upward trend in the MC shows increase in TVC, on the one hand, and decreasing marginal productivity of the variable input, on the other.

Some important relationships between costs used in analysing the short-run cost-behaviour may now be summed up as follows:

NOTES

(*a*) Over the range of output both *AFC* and *AVC* fall, *AC* also falls because $AC = AFC + AVC$.

(*b*) When *AFC* falls but *AVC* increases, change in *AC* depends on the rate of change in *AFC* and *AVC*.

(*i*) if decrease in *AFC* > increase in *AVC*, then *AC* falls,

(*ii*) if decrease in *AFC* = increase in *AVC*, *AC* remains constant, and

(*iii*) if decrease in *AFC* < increase in *AVC*, then *AC* increase.

(*c*) The relationship between *AC* and *MC* is of a varied nature. It may be described as follows:

(*i*) When *MC* falls, *AC* follows, over a certain range of initial output. When *MC* is falling, the rate of fall in *MC* is greater than that of *AC*, because in the case of *MC* the decreasing marginal cost is attributed to a single marginal unit while, in case of *AC*, the decreasing marginal cost is distributed over the entire output. Therefore, *AC* decreases at a lower rate than *MC*.

(*ii*) Similarly, when *MC increases, AC also increases but at a lower rate* for the reason given in (*i*). There is, however, a range of output over which the relationship does not exist. Compare the behaviour of *MC* and *AC* over the range of output from 6 to 10 units (Fig. 2.22). Over this range of output, *MC* begins to increase while *AC* continues to decrease. The reason for this can be seen in Table 2.4: when *MC* starts increasing, it increases at a relatively lower rate which is sufficient only to reduce the rate of decrease in *AC*—not sufficient to push the *AC* up. That is why *AC* continues to fall over some range of output even if *MC* increases.

(*iii*) The *MC curve intersects the AC at its minimum point*. This is simply a mathematical relationship between *MC* and *AC* curves when both of them are obtained from the same *TC* function. In simple words, when *AC* is at its minimum, it is neither increasing nor decreasing: it is constant. When *AC* is constant, *AC* = *MC*. That is the point of intersection.

Output Optimization in the Short-Run

Optimization of output in the short-run has been illustrated graphically in Fig. 2.22. Optimization technique is repeated here for the sake of completeness.

Let us suppose that a short-run cost function is given as

$$TC = 200 + 5Q + 2Q^2 \qquad \ldots(2.48)$$

We have noted above that an optimum level of output is one that equalizes *AC* and *MC*. In other words, at optimum level of output, *AC* = *MC*. Given the cost function in Eq. (2.48),

$$AC = \frac{200 + 5Q + 2Q^2}{Q} = \frac{200}{Q} + 5 + 2Q \qquad \ldots (2.49)$$

and

$$MC = \frac{\partial TC}{\partial Q} = 5 + 4Q \qquad \ldots (2.50)$$

By equating *AC* and *MC* equations, i.e., Eqs. (2.49) and (2.50), respectively, and solving them for *Q*, we get the optimum level of output. Thus,

$$\frac{200}{Q} + 5 + 2Q = 5 + 4Q,$$

$$\frac{200}{Q} = 2Q$$

$$2Q^2 = 200 \quad \text{or} \quad Q = 10$$

Thus, given the cost function (2.48), the optimum output is 10.

NOTES

Long-Run Cost-Output Relations

By definition, long-run is a period in which all the inputs–specifically, labour and capital–become variable. The variability of inputs is based on the assumption that in the long-run supply of all the inputs, including those held constant in the short-run, becomes elastic. The firms are, therefore, in a position to expand the scale of their production by hiring a larger quantity of both labour and capital. The long-run-cost-output relations, therefore, imply the relationship between the changing scale of the firm and the total output, whereas in the short-run this relationship is essentially one between the total output and the variable cost (labour).

Derivation of Long-run Cost Curves

To understand the long-run-cost-output relations and to derive long-run cost curves it will be helpful to imagine that a long-run is composed of a series of short-run production decisions. As a corollary of this, long-run cost curve is composed of a series of short-run cost curves. Based on this principle, we may now derive the long-run cost curves and study their relationship with output.

Long-run Total Cost Curve (LTC). In order to draw the long-run total cost curve, let us begin with a short-run situation. Suppose that a firm having only one plant has its short-run total cost curve as given by STC_1, in panel (*a*) of Fig. 2.23. Let us now suppose that the firm decides to add two more plants to its size over time, one after the other. As a result, two more short-run total cost curves are added to STC_1, in the manner shown by STC_2 and STC_3 in Fig. 2.23 (*a*). The *LTC* can now be drawn through the minimum points of STC_1, STC_2 and STC_3 as shown by the *LTC* curve corresponding to each *STC*.

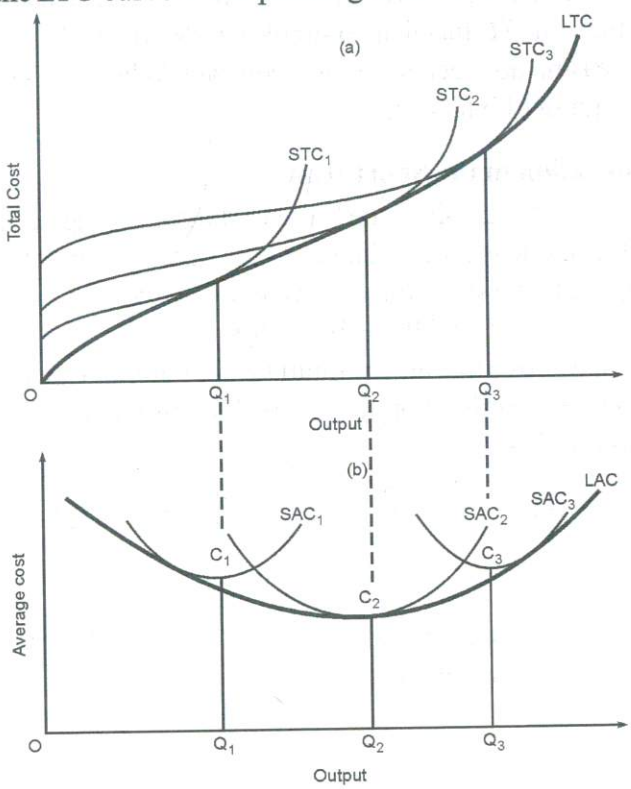


Fig. 2.23 Long-run Total and Average Cost Curves

Long-run Average Cost Curve (LAC). The long-run average cost curve (*LAC*) is derived by combining the short-run average cost curves (SAC_s). Note that there is one *SAC* associated with each *STC*. Given the STC_1, STC_2, STC_3 curves in panel (*a*) of Fig. 2.23 there are three corresponding *SAC* curves as given by SAC_1, SAC_2, and SAC_3 curves in panel (*b*) of Fig. 2.23. Thus, the firm has a series of *SAC* curves, each having a bottom point showing the minimum *SAC*. For instance, C_1Q_1 is minimum *AC* when the firm has only one plant. The *AC* decreases to C_2Q_2 when the second plant is added and then rises to C_3Q_3 after the addition of the third plant. The *LAC* curve can be drawn through the bottom points of the SAC_1, SAC_2 and SAC_3 as shown in Fig. 2.23(*b*) The *LAC* curve is also known as the **'Envelope Curve'** or **'Planning Curve'** as it serves as a guide to the entrepreneur in his plans to expand production.

Alternatively, the *SAC* curves can be derived from the data given in the *STC* schedule, from *STC* function or straightaway from the *STC* curve Similarly, *LAC* and can be derived from *LTC*-schedule, *LTC* function or form *LTC*-curve.

The relationship between *LTC* and output, and between *LAC* and output can now be easily derived. It is obvious from the *LTC* that the long-run cost-output relationship is similar to the short-run cost-output relation. With the subsequent increases in the output, *LTC* first increases at a decreasing rate, and then at an increasing rate. As a result, *LAC* initially decreases until the optimum utilization of the second plant and then it begins to increase. These cost-output relations follow the 'laws of returns to scale'. When the scale of the firm expands, unit cost of production initially decreases, but ultimately increases as shown in Fig. 2.23(*b*). The decrease in unit cost is attributed to the internal and external economies of scale and the eventual increase in cost is linked to the internal and external diseconomies. The economies and diseconomies of scale are discussed in the following section.

Long-run Marginal Cost Curve (LMC). The *long-run marginal cost curve* (*LMC*) is derived from the short-run marginal cost curves (*SMCs*). The derivation of *LMC* is illustrated in Fig. 2.24 in which *SACs* and *LAC* are the same as in Fig. 2.23 (*b*) To derive the *LMC*, consider the points of tangency between *SACs* and the *LAC*, i.e., points *A*, *B* and *C*. In the long-run production planning, these points determine the output levels at the different levels of production. For example, if we draw perpendiculars from points *A*, *B* and *C* to the *X*-axis, the corresponding output levels will be OQ_1, OQ_2 and OQ_3. The perpendicular AQ_1 intersects the SMC_1 at point *M*. It means that at output OQ_1, *LMC* is MQ_1. If output increases to OQ_2, *LMC* rises to BQ_2. Similarly, CQ_3 measures the *LMC* at output OQ_3. A curve drawn through points *M*, *B* and *N*, as shown by the *LMC*, represents the behaviour of the marginal cost in the long-run. This curve is known as the long-run marginal cost curve, *LMC*. It shows the trends in the marginal cost in response to the changes in the scale of production.

Some important inferences may be drawn from Fig. 2.24. The *LMC* must be equal to *SMC* for the output at which the corresponding *SAC* is tangent to the *LAC*. At the point of tangency, *LAC* = *SAC*. Another important point to notice is that *LMC* intersects *LAC* when the latter is at its minimum, i.e., point *B*. There

NOTES

is one and only one short-run plant size whose minimum *SAC* coincides with the minimum *LAC*. This point is *B* where

$$SAC_2 = SMC_2 = LAC = LMC$$

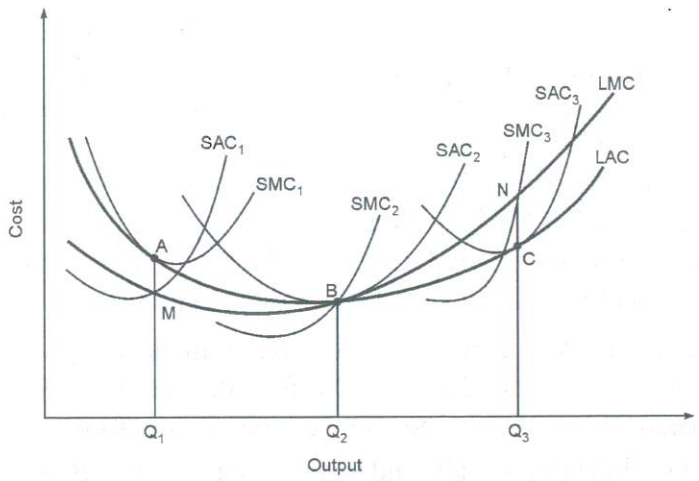


Fig. 2.24 Derivation of LMC

Conceptually, the optimum size of a firm is one which ensures the most efficient utilization of resources. Practically, the optimum size of the firm is one which minimizes the *LAC*. Given the state of technology over time, there is technically a unique size of the firm and level of output associated with the least-cost concept. In Fig. 2.24, the optimum size consists of two plants which produce OQ_2 units of a product at minimum long-run average cost (*LAC*) of BQ_2. The downtrend in the *LAC* indicates that until output reaches the level of OQ_2, the firm is of less than optimal size. Similarly, expansion of the firm beyond production capacity OQ_2, causes a rise in *SMC* and, therefore, in *LAC*. It follows that given the technology, a firm aiming to minimize its average cost over time must choose a plant which gives minimum *LAC* where $SAC = SMC = LAC = LMC$. This size of plant assures the most efficient utilization of the resource. Any change in output level—increase or decrease—will make the firm enter the area of inoptimality.

2.3.3 Modern Theory of Cost

Some economists, especially George Stigler, have questioned theoretically as well as empirically the U-shaped cost curves of 'the traditional theory of cost' and have attempted to establish that the shape of the cost curves, at least in the long run, is L-shaped. However, this point of view does not appear to have received a general recognition by the economists or as much attention as the traditional theory of cost, at least in the context of pricing theory. One possible reason is that the traditional theory of cost has a greater application to the theory of price determination and has a greater predicting power than the 'modern theory'. However, this section provides a brief description of the 'modern approach' to the theory of cost. Incidentally, like traditional theory of cost, modern theory too analyses cost-output relationships in the short-run and long-run framework.

Modern Approach to Short-run Cost Behaviour

NOTES

Like traditional theory of cost, modem theory recognizes that in the short-run,

$$TC = TFC + TVC$$

and

$$AC = AFC + AVC$$

In traditional as well as in modem theory of cost, *TFC* includes the following elements of costs:

(*i*) the salaries of administrative staff and related expenses;

(*ii*) the salaries of direct production labour paid on fixed-term basis;

(*iii*) standard depreciation allowance; and

(*iv*) maintenance cost of land and building.

This point onwards, the modem theory deviates from the traditional theory. *Traditional theory* assumes optimum capacity of a plant to be technically given (where *SAC* in minimum) and a cost-minimizing firm has no choice but to utilize the plant to its optimum capacity. On the other hand, *modern theory of cost* emphasizes that firms, in their production planning, choose a plant with flexible capacity, i.e., a plant with ***built-in reserve capacity.*** According to the modern theory, firms want to have some *reserve capacity,* as a matter of *planning* for the following reasons:

(*i*) to meet the 'seasonal' and eventual increase in demand;

(*ii*) to avoid loss of production due to break-down and repair works;

(*iii*) to have provision for meeting anticipated growth in demand;

(*iv*) to take the advantage of technology providing built-in reserve capacity;

(*v*) to build excess capacity in land and building for expansion, if required; and

(*vi*) to make full utilization of excess 'organizational and administrative' capacity.

Under these conditions, a firm does not necessarily choose a plant that gives the lowest cost of production. Instead, it chooses a plant (a set of machinery) that gives 'maximum flexibility' in production with minor adjustment in technique. For example, let us suppose that the firm has the option of setting up a plant which has an *absolute limit* to produce a commodity at the minimum cost. This absolute limit is shown by the quantity OQ_2 in Fig. 2.25. Note that if the firm chooses this plant, it can produce a maximum quantity of OQ_2 at the minimum *AFC* (average fixed cost) as shown in the figure by the boundary line BQ_2. Since there is no excess capacity, the firm cannot produce any quantity beyond OQ_2 even if demand increases and hence the firm will not be able to take the advantage of rising demand for its product. Therefore, the firm chooses a flexible plant capable of producing more than OQ_2 with minor adjustment or alternation in the production technique. For example, let us suppose that the firm chooses a flexible plant with absolute limit of output OQ_1 as shown by the boundary line AQ_1. Now let the firm anticipate a rise in demand for its product and add a small-unit machinery to its flexible plant at the output level OQ_1. As Fig 2.25 shows, with the addition of a

NOTES

small-unit machinery, firm' *AFC* increases from point *c* to point *a* on the boundary line AQ_1. But what is important from the firm's point of view is that the firm can increase its production beyond OQ_2 to meet the anticipated increase in demand. Though its *AFC* increases initially, it declines as production increases, as shown by the curve *ab* and goes below the limit set by the inflexible plant and the firm is a gainer.

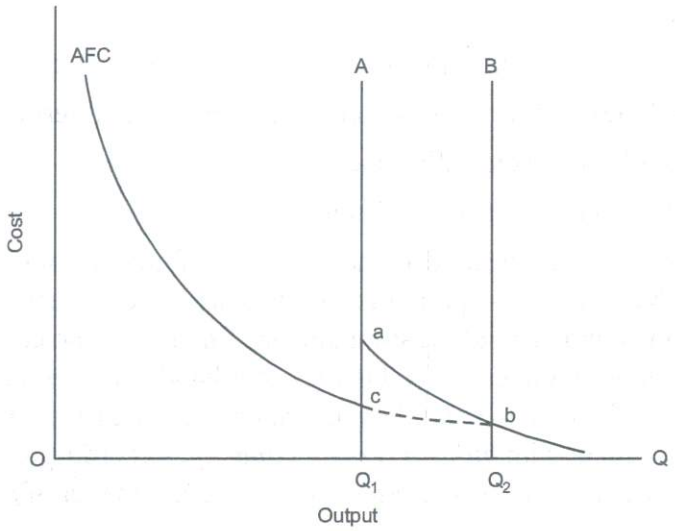


Fig. 2.25 The Built-in Reserve Capacity and AFC Curve

What Happens to the Average Variable Cost? The *average variable cost,* as in traditional theory, includes *average cost* of (*a*) direct labour, (*b*) raw materials, and (*c*) running cost of machinery. There is however a difference between the *short-run average variable cost* (*SAVC*) curves of the traditional and modem cost theories. While in traditional theory, the *SAVC* curve is U-shaped, in modem theory, it is saucer-shaped or bowl-shaped. Part (*a*) of Fig. 2.26 shows *SAVC* curve of the traditional theory, and part (*b*) shows *SAVC* curve of the modern theory.

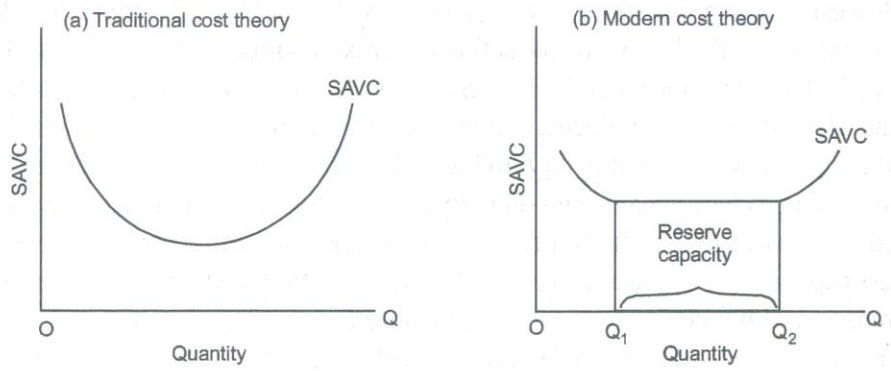


Fig. 2.26 The Traditional and Modern SAVC Curves

As part (*b*) of Fig. 2.26 shows, according to the modem theory cost, the *SAVC* remains constant over a long stretch of output between OQ_1 and OQ_2. The constancy of *SAVC* in the modem theory is attributed to the *built-in reserve*

NOTES

capacity of the flexible plant. The utilization of the built-in reserve pacity keeps the *SAVC* constant. This is an 'innovative' aspect of the modern theory of cost. In the traditional theory, there is no such built-in reserve capacity and therefore *SAVC* begins to rise once the technically efficient level of output is reached.

SAVC and SMC curves: A more important aspect of the modem theory of cost is the nature of and relationship between the *SAVC* and the *SMC* curves. The derivation of *SAVC* and the *short-run marginal cost* (*SMC*) curves is illustrated in Fig. 2.27. The *SAVC* curve is the same as in Fig. 2.27. The *SMC* curve follows the pattern of the traditional theory. The *SMC* decreases with increase in output upto a certain level.

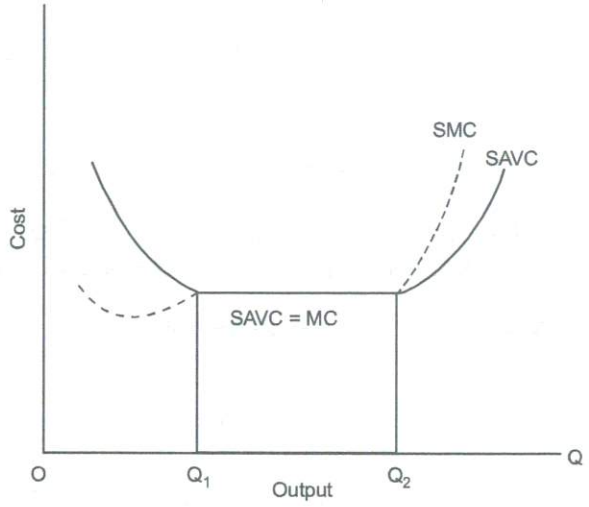


Fig. 2.27 Modern SAVC and SMC Curves

This behaviour of *SMC* curve is shown in Fig. 2.27 till the output OQ_1. However, in the range of output, between OQ_1 and OQ_2, the *SAVC* is constant. It is therefore equal to *SMC*. We know that when *SMC* begins to rise, it rises faster than *SAVC*. This behaviour of *SMC* is shown begin at output OQ_2 and continue beyond. Beyond output OQ_2, the *SMC* begins to rise and it rises faster than the *SAVC* as is the case in the traditional-theory.

Short-run Average Cost (SAC) Curves: As in traditional theory, in modem theory of cost, $SAC = AFC + SAVC$. The *AFC* includes *normal profit*. Derivation of the *SAC* curve in the modem theory is illustrated in Fig. 2.28. The *SAVC* curve (and also the *SMC* curve) is similar to one given in Fig. 2.27. For the derivation of the *SAC* curve, the *AFC* curve is added to Fig. 2.28. The *SAC* curve is the vertical summation of the *SAVC* and *AFC* curves.

As Fig. 2.28 shows, *AFC* falls continuously whereas *SAVC* decreases till output OQ_1 and remains constant between output OQ_1 and OQ_2. Therefore, a vertical summation of *AFC* and *SAVC* curves gives the *SAC* curve which declines continuously till output OQ_2. Thus, in modem theory of cost, *SAC* decreases until the *built-in reserve capacity* is fully exhausted. The *reserve capacity* is exhausted at output OQ_2.

NOTES

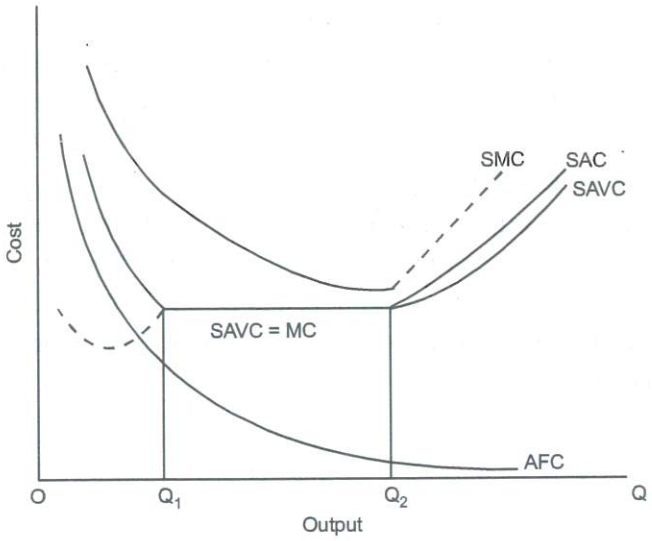


Fig. 2.28 Derivation of the Modern SAC Curve

Beyond output OQ_2, therefore, *SAC* begins to increase and goes on increasing following the increase in *SAVC* while decreasing *AFC* loses its significance.

Modern Approach to Long-run Cost Behaviour: The L-shaped Scale Curve

In respect of long-run cost behaviour, the modern theory of cost distinguishes between *production costs* and *managerial costs*. Both these costs are variable in the long run. The behaviour of these costs determines the shape of the long-run average cost curve (*LAC*). According to the modern theory, the long-run *LAC* is broadly L-shaped. Let us now look at the behaviour of the *production* and *managerial costs* in the long run and how they determine the shape of the *LAC* curve.

Production Cost Behaviour. Production cost decreases steeply in the beginning with the increase in the scale of production but the rate of decrease slows down as the scale increases beyond a certain level of production. The decrease in the production costs is caused by the technical economies which taper off when the scale of production reaches its *technical optimum scale.* Nevertheless, some economies of scale are always available to the expanding firms due to (*i*) 'decentralization and improvement in skills'; and (*ii*) decreasing cost of repairs per unit of output. In addition, in case of multi-product firms producing some of their raw materials and equipments have economies in material cost compared to purchases made from outside.

Managerial Cost Behaviour. The modern theory of cost assumes that, in modern management technology, there is a fixed managerial or administrative set up with a certain scale of production. When the scale of production increases, management set up has to be accordingly expanded. It implies that there is a link between the scale of production and the cost of management. According to the modern theory, the managerial cost first decreases but begins to increase as the scale of production is expanded beyond a certain level.

What Makes LAC L-shaped? *The net effect of decreasing production cost and increasing managerial cost determines the shape of the long-run average cost (LAC).* Recall that production cost continues to decrease, though slowly beyond a certain scale of production and managerial cost too decreases initially but rises later. In the initial stage of production, therefore, the *LAC* decreases very steeply. Beyond a certain scale of production, however, while production cost continues to decline, management cost begins to rise. According to the modern theory of cost, the rise in managerial cost is more than offset by the decrease in the production costs. Therefore, the *LAC* continues to fall but very slowly. *In case the decrease in production cost* is *just sufficient to offset the rise in the managerial cost, the LAC becomes constant.* This makes *LAC* an L-shaped curves.

Derivation of the LAC Curve: The derivation of the *LAC* curves is illustrated in Figs. 2.29 and 2.30. Fig. 2.29 shows the decreasing *LAC* curve. Let us suppose that, given the technology, the optimum scale of production consists of four plants and *SAC* curves from SAC_I to SAC_4 in Fig. 2.29 represent the addition of four plants to the production scale in each period of time. Clower and Due have found that firms use 'normally' only 2/3 to 3/4 of the plant size. This is called 'load factor'. The *load factor* is the "ratio of average actual rate of use to the capacity or best rate of use, and this load factor will generally be smaller than one". The points *A, B, C* and *D* on the *SACs* mark the 'load factor' in case of each plant, respectively—it may be any value between 2/3 and 3/4 of the plant size. By drawing a curve through the 'load-factor' points, we get the *LAC* curve. If there is a larger number of plants, we will get much larger number of 'load factor' points and draw a smooth *LAC* curve as shown in Fig. 2.29.

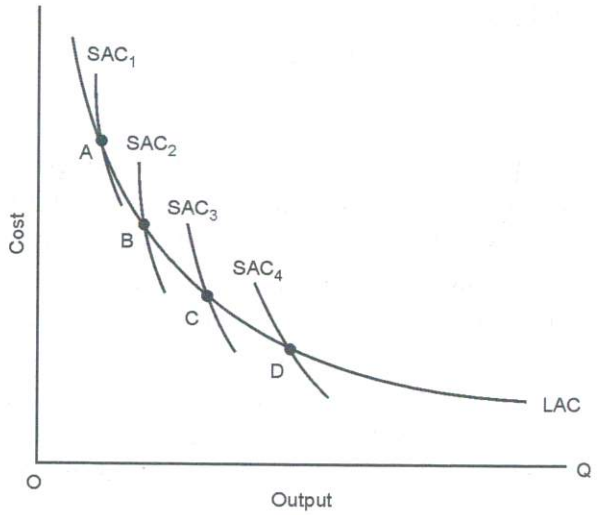


Fig. 2.29 Derivation of LAC Curve in Modern Theory

To compare the *LAC* of the modern and traditional theories of cost, two points need to be noted: (*i*) the *LAC* curve of the modern theory does not show the tendency to turn up even at a very large scale of production whereas the traditional *LAC* curve does turn up; (*ii*) unlike traditional *LAC* forming an *envelope curve,* modern *LAC* intersects the *SACs.*

NOTES

If scale of production has a *minimum optimal scale of plant,* as shown by output level *OQ* in part (*b*) of Fig. 2.30, all economies of scale are achieved at output *OQ* and the *LAC* becomes constant even if scale of production is expanded. In this case, the *LMC* lies below the *LAC* till the *minimum optimal scale of plant* is reached, as shown in part (*a*) of Fig. 2.30. When the firm operates in the range of no-seale-economies, i.e., beyond output *OQ* in part (*b*) of Fig. 2.30, the *LAC* becomes constant and the *LAC* curve takes the shape of a horizontal line. Both the parts (declining and constant) of the *LAC* curve put together make it roughly L-shaped.

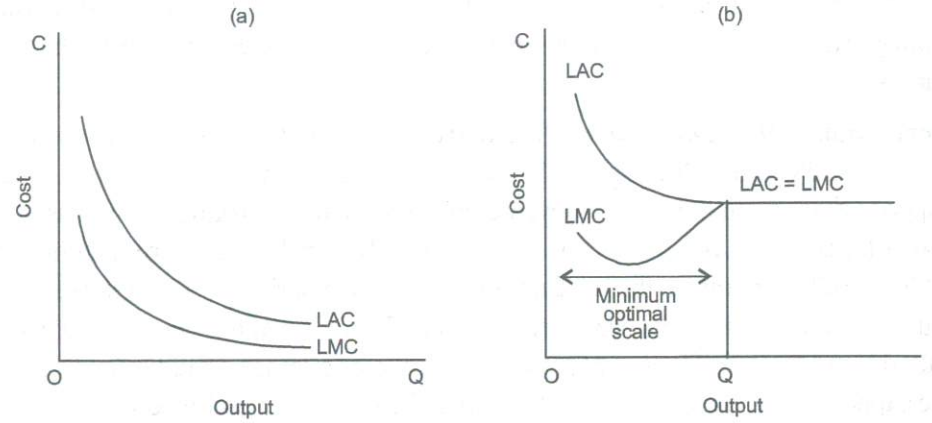


Fig. 2.30 Derivation of the L-shaped LAC Curve

From practical point of view, the modern *LAC* curve is regarded to be more realistic. But from analytical and prediction point of view, the traditional cost curves still hold the ground firmly. In fact, the so called 'modern theory of cost' is a modification of the traditional theory on the basis of empirical data in some manufacturing industries of some countries.

Learning Curve

In many industries, some firm specialize in one particular good or service and continue to produce it over time. As a result, they gain additional knowledge, expertise and efficiency and are able to make technological advances in the field of production. The economists have found that these factors result in continuous decline in their average cost with increase in output. A curve drawn to show the continuous decline in the average cost is called *learning curve.* The *learning curve* is, in fact, the long-run average cost curve showing a continuous decline in average cost with increase in output over time, as shown in Fig. 2.31(a) and (b). The shape of the learning curve depends on the nature of the product. As shown by Mansfield *et. al.* Fig. 2.31(a) shows the learning curve for the production of portable turbine and Fig. 2.31(b) shows the learning curve for production of optical equipment. The continuous decrease in long-run average cost, as shown in Fig. 2.31, results from the continuous technological improvement, on the one hand, and knowledge, expertise and efficiency gained by the firms by producing the same commodity over time, on the other.

It is ***important*** to note here that average cost is estimated by dividing ***cumulative total cost*** by ***cumulative total output*** increasing at constant rate.

This implies that when firm continues to produce the same commodity over time, its cumulative cost increases less than proportionate to a constant increase in output.

NOTES

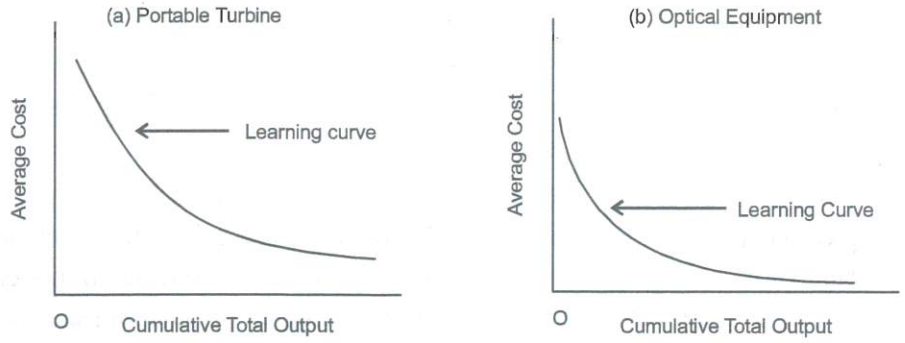


Fig. 2.31 The Learning Curve

The learning curve is generally expressed by the following form of *AC* function.

$$AC = BQ^a$$

where *AC* average cost; *Q* is output; *B* is the cost of the first unit of output; and power *a* is a *negative* constant.

In its double logarithmic form, the *AC* function for learning curve can be expressed as

$$\text{Log } AC \quad = \log B + a \log Q$$

In its logarithmic form of the learning cost function, *a* gives the slope of the learning curve. The numerical values of parameters *B* and *a* are estimated by regression techniques by using historical data of cumulative output and cost.

Many firms in both manufacturing and service sectors have adopted pricing strategy based on learning curve, mainly by the firms manufacturing airplanes, ships, semi-conductor chips, domestic appliances, refined petroleum products and power plants. The learning curve is used to assess and forecast the future requirements for inputs including manpower, raw materials, and machinery.

Check Your Progress

4. Write any two examples of explicit costs.
5. How can we calculate average cost?
6. State one difference between the traditional and modern theory of cost.
7. What does the average variable cost under the traditional theory of cost consist of?

2.4 ANSWERS TO 'CHECK YOUR PROGRESS'

1. A fixed input is one whose quantity remains constant for a certain level of output, e.g., plant, building and machinery.
2. The three kinds of isoquants are linear isoquants, L-shaped isoquants and kinked or linear programming isoquant.

NOTES

3. Cobb-Douglas production function in its general form, $Q = K^aL^{1-a}$ indicates that at zero cost, there will be zero production.
4. The payments on account of wages, salaries, utilities, interest, purchase of materials, licence fee, insurance premium and depreciation charges are the examples of explicit costs. (Write any two examples.)
5. Average cost is obtained simply by dividing the total cost (TC) by the total output (Q), i.e., TC/Q = average cost.
6. Traditional theory assumes optimum capacity of a plant to be technically given (where SAC in minimum) and a cost-minimizing firm has no choice but to utilize the plant to its optimum capacity. On the other hand,modern theory of cost emphasizes that firms, in their production planning, choose a plant with flexible capacity, i.e., a plant with built-in reserve capacity.
7. The average variable cost, as in traditional theory, includes average cost of (a) direct labour, (b) raw materials, and (c) running cost of machinery.

2.5 SUMMARY

- An input is simply anything which the firm buys for use in its production or another process. An 'output' is any commodity which the firm produces or processes for sale.
- A fixed input is one whose quantity remains constant for a certain level of output, e.g., plant, building and machinery.
- A variable input is defined as one whose quantity changes with change in output.
- In the short-run, production of a commodity can be increased by increasing the use of variable inputs, like labour and raw materials.
- The production function describes the technological relationship between inputs and outputs in physical terms.
- A production function may take the form of a schedule or table, a graphed line or curve, an algebraic equation or a mathematical model.
- It is a universal fact that increasing quantity of the inputs (labour and capital) leads to an increase in the output and vice versa.
- The short-run laws of production are also known as the laws of diminishing returns, the laws of variable proportions, and the laws of returns to variable inputs.
- The law of diminishing returns states that if more and more units of a variable input are applied to a given quantity of fixed inputs, the total output may initially increase on increasing rate, but beyond a certain level, output increases at a diminishing rate.
- An isoquant curve is a locus of points representing the various combinations of two inputs—capital and labour—yielding the same output.

NOTES

- A convex isoquant is the most widely used isoquant in traditional economic theory.
- The shape of an isoquant, in fact, depends on the assumption regarding the degree of substitutability between the factors in the production function.
- The elasticity of substitution (s) is formally defined as the percentage change in the capital- labour ratio (K/L) divided by the percentage change in the marginal rate of technical substitution (MRTS).
- The three plausible reasons for increasing returns to scale are Technical and managerial indivisibilities, higher degree of specialization and dimensional relations.
- Actual costs are those which are actually incurred by the firm in payment for labour, material, plant, building, machinery, equipment, travelling and transport, etc.
- Business costs include all the expenses which are incurred in carrying out a business.
- Full cost includes two other costs: opportunity cost and normal profit.
- Cost function is a symbolic statement of the technological relationship between cost and output.
- The theory of cost deals with the behaviour of cost in relation to a change in output. In other words, the cost theory deals with cost-output relations.
- Depending on whether cost analysis pertains to short-run or to long run, there are two kinds of cost functions: (*i*) short-run cost functions, and (*ii*) long-run cost functions,
- The basic analytical cost concepts used in the analysis of cost behaviour are Total, Average and Marginal costs.
- The cost-output relations are determined by the cost function and are exhibited through cost curves. The shape of the cost curves depends on the nature of the cost function.
- The concept of marginal cost (*MC*) is useful particularly in economic analysis.
- By definition, long-run is a period in which all the inputs–specifically, labour and capital–become variable. The variability of inputs is based on the assumption that in the long-run supply of all the inputs, including those held constant in the short run, becomes elastic.
- Some economists, especially George Stigler, have questioned theoretically as well as empirically the U-shaped cost curves of 'the traditional theory of cost' and have attempted to establish that the shape of the cost curves, at least in the long run, is L-shaped.
- The *average variable cost,* as in traditional theory, includes *average cost* of (*a*) direct labour, (*b*) raw materials, and (*c*) running cost of machinery.

NOTES

- A more important aspect of the modem theory of cost is the nature of and relationship between the *SAVC* and the *SMC* curves.
- In respect of long-run cost behaviour, the modern theory of cost distinguishes between *production costs* and *managerial costs.*
- Production cost decreases steeply in the beginning with the increase in the scale of production but the rate of decrease slows down as the scale increases beyond a certain level of production.
- The net effect of decreasing production cost and increasing managerial cost determines the shape of the long-run average cost (LAC).
- In many industries, some firm specialize in one particular good or service and continue to produce it over time. As a result, they gain additional knowledge, expertise and efficiency and are able to make technological advances in the field of production.
- The learning curve is used to assess and forecast the future requirements for inputs including manpower, raw materials, and machinery.

2.6 KEY TERMS

- **Production:** It refers to the process by which tangible and intangible inputs or factors of production are converted or transformed into useful tangible or intangible output.
- **Production function:** It describes the technological relationship between inputs and outputs in physical terms.
- **Law of diminishing returns:** It states that if more and more units of a variable input are applied to a given quantity of fixed inputs, the total output may initially increase on an increasing rates, but beyond a certain level, output increases at a diminishing rate.
- **Law of returns to scale:** It explains the behaviour of the total output in response to changes in the scale of the firm, i.e., in response to a simultaneous and proportional increase in all the inputs it uses.
- **Marginal cost:** It is the additional cost incurred for the production of an additional unit of output.
- **Average cost:** It is a cost accounting term that is sometimes referred to as unit cost or weighted average cost.
- **Learning curve:** It is the long-run average cost curve showing a continuous decline in average cost with increase in output over time.
- **Isoquant:** It is a graph of all possible combinations of inputs that result in the production of a given level of output.
- **Explicit Costs:** These costs fall under actual or business costs entered in the books of accounts.
- **Fixed Costs:** These costs are fixed in volume for a certain given output.

NOTES

2.7 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. What are short-run production conditions and long-run production conditions?
2. How can a firm apply the law of diminishing returns?
3. Differentiate between the follows:
   (a) Short-run and long-run costs
   (b) Business cost and full costs
4. State the formula for cubic cost function.
5. Write a short note on output optimization in the short-run.
6. State the difference between short-run average variable cost curves of the traditional and modern cost theories.
7. Mention the uses of the learning curve in economics.

Long Answer Questions

1. Describe the nature and forms of a production function.
2. Explain the laws of variable proportions with the help of an example.
3. Discuss the properties of isoquant curves.
4. Express the general criteria for the least-cost input combination in both physical and value terms.
5. Analyse the causes of increasing returns to scale.
6. Explain linear cost function with the help of diagrams.
7. Discuss the relationship between cost curves and the law of diminishing returns.
8. Discuss the 'innovative' aspect of the modern theory of cost.
9. Explain the modern approach to long-run cost behaviour.

2.8 FURTHER READING

Dwivedi, D.N. 2015. *Managerial Economics,* Eighth Edition. New Delhi: Vikas Publishing.

Keat, Paul G. and Philip, K.Y. Young. 2003. *Managerial Economics: Economics Tools for Today's Decision Makers,* Fourth Edition. Singapore: Pearson Education.

Keating, B. and J.H. Wilson. 2003. *Managerial Economics: An Economic Foundation for Business Decisions,* Second Edition. New Delhi: Biztantra.

NOTES

Peterson, H.C. and W.C. Lewis. 1999. *Managerial Economics*, Fourth Edition. Singapore: Pearson Education.

Salvantore, Dominick. 2001. *Managerial Economics in a Global Economy*, Fourth Edition. Australia: Thomson-South Western.

Thomas, Christopher R. and Maurice S. Charles. 2005. *Managerial Economics: Concepts and Applications*, Eight Edition. New Delhi: Tata McGraw-Hill.

NOTES

UNIT 3 MARKET STRUCTURE AND PRICING STRATEGIES

Structure

3.0 INTRODUCTION

The nature and degree of competition in the market for goods and services collectively is referred to as market structure. The structures of market including goods and services are determined by the nature of competition prevailing in a particular market. There are a number of determinants of market structure for a specific good. They are (i) number and nature of sellers, (ii) number and nature of buyers and (iii) nature of the product. The four major types of market structures are perfect competition, oligopoly market, monopoly market, and monopolistic competition.

This unit will explain the market structure and its types. In a free market, the forces of demand and supply determine the prices. In the time of high competition, firms follow different pricing strategies when setting prices for their products or services with the aim of profit maximization. Senior executives of a company must identify the company's pricing position, pricing segment, pricing capability and their competitive pricing reaction strategy to determine the most effective pricing strategy for a company. Pricing strategies and tactics vary from company to company, as well as across countries, cultures, industries and over time, as industries and markets mature and the economy evolves. In addition, it will discuss various pricing strategies in detail.

NOTES

3.1 OBJECTIVES

After going through this unit, you will be able to:

- Explain the concept and types of market structures
- Describe price-out determination under different market structures
- Discuss various pricing strategies

3.2 PRICE-OUTPUT AND MARKET STRUCTURE

The modern theory of price came into existence during the 1930s with Joan Robinson's *The Economics of Imperfect Competition*, and Edwin H Chamberlin's *The Theory of Monopolistic Competition,* both written independently in 1933. Earlier, the theory of price determination was in the form of the 'Theory of Value' attributed to Alfred Marshall and his *Principles of Economics*. The theory of value propounded by Marshall on the assumptions of perfect competition and a static equilibrium system was regarded to provide answers to all questions regarding price and output determination. The existence of perfect competition was, however, challenged by Piero Sraffa. He showed that perfect competition was not logically consistent with partial equilibrium analysis. This led to the abandonment of the assumption of perfect competition. Joan Robinson and Chamberlin developed independently the theory of imperfect competition and theory of monopolistic competition, respectively. Joan Robinson and Chamberlin demonstrated that price and output are determined by individual decisions under the condition of imperfect competition. They had, however, retained the earlier assumption of profit maximization. This assumption was later challenged, and many new theories of firms were suggested. None of the theories has, however, received a universal acceptance.

Let us the theory of the firm with the analysis of price and output determination in a perfectly competitive market.

In contrast to perfect competition, monopoly is an extremely opposite form of a product market. In case of perfect competition, the number of sellers is so large that no one of them has any power whatsoever to influence the market price. A monopoly firm, on the other hand, has the sole power to influence the market price. While under perfect competition, no seller can afford to discriminate between the buyers of different categories, the monopolists practice price discrimination as a matter of policy.

Until the early 1920s, 'the theories of perfect competition and monopoly constituted "classical" microeconomic theory from Marshall to Knight.' In the late 1920s and the early 1930s, however, economists expressed their dissatisfaction with perfect competition and pure monopoly models as tools of analysing real-world business behaviour. Piero Shraffa was one of those who first pointed out the limitations of perfect competition and pure monopoly models. He was followed by Hotelling and Zeuthen who urged that neither perfect competition nor monopoly represents the real-world business phenomenon and that the real business world

NOTES

conforms to a middle ground between monopoly and perfect competition, i.e., *monopolistic competition*. The two most notable contributions that broke new ground in the price theory were made by Joan Robinson and Edward H. Chamberlin.

Oligopoly is similar to monopolistic competition. However, oligopoly differs from monopolistic competition in the following two ways. (i) under monopolistic competition, the number of sellers is very large, while under oligopoly it is small, and (ii) under oligopoly competition between firms is much more intensive as compared to monopolistic competition.

Market Structure

Apart from the firm's objective, another factor that plays an important role in firm's choice of price and output is the market structure. The term 'market structure' refers to the organizational features of an industry that influence a firm's behaviour in its choice of price and output. Market structure is an economically significant feature of the market. It affects the behaviour of firms in respect of their production and pricing behaviour.

The market structure is classified on the basis of organizational features of the industry, more specifically, on the basis of degree of competition among the firms. In general, the organizational features include the number of firms, distinctiveness of their products, elasticity of demand and the degree of control over the price of the product.

The market structure is generally classified on the basis of the degree of competition as follows:

(i) Perfect competition

(ii) Imperfect competition
  (a) Monopolistic competition
  (b) Oligopoly with and without product differentiation
  (c) Duopoly

(iii) Monopoly

The basic features of these kinds of market are summarized as follows:

Table 3.1 *Kinds of Market Structure*

Type of Market	*No. of Sellers*	*Nature of Product Firms*	*Firm's Over Price Control*
(i) **Perfect Competition**	Very large	Homogeneous (wheat sugar, vegetables....)	None
(ii) **Imperfect Competition**			
(a) **Monopolistic Competition**	Many (most retail trade)	Real or perceived difference in product	Some
(b) **Oligopoly**	Few	(i) Product without differentiation, e.g. aluminium, steel, and chemicals, etc. (ii) Differentiated products (tea, TV	Some

Contd...

NOTES

		Refrigerator, toothpastes, soaps, detergents, automobiles)	
Monopoly	Single	Products without close substitutes, like gas, electricity and water supply	Full but usually regulated

(i) **Perfect competition:** Perfect competition is a market situation in which a large number of producers offer a homogeneous product to a very large number of buyers of the product. The number of sellers is so large that each seller offers a very small fraction of the total supply and therefore has no control over the market price. Likewise, the number of buyers is so large that each buyer buys an insignificant part of the total supply and has no control over the market price. Both buyers and sellers are 'price-takers', not 'price-makers'. The price of a commodity is determined in this kind of market by the market demand and market supply. Each seller faces a horizontal demand curve (with $e = \infty$), which implies that a seller can sell any quantity at the market-determined price. Each firm is in competition with so many firms that there is virtually no competition. This kind of market is, however, more of a hypothetical nature rather than being a common or realistic one. Some examples of a perfectly competitive market include stock markets, and vegetable markets, wheat and rice mandis, where goods are sold by auction.

(ii) **Imperfect competition:** Perfect competition, in strict sense of the term, is a rare phenomenon. In reality, markets for most goods and services have imperfect competition. Imperfect competition is said to exist when a number of firms sell identical or differentiated products with some control over the price of their products. Barring a few goods like shares and vegetable markets, its market is imperfect. In spite of a large number of dealers *(arhatias)* in the wheat market, the Food Corporation of India is the biggest buyer and seller of wheat in India, with a great degree of control over wheat prices.

Imperfect competition creates two different forms of markets with different number of producers and with different degrees of competition, classified as (a) monopolistic competition, (b) oligopoly, and (c) monopoly.

(a) **Monopolistic competitions:** Monopolistic competition is a kind of market in which a large number of firms supply differentiated products. The number of sellers is so large that each firm can act independently of others, without its activities being watched and countervailed by others. It is extremely difficult to keep track of competitors' strategy. Moreover, but also it is not of any avail. In this respect, it is similar to perfect competition. It differs from perfect competition in which the products under monopolistic competition are somewhat differentiated, whereas they are identical under perfect competition. There is free entry and free exit.

(b) **Oligopoly:** Oligopoly is an organizational structure of an industry in which a small number of firms supply the entire market, each seller having a considerable market share and control over the price. Most industries in our country are oligopolistic. A small number of companies supply tea, medicines, cosmetics, refrigerators, TV and VCRs, cars, trucks, jeeps, and so on. The producers of all these goods have some control over the price of their products. Their products are somewhat differentiated, at least made to look different in the consumers' perception. Therefore, demand curve for their product has high elasticity but less than infinity, unlike under perfect competition.

(c) **Monopoly:** Monopoly is the market of a single seller with control over his price and output. It is antithesis of perfect competition. Absolute monopolies are rare these days. They are found mostly in the form of government monopolies in public utility goods, e.g., electricity, radio broadcasting, water, rail and postal services.

Why markets are imperfect?

Imperfect competition arises mainly from the barriers to entry. Barriers to entry are created by the following factors:

(i) The large-sized firms which enjoy economies of scale can cut down their prices to the extent that can eliminate new firms or prevent their entry to the industry, if they so decide.

(ii) In some countries, like India, licencing policy of the government creates a barrier for the new firms to enter an industry.

(iii) Patenting of rights to produce a well-established product or a new brand of a commodity prevents new firms from producing that commodity.

(iv) Sometimes entry of new firms to an industry is prevented by a law, with a view to enabling the existing ones to have economies of scale so that prices are low.

In the following subsections you will learn about the different market structures and the price-output relationships prevalent in them.

3.3 PRICE-OUTPUT UNDER PERFECT COMPETITION

A perfectly competitive market is characterized by complete absence of rivalry among individual firms. In fact, under perfect competition, as conceived by the economists, competition among the individual firms is so widely dispersed that it amounts to no competition. Perfect competition is characterized by the following assumptions:

1. **Large number of buyers and sellers:** Under perfect competition, the number of sellers is assumed to be so large that the share of each seller in the total supply of a product is so small that no single firm can influence the market price by changing its supply. Similarly, the number of buyers is so large that the share of each buyer in the total demand is so small that no

NOTES

single buyer or a group of buyers can influence the market price by changing buyer's individual or group demand for a product.

2. **Homogeneous product:** The commodities supplied by all the firms of an industry are assumed to be *homogeneous* or approximately identical. Homogeneity of the product implies that buyers do not distinguish between products supplied by the various firms of an industry. Product of each firm is regarded as a perfect substitute for the products of other firms. Hence, no firm can gain any competitive advantage over the other firms. This assumption limits the power of any firm to charge a price which is even slightly higher than the market price.
3. **Perfect mobility of factors of production:** Another important characteristic of perfect competition is that the factors of production (especially, labour and capital) are freely mobile between the firms. Labour can freely change the firms as there is no barrier on labour mobility—legal, language, climate, skill, distance or otherwise. There is no trade union. Capital can also move freely from one firm to another. No firm has any kind of monopoly over any industrial input. This assumption guarantees that factors of production—labour, capital, and entrepreneurship—can enter or quit a firm or the industry whenever it is found desirable.
4. **Free entry and free exit:** There is no legal or market barrier on the entry of new firms into the industry. In addition, there is no restriction on the exit of the firms from the industry. That is, a firm may enter the industry and quit it at its will. Thus, when normal profit of the industry increases, new firms enter the industry, and if profits decrease and better opportunities are available, firms leave the industry.
5. **Perfect knowledge about the market conditions:** There is perfect knowledge about the market conditions. All the buyers and sellers have full information regarding the prevailing and future prices and the availability of commodity. As Marshall puts it, '... though everyone acts for himself, his knowledge of what others are doing is supposed to be generally sufficient to prevent him from taking a lower or paying a higher price than others are doing.' Information regarding market conditions is available free of cost. There is no uncertainty.
6. **No government interference:** Government does not interfere in any way with the functioning of the market. There are no taxes or subsidies; no licencing system, no allocation of inputs by the government, or any kind of other direct control. That is, the government follows the *free enterprise* policy. Where there is intervention by the government, it is intended to correct the market imperfections.
7. **Absence of collusion and independent decision making:** Perfect competition assumes that there is no collusion between the firms, i.e., they are not in league with one another in the form of *guild* or *cartel.* In addition, buyers are not in collusion among themselves. There are no consumers' associations, etc. This condition implies that buyers and sellers take their decisions independently and they act independently.

NOTES

Perfect vs pure competition

Sometimes a distinction is made between perfect competition and pure competition. The difference between the two is a matter of degree. While 'perfect competition' has all the features mentioned above, 'pure competition' does not assume *perfect mobility* of factors and *perfect knowledge*. That is, *perfect competition* less *perfect mobility and knowledge* is pure competition. 'Pure competition' is 'pure' in the sense that it has absolutely no element of monopoly.

The perfect competition, as characterized above, is considered as a rare phenomenon in the real business world. However, the actual markets that approximate the conditions of perfectly competitive market include the security markets for stocks and bonds and agricultural markets like local vegetable markets. Despite its limited scope, a perfect competition model has been the most popular model used in economic theories due to its analytical value.

Equilibrium of a Firm

According to the traditional theory of the firm, a firm is in equilibrium when its profit is maximum. Maximization of profits depends on the revenue and cost conditions. Revenue and cost conditions vary according to whether the period under reference is short or long. The equilibrium of the firm under *short-run* and *long-run conditions* is discussed below.

A. Equilibrium of a firm in the short-Run

Here, short-run refers to a period of time during which (i) price of the product is given in the market and the firm can sell any quantity at the prevailing price; (ii) plant size of the firm is given; and (iii) the firm is faced with given short-run cost curves.

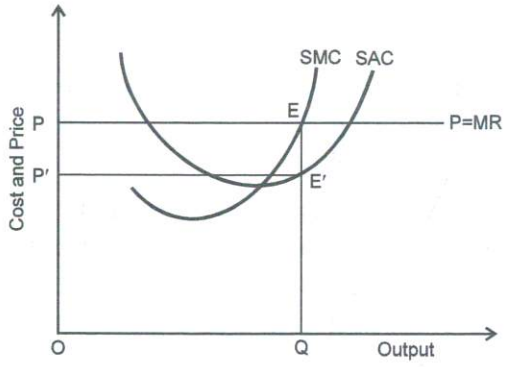


Fig. 3.1 *Equilibrium of the Firm*

Where,

AR = Average revenue

MR = Marginal revenue

The firm's equilibrium in the short-run is illustrated in Figure 3.1. Price of a commodity is fixed by the market forces in a perfectly competitive market. The firms, therefore, face a straight-line, horizontal demand curve, as shown by the line $P = MR$. It implies that price equals marginal revenue, i.e., $AR = MR$. At firm's sort-run average and marginal cost curves are shown by *SAC* and *SMC*, respectively.

It can be seen from Figure 3.1 that *SMC* curve intersects the $P = MR$ line at point *E*, from below, where $SMC = MR$. A perpendicular drawn from point *E* to the output axis determines the equilibrium output at *OQ*. It means that output *OQ* meets both the first and second order conditions of profit maximization. At output *OQ*, therefore, profit is maximum. The output *OQ* is thus the equilibrium output. At this output, the firm is in equilibrium and is making maximum profit. The firm's maximum pure profit is shown by the area $PEE'P'$ which equals $PP' \times OQ$, where PP' is the per unit abnormal profit at output *OQ*.

Does a firm always make profit in the short-run?

In the short-run equilibrium, a firm may not always make profits. In the short-run, it may earn just a normal profit or even make losses. Whether a firm makes abnormal profits, normal profits, or losses depends on its cost and revenue conditions. If its short-run average cost (*SAC*) is below the price ($P = MR$) at equilibrium, as shown in Figure 3.1, the firm makes abnormal or pure profits. If its *SAC* is tangent to $P = MR$ [Figure 3.2(a)], the firm makes only normal profit as it covers only its *SAC* which includes normal profit. But, if its *SAC* falls above the price $(P = MR)$, the firm makes losses [Figure 3.2(b)]. The total loss is shown by the area $PP'EE'$ $(= P'P \times OQ)$, while per unit loss is $PP' = EE'$.

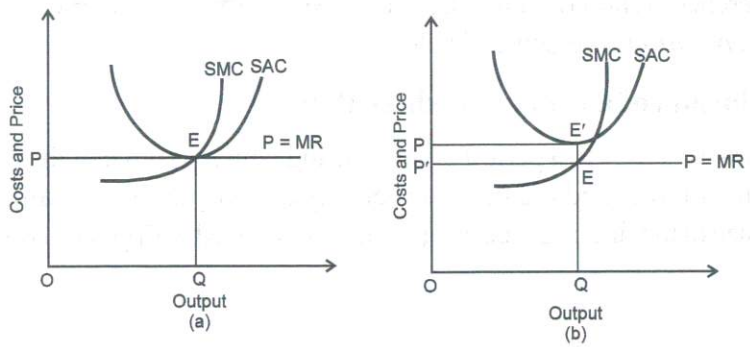


Fig. 3.2 Short-run Equilibrium of a Firm

Shut-down or close-down point

In case a firm is making loss in the short-run, it must minimize its losses. In order to minimize its losses, it must cover its short-run average variable cost (*SAVC*). The behaviour of short-run average variable cost is shown by the curve *SAVC* in Figure 3.3. A firm unable to recover its minimum *SAVC* will have to close down. The *SAVC* is minimum, where it equals the *MC*. The *SMC* intersects *SAVC* at its minimum level, as shown in Figure 3.3.

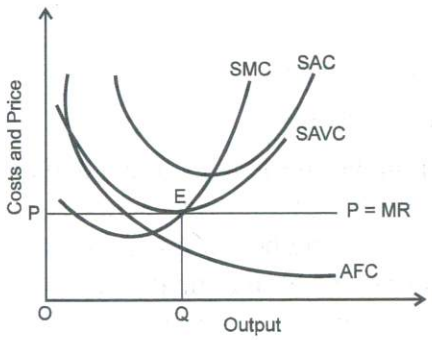


Fig. 3.3 Shut-down Point

Another condition that must be fulfilled is $P = MR = SMC$. That is, for loss to be minimum, $P = MR = SMC = SAVC$. This condition is fulfilled at point E in Figure 3.3. Point E denotes the 'shut-down point' or 'breakdown point' because at any price below OP, it pays firms to close down as it minimizes its losses.

(i) Derivation of supply curve of the firm

The supply curve of an individual firm is derived on the basis of its equilibrium output at different levels of the price. The equilibrium output is determined by the intersection of MR and MC curves. The derivation of supply curve of a firm is shown in Figure 3.4(a) and (b). The equilibrium level of output in the short-run is given at point M—the point of intersection between MC and $SAVC$. This is 'breakdown point' which gives the minimum supply of the firm in the short-run. The equilibrium level of output at this point is OQ_1. Let us suppose that price increases to OP_2. As a result, the equilibrium point shifts to R and output increases to OQ_2. Let the price further increase to OP_3 so that equilibrium output increases to OQ_3. When price rises to OP_4, the equilibrium output rises to OQ_4. It may thus be concluded that as price increases, firm's supply goes on increasing, which means that there is positive relation between price and supply. The price and output information contained in Figure 3.4(a) is presented in the form of a supply curve, MS, as shown in Figure 3.4(b).

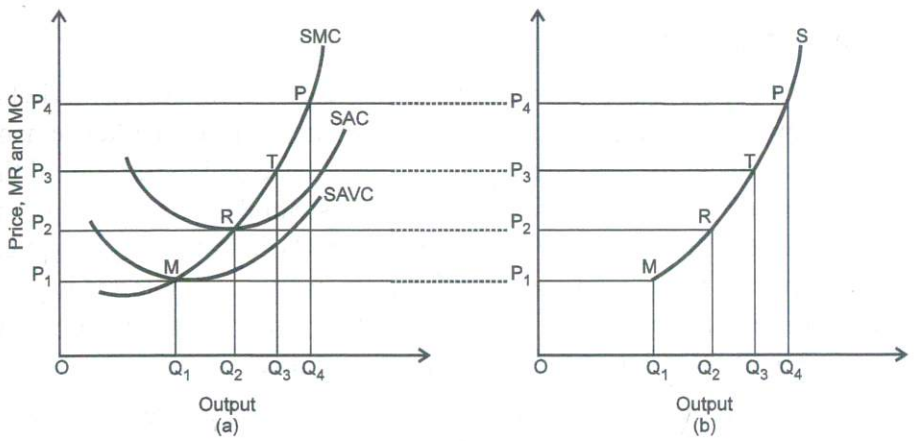


Fig. 3.4 *Derivation of Firm's Supply Curve*

(ii) Derivation of supply curve of the industry

Just as the market demand curve is a horizontal summation of individual demand curves, the industry supply curve or market supply curve is the horizontal summation of the supply curves of the individual firms. If cost curves of the individual firms of an industry have identical shape, their individual supply curves would also be identical. In that case, industry supply curve can be obtained by multiplying the individual supply at various prices by the number of firms. In the short-run, however, the individual supply curves may not be identical. If so, the market supply curve can be obtained by summing horizontally the individual supply curves. Let us suppose that there are only two firms having their individual supply curves S_1 and S_2, as shown in Figure 3.5(a). At price OP_1, the industry supply equals $P_1A + P_1B$, which equals P_1M in Figure 3.5(b).

NOTES

Similarly, at price OP_2, the industry supply equals $P_2C + P_2C$ or $2(P_2C)$, which equals P_2N in Figure 3.5(b). In the same way, point T is located. By joining the points M, N and T, we get the market or industry supply curve, SS'. The market supply curve so derived is used to show the determination of market price.

Now let us study price determination in a perfectly competitive market.

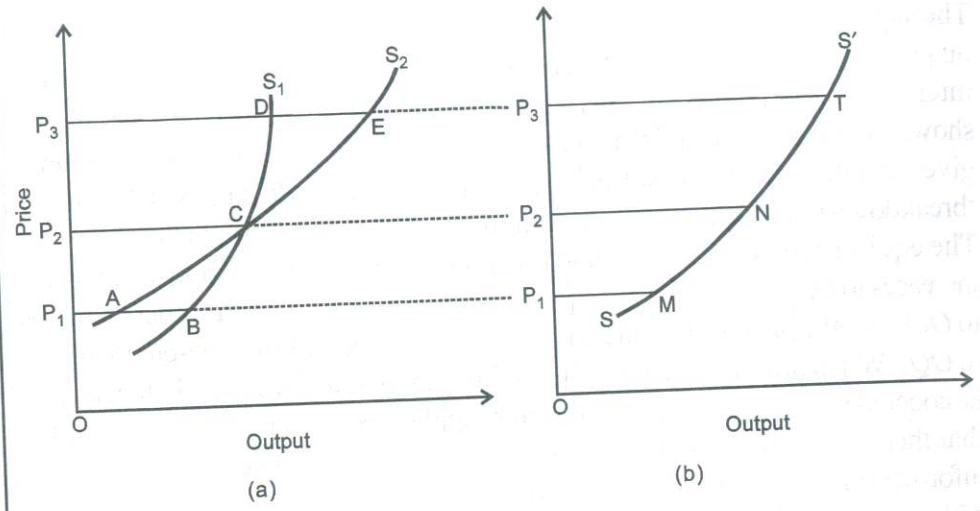


Fig. 3.5 Derivation of Industry Supply Curve

Price and output determination

Under perfect competition, market price in a perfectly competitive market is determined by the market forces, viz., demand and supply. Here, market demand refers to the demand for the industry as a whole. It is equal to the sum of the quantity demanded by individuals at different prices. Similarly, market supply is the sum of quantity supplied by the individual firms in the industry at a given price. The market price is therefore determined for the industry as a whole and is given for each individual firm and for each buyer. Thus, every seller in a perfectly competitive market is a 'price-taker', not a 'price-maker'.

In a perfectly competitive market, therefore, the main problem of a firm is not to determine the price of its product but to find its output at the given price so that profit is maximized.

The role of market forces and the mode of price determination depends on the time taken by supply position to adjust itself to the changing demand conditions. Price determination is analysed under three different time periods: (i) market period or very short-run; (ii) short-run; and (iii) long-run.

(i) Price determination in very short-Run

The market period or very short-run refers to a time period in which quantity supplied is absolutely fixed or, in other words, supply response to change in price is nil. In the market period, therefore, the total output of the product is fixed. Each firm has a given quantity of commodity to sell. The aggregate supply of all the firms makes the market supply. The supply curve is *perfectly inelastic*, as shown by line SQ in Figure 3.6. In this situation, price is determined entirely by the demand conditions. For instance, suppose that the number of marriage houses (or tents)

available per month in a city is given at OQ (Figure 3.6), so that the supply curve takes the shape of a vertical straight line SQ. Let us also suppose that the monthly demand curve for marriage houses is given by the demand curve, D_1. Demand and supply curves intersect each other at point M, determining the rental at MQ. Let us now suppose that during a particular month, demand for marriage houses suddenly increases because a relatively large number of parents decide to celebrate the marriage of their daughters and sons due to, say, non-availability of auspicious dates for some time to come. Consequently, the demand curve shifts upward to D_2. The demand curve D_2 intersects the supply curve at point P. The equilibrium rate of rental is thus determined at PQ. This becomes parametric price for all the buyers. Note that the rise in the rental from MQ to PQ is caused by the upward shift in the demand curve and that market supply curve remains perfectly inelastic in the market period. The other example of very short-run markets may be of perishable commodities like fish, milk and vegetable and non-perishable commodities like shares and bonds.

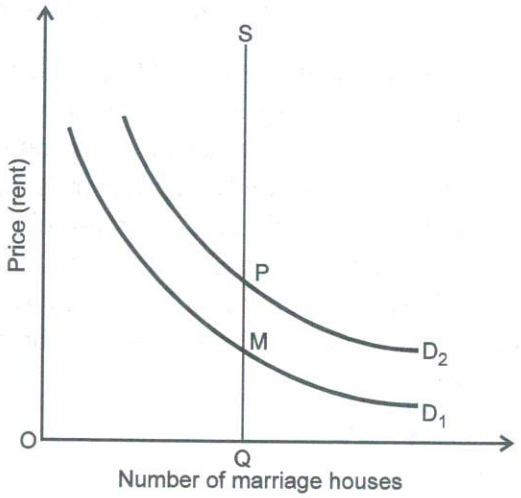


Fig. 3.6 *Determination of Market Price*

(ii) Price determination in the short-run

While in market period (or very short-run), supply is absolutely fixed. In the short-run, it is possible to increase (or decrease) the supply by increasing (or decreasing) the variable inputs. In the short-run, therefore, supply curve is elastic, unlike a straight vertical line in the market period. Supply curve in the short-run approximates the *SMC* curve.

Under competitive conditions, the process of price determination and output adjustment in the short-run is given in Figure 3.7(a) and 3.7(b). Figure 3.7(a) shows that demand curve DD and supply curve SS intersect at point P determining the price at OP_1. This price is fixed for all the firms in the industry.

Given the price $PQ (= OP_1)$, in Figure 3.7(a), an individual firm can produce and sell any quantity at this price. However, any quantity will not yield maximum profit. The firms will have to adjust their output to the price OP_1. The process of output determination is presented through Figure 3.7(b).

NOTES

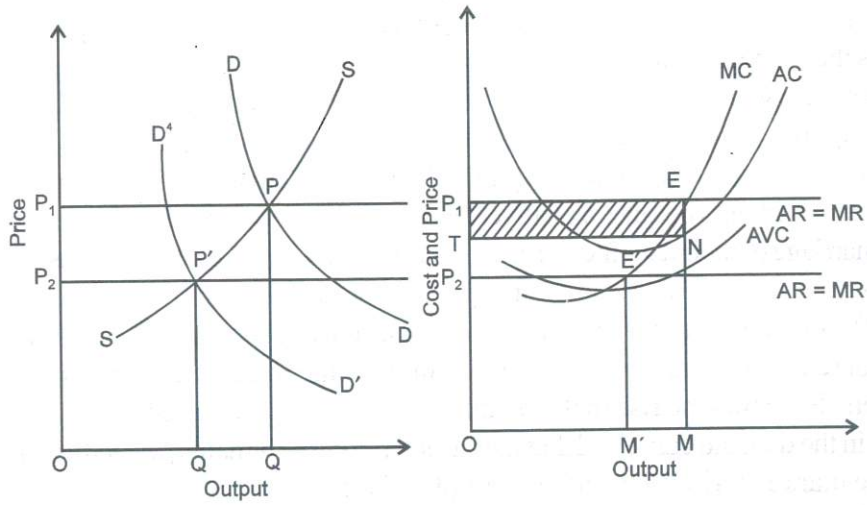


Fig. 3.7 Pricing under Perfect Competition: Short-run

Since a firm can sell any quantity at price OP_1, the demand for the firm's product is given by a horizontal straight line, $AR = MR$. Price being constant, its average revenue and marginal revenue are equal. Firm's upward sloping MC curve beyond its AVC curve represents its supply curve. Firm's MR and MC curves intersect each other at point E. This is the firm's equilibrium point. The perpendicular EM determines the profit-maximizing output at OM. At this output, firm's $MR = MC$, which satisfies both the first order and the second order conditions of maximum profit. The total maximum profit is shown by the area $P_1 TNE$. The total profit (h) may be calculated as:

$$h = (AR - AC)Q$$

In Figure 3.7(b),

$$AR = EM;$$

$$AC = NM;$$

and $$Q = OM.$$

By substituting the values from Figure 3.7(b), we get

$$h = (EM - NM)\, OM$$

Since $EM - NM = EN$,

$$h = EN \times OM$$

This is the maximum profit that a firm can make, given the cost and revenue conditions as presented in Figure 3.7(b).

Now, if price falls to OP_2 due to downward shift in the demand curve to $D'D'$, the firm will be in equilibrium at point $E¢$. Here again, the firm's $AR' = MR' = MC$, and its $AR < AC$. Therefore, the firm incurs loss. However, in the short-run, it may not be desirable to close down so long as it covers its MC.

Short-Run equilibrium of the industry

Now let us discuss the short-run equilibrium of the industry.

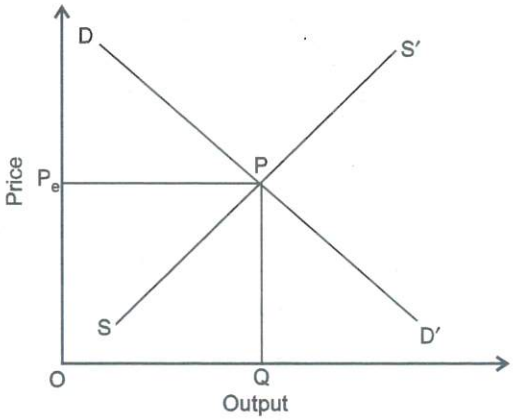


Fig. 3.8 Equilibrium of the Industry

An industry is in equilibrium in the short-run when market is cleared at a given price, i.e., when the total supply of the industry equals the total demand for its product. The price at which market is cleared is equilibrium price. The industry being in equilibrium, there is no tendency to expand or to contract the output. The equilibrium of the industry is shown at point P in Figure 3.8. The industry demand and supply curves intersect at point P, determining equilibrium price OP_e. The industry is supplying as much as consumers demand. In the short-run equilibrium of the industry, individual firms may make pure profits, normal profits or losses, depending on their cost conditions.

(iii) Price and output determination in the long-run

Unlike in the short-run, the supply curve in the long-run is supposed to be more elastic. Long-run brings in two additional factors in operation which make the supply curve more elastic. **First,** in the long-run, it becomes possible for the existing firms to increase their output by increasing the size of their plant. **Second,** new firms may enter and some existing ones may leave the industry. Entry and exit of firms bring about the long-run variation in the output. If cost and revenue conditions in the long-run are such that some firms are making losses and are not able to adjust their plant size and cost structure to the market price, such firms leave the industry. This makes the market supply curve shift leftward, causing a rise in the price. The increase in market price increases the excess profit of the profit-making firms. Under the conditions of the perfect competition (i.e., free entry and exit), the pure profit would invite many new firms to the industry. This will make supply curve shift rightward, causing a decrease in the price, which will eventually take away the excess or pure profits. All firms earn only *normal profit*.

Price determination in the long-run

As in the short-run, market price is determined in the long-run by the market forces of demand and supply. Let us suppose that the market demand curve is given by DD' which is relevant for both short-run and long-run, and short-run supply curve is given by SS_1 in Figure 3.9(a). The market demand curve DD' and market supply curve SS_1 intersect each other at point P_1 and the short-run market price is determined at OP_0. At this price, the firms find their short-run equilibrium at point E_1 and each of them produces output OQ_1. The total market supply equals $OQ_1 \times$ Number of firms $= ON_1$ [in panel (a) of Figure 3.9] and the industry is in short-run equilibrium.

NOTES

Given the cost and revenue conditions in Figure 3.9(b), the firms are making super normal profit of E_1M per unit. The existence of super normal profit in the short-run leads to increase in the market supply on two accounts: *one,* new firms will enter the industry attracted by the super normal profits, and *two*, the existing firms would expand their plant size because returns to scale would increase, as shown by the LAC. As a result, the market supply would increase so that supply curve shifts rightward to SS_2 [Figure 3.9(a)]. The shift in supply curve brings down the market price to OP' which is the long-run equilibrium price. Thus, equilibrium price is once again determined in the market.

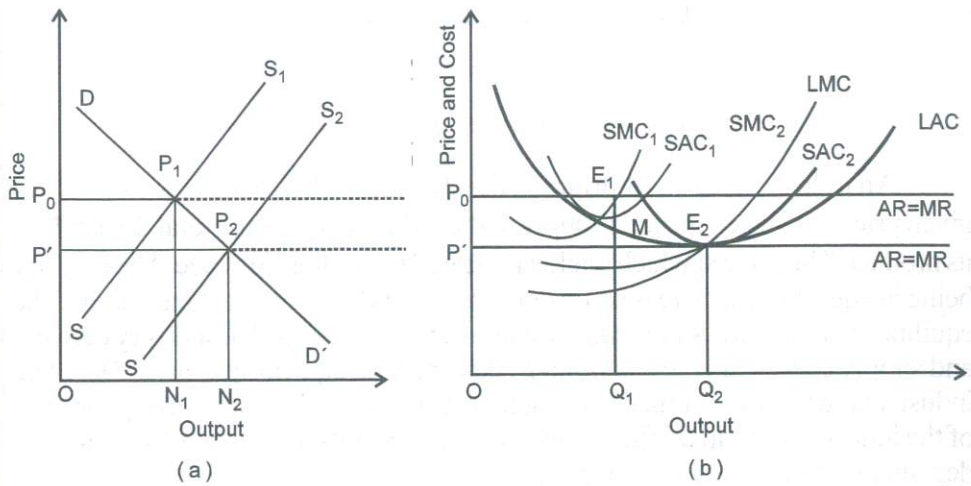


Fig. 3.9 Long-run Equilibrium of the Firm

B. Equilibrium of the firm in the long-run

The firms are in equilibrium in the long-run when their $AR = MR = LMC = LAC = SMC = SAC$.

This means that the firms of an industry reach their equilibrium position in the long-run where both short-run and long-run equilibrium conditions coincide. In a perfectly competitive market, the cost and revenue conditions are given for the firms. What the firms can do, therefore, is to adjust their output to the given revenue and cost conditions in order to maximize their profit. Let us now illustrate the process of adjustment of output so as to reach the equilibrium in the long-run.

Suppose that the firms are in equilibrium at point E_1 in Figure 3.9(a) where they make excess profits $AR - SAC_1 = EM$ per unit. This gives incentives to the firms to expand their scale of production, i.e., they add more plants to the existing ones. As a result, market supply increases. Besides, supply also increases because new firms enter the industry. Therefore, the market supply curve SS_1 tends to shift rightward, causing a fall in price to OP'. On the other hand, due to increase in demand for inputs, cost tends to rise. But so long as economies of scale are greater than the diseconomies of scale, the LAC tends to decrease and it pays firms to expand their plant size. When a stage is arrived where $P < LAC$, firms incur losses. The firms that are not able to make adjustment in the plant size or scale of production leave the industry. This works in two directions. On the one hand, supply decreases and price increases, and on the other, demand for inputs decreases, which causes a decrease in the input prices. This process of adjustment

continues until LAC is tangent to $P = AR = MR$ for each firm in the industry. This position is shown at point E_2 in Figure 3.9(b). Eventually, at point E_2, i.e., at the point of equilibrium,

$$P = MR = LMC = LAC = SMC = SAC$$

Since $P = LAC$, the firms make only normal profits in the long-run. If firms deviate from point E_2, due to some short-run disturbances, the market forces will make them return to this point only.

Equilibrium of the industry

An industry is in equilibrium when its market demand equals its market supply. When an industry is in equilibrium, all its firms are supposed to be in equilibrium [as shown in Figure 3.9(b)]. When an industry is in equilibrium, all its firms earn only normal profits, because under the conditions of perfect competition, all the firms are assumed to achieve the same level of efficiency in the long-run. Since industry yields only normal profits, there is no incentive for new firms to enter the industry. These conditions are fulfilled at price OP' in Figure 3.9(a) and (b).

$$LMC = LMR = SMC = SAC = P = LAC$$

Since $P = LAC$, all the firms are earning only normal profits. At industry's equilibrium output ON_2, market demand equals market supply [Figure 3.9(a)]. At price OP', therefore, market is cleared. The output ON_2 has a fair chance to remain stable in the long-run. For, there is no incentive for new firms to enter the industry and for existing ones to leave the industry. The industry is, therefore, in equilibrium.

Long-run supply curve of a competitive industry

We have earlier derived the short-run supply curve of the industry by summing horizontally the individual supply curves of the firms, which are based on their short-run MC curves. The long-run supply curve of a competitive industry, however, has nothing do with the LMC curves. The shape of the long-run supply curve of an industry, under perfect competition, depends on whether factor prices remain *constant* or *increase* when demand for inputs increases in the long-run as a result of expansion in the output of the industry. According to whether cost is constant, increasing or decreasing, an industry is termed as *constant cost, increasing cost* or *decreasing cost* industry. Let us now derive the long-run supply curve of each of such industries.

(a) Constant cost industry

An industry in which factor prices are independent of the rate of increase in factor demand is referred to as *constant cost industry*. In other words, when the expansion of output in an industry does not entail an increase in factor prices, the industry is said to be a constant cost industry. The derivation of the supply curve of such an industry is illustrated in Figure 3.10(a) and (b). Let us suppose that the industry is in equilibrium at P where demand curve DD_1 and supply curve SS_1 intersect each other. The industry is in equilibrium at price OP_1 and output OQ_1. At price OP_1, all firms are in equilibrium as their $LMC = P = MR = SMC = SAC$.

Suppose now that demand curve shifts to DD_2 due to increase in consumers' income or increase in population or both. As a result, market price increases to OP_2. In the short-run, this increase in price causes increase in supply by the firms from OM to ON [Figure 3.10(b)], determined by the point of intersection of firm's SMC and new price line through P_2. The firms enjoy abnormal profits in the short-run.

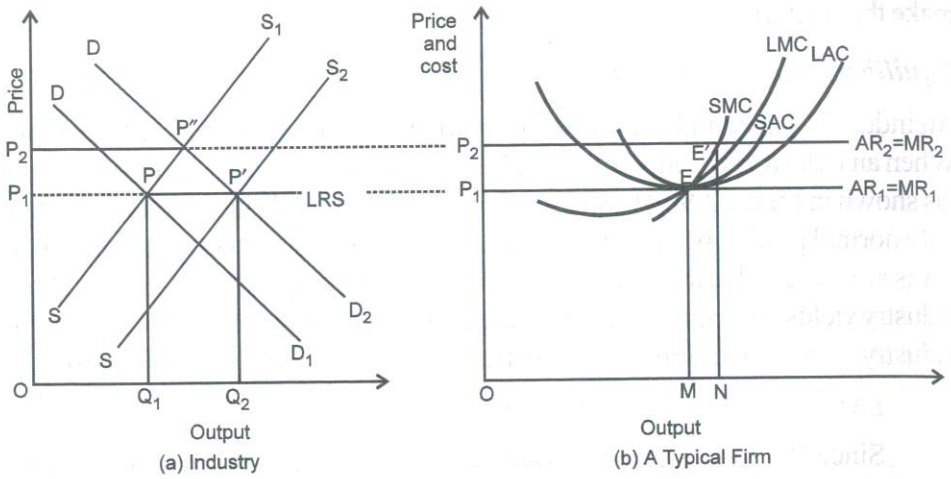


Fig. 3.10 *Long-run Supply Curve of the Constant Cost Industry*

The abnormal profits attract new firms into the industry. The entry of new firms leads to increase in demand for factors. The industry being a constant-cost industry, factor prices do not increase. Cost of production for both new and old firms remain constant at the previous level. However, due to the entry of new firms, market supply increases and market supply curve shifts to SS_2 [Figure 3.10(a)]. Consequently, in the long-run, market price falls to its previous level, OP_1, and individual firms return to their previous equilibrium point E. But the industry output increases form OQ_1 to OQ_2, since industry moves to a new equilibrium P'. By joining the two industry-equilibrium points, P and P', we get long-run supply curve (LRS) of the constant cost industry. Obviously, the long-run supply curves (LRS) of a constant cost industry is a horizontal straight line, as given by the line LRS.

(b) Increasing cost industry

An industry is referred to as an *increasing cost industry* if factor prices increase due to increase in demand for inputs. The long-run supply curve of an increasing cost industry has a positive slope.

The derivation of long-run market supply curve under increasing cost condition is demonstrated in Figure 3.11. Let the original demand and supply curves of the industry be represented, by DD_1 and SS_1, respectively, and industry be in equilibrium at point A. Let us now suppose that for some reasons, demand curve DD_1 shifts rightward to DD_2. As a result, short-run market price increases from OP_1 to OP_3. With this increase in price, the demand curve for the individual firms shifts upward to $AR_3 = MR_3$ [Figure 3.11(b)]. The firms, therefore, enjoy super normal or economic profits. This profit attracts new firms into the industry and demand for inputs increases. Since in an increasing cost industry, the supply

of inputs is assumed to be less than infinitely elastic, the entry of new firms causes an increase in the input prices. Consequently, cost curves, both short-run and long-run, shift upward from LAC_1 to LAC_2. In this process of adjustments, however, industry supply increases so that market supply curve SS_1 shifts rightward to SS_2. With this shift in supply curve, the industry reaches another equilibrium position at point C where new demand and supply curves intersect each other. A new market price OP_2 is determined. At price OP_2 [Figure 3.11(b)], the long-run and short-run cost curves (LAC_2, etc.) are tangent to the price line (OP_2). The firms shift to a new long-run equilibrium, E_2, its output remaining the same. Whether equilibrium output of the firms remains constant, increases or decreases, depends, on whether cost curves shift upward vertically, upward to the right or upward to the left, respectively.

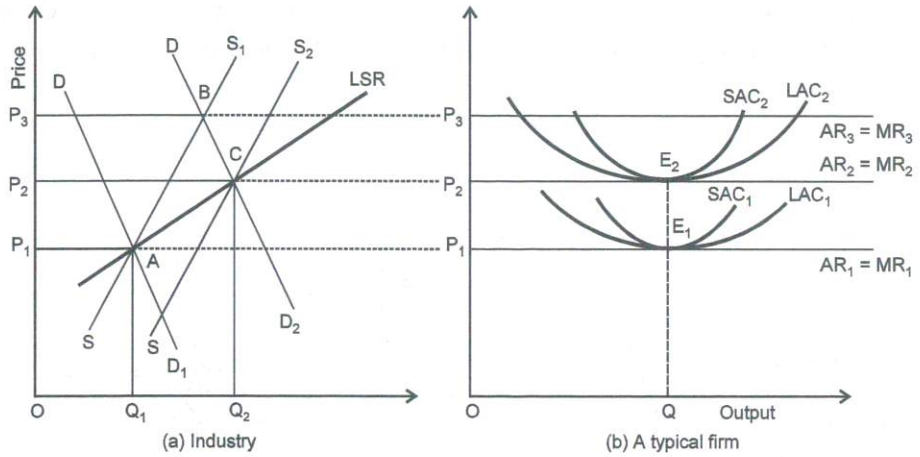


Fig. 3.11 *Long-run Supply Curve of an Increasing Cost Industry*

Note that at price OP_2, both industry and individual firms are in equilibrium. In the absence of any further disturbance, the equilibrium of both firms and industry has a fair chance to remain stable. Thus, at the new equilibrium price OP_2, the industry output increases from OQ_1 to OQ_2 and corresponding equilibrium points are A and C, respectively. By joining the long-run equilibrium points A and C, we get the long-run supply curve for the industry, as shown by the curve LRS. Obviously, the LRS has a positive slope in an increasing cost industry.

(c) Decreasing cost industry

If expansion of output of the an industry is associated with or leads to decrease in the input prices, the industry is referred to as a decreasing cost industry. A decreasing cost industry has a long-run industry supply curve with a negative slope, since input prices decrease with the expansion of the industry output.

The derivation of long-run industry supply curve (LRS) for a decreasing cost industry is illustrated in Figure 3.12(a) and (b). Let the industry be initially in equilibrium at point A [Figure 3.12(a)] and firms at be initially in equilibrium E_2 [Figure 3.12(b)]. Now suppose that demand curve shifts from DD_1 to DD_2 and, consequently, price rises from OP_2 to OP_3. The short-run equilibrium of firms at price OP_3 [Figure 3.12(b)] moves upward on SMC_2, where the firms make abnormal profits. The abnormal profits attract new firms to the industry, causing

increase in demand for inputs. If input industries are enjoying increasing returns to scale due to economies of scale, the increase in demand for inputs would encourage increased supply of inputs. Increase in the supply of inputs causes input prices to fall. The industry, therefore, enjoys the external economies to scale. As a result, their long-run and short-run cost curves shift downward, from LAC_2 to LAC_1 [Figure 3.12(b)].

From the industry's point of view, industry supply increases due to the entry of the new firms, even if the existing firms maintain their old level of output. Therefore, the industry supply curve shifts from SS_1 to SS_2 which intersects the new demand curve DD_2 at point C. Thus, equilibrium of the industry shifts from point A to point C. Industry output increases from OQ_1 to OQ_2. In the absence of any external disturbance, the industry equilibrium point C would tend to stabilize. By joining the two equilibrium points A and C, we get the long-run supply curve of the industry, LRS. The LRS has a negative slope.

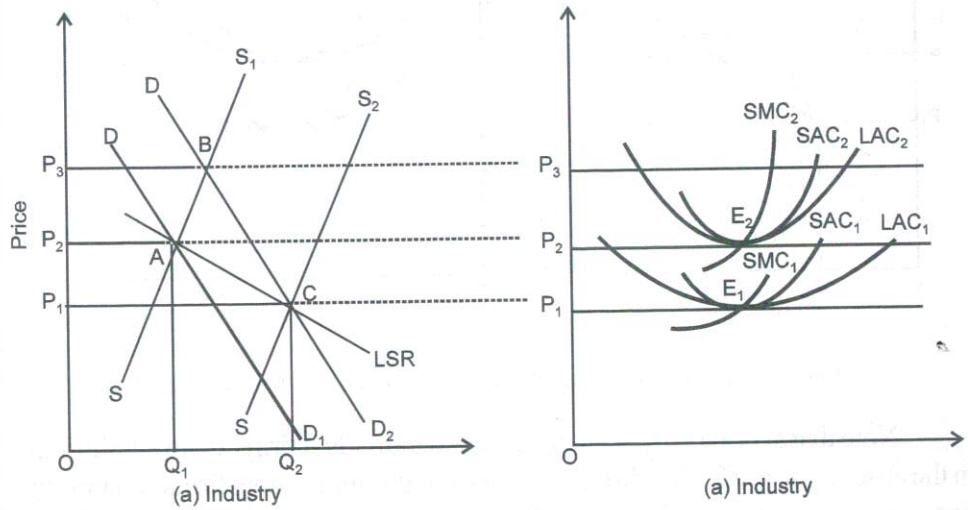


Fig. 3.12 Long-run Supply Curve of a Decreasing Cost Industry

Is decreasing cost a reality?

Some authors argue that the phenomenon of decreasing cost... is not consistent with all the requirements of the perfect competition. However, the possibility of a decreasing cost industry cannot be ruled out in a very long period. One reason for this is the likelihood of existence of large *external economies* of sale, particularly in case of young industries in undeveloped areas. An increase in the number of industries and the consequent growth of transportation, marketing facilities and financial institutions may reduce the industry's cost of production. Nevertheless, it depends on how substantial are the external economies of scale. RG Lipsey has cited the car industry of England as an example of decreasing cost industry. In his words, As the output of cars increased, the industry's demand for tyres grew greatly. This would have increased the demand for rubber and tended to raise its price, but it also provided the opportunity for tyre manufacturers to build large modern plants and reap the benefits of increasing returns in tyre production. At first, these economies were large enough to offset any factor-

price increases and tyre-price charged to car manufacturers fall. Thus car costs fell because of lower prices of an important input. Whether industry costs are constant or decreasing also depends on the proportion of total input supplies used by the industries. For example, output of the pencil industry can be increased without substantially affecting the lumber prices, as the pencil industry uses a small proportion of the total lumber output. However, increase in the output of furniture industry will affect the lumber price. Similarly, output of the pin industry can be substantially increased without affecting the steel price. Whereas, the output of the car industry cannot be substantially increased without affecting steel prices. Another factor that may cause rise in input prices is whether or not input industries enjoy economies of scale.

Moreover, the most common cases are of the constant and increasing cost industries. Decreasing cost industries are most unlikely to exist. The constant and decreasing cost industries tend, over time, to become increasing cost industries. Because external economies have a limit to which cost can be decreased, external diseconomies would exceed the external economies in the very long period.

Check Your Progress

1. Classify the market structures on the basis of the degree of competition.
2. List any two characteristics of perfect competition.
3. How is the supply curve of an individual firm derived?

3.4 PRICE AND OUTPUT UNDER MONOPOLY

The term *pure monopoly* signifies an absolute power to produce and sell a product which has no close substitute. In other words, a monopoly market is one in which there is only one seller of a product having no close substitute. The cross elasticity of demand for a monopolized product is either zero or negative. In a monopolized market structure, the industry is a single-firm industry. Firm and industry are identical in a monopoly setting.

Moreover, the precise definition of monopoly has been a matter of opinion and purpose. For instance, in the opinion of Joel Dean, a monopoly market is one in which a product of lasting distinctiveness is sold. The monopolised product has distinct physical properties recognized by its buyers and the distinctiveness lasts over many years. Such a definition is of practical importance if one recognizes the fact that most of the commodities have their substitutes varying in degree and it is entirely for the consumers or users to distinguish between them and to accept or reject a commodity as the substitute. Another concept of pure monopoly has been advanced by DH Chamberlin who envisages the control of all goods and services by the monopolist. However, such a monopoly has hardly ever existed, hence his definition is unrealistic. In the opinion of some others, any firm facing a sloping demand curve is a monopolist. This definition includes all kinds of firms except those under perfect competition.

NOTES

Features of monopoly

Some important features of monopoly are as follows:

1. There is a single seller of a product that has no close substitute.
2. A monopoly firm is a price-maker, not a price-taker.
3. Under monopoly, there is absence of supply curve.
4. A monopoly makes a single-firm industry.

Sources and kinds of monopolies

The emergence and survival of a monopoly is attributed to the factors that prevent the entry of other firm into the industry. The barriers to entry are, therefore, the sources of monopoly power. The major sources of barriers to entry to a monopolised market are as follows:

(i) **Legal restrictions:** Some monopolies are created by the law in the public interest. Most of the state monopolies in the public utility sector, including postal, telegraph and telephone services, radio, generation and distribution of electricity, railways, airlines and state roadways are public monopolies that are created by the public law. The state may create monopolies in the private sector also by restricting entry of other firms by law or by granting patent rights. Such monopolies are intended to reduce cost of production to the minimum by enlarging the size and investing in technological innovations. Such monopolies are known as *franchise monopolies*.

(ii) **Control over key raw materials:** Some firms acquire monopoly power because of their traditional control over certain scarce and key raw materials, which are essential for the production of certain other goods, e.g., bauxite, graphite and diamond. For instance, Aluminium Company of America had monopolised the aluminium industry before World War II because it had acquired control over almost all sources of bauxite supply. Such monopolies are often called '*raw material monopolies*'. The monopolies of this kind emerge because of monopoly over certain specific knowledge or technique of production.

(iii) **Efficiency:** A primary and technical reason for the growth of monopolies is the economies of scale. In some industries, long-run minimum cost of production or the most efficient scale of production almost coincides with the size of the market. Under this condition, the large-size firm finds it profitable in the long-run to eliminate the competition by cutting down its price for a short period. Once monopoly is established, it becomes almost impossible for the new firms to enter the industry and survive. Monopolies existing on account of this factor are known as natural monopolies. A natural monopoly emerges either due to technical efficiency or is created by the law on efficiency grounds.

(iv) **Patent rights:** Another source of monopoly is the patent right of the firm for a product or for a production process. Patent rights are granted by the

government to a firm to produce a commodity of specified quality and character or to use a specified technique of production. Patent rights give a firm exclusive rights to produce the specified commodity or to use the specified technique of production. Such monopolies are called *patent monopolies*.

(i) Demand curve under monopoly

The nature of revenue curves under monopoly depends on the nature of demand a monopoly firm faces. It has been seen that in a perfectly competitive market, firms face a horizontal, straight-line demand curve. It signifies that an individual firm of an industry can sell any quantity at the prevailing price. Under monopoly, however, there is no distinction between the firm and the industry. The monopoly industry is a single-firm industry. The monopoly firm is, therefore, capable of influencing the industry price by changing the level of its production, which is eventually the industry output. Besides, a monopoly firm is free to choose between price quantity combination. It can fix higher price and sell a lower quantity and *vice versa*. For these reasons, *a monopoly firm faces a demand curve with a negative slope*. What is important in the context of monopoly pricing is the relation between firm's average revenue (*AR*) curve and its marginal revenue (*MR*) curve. The analysis is, therefore, repeated here.

Relation between AR and MR

The relationship between *AR* and *MR* plays an important role in price and output determination under monopoly. Therefore, let us look at technical relationship between *AR* and *MR*. The relationship between *AR* (= *P*) and *MR* can be specified as follows:

Recall that total revenue, *TR*, equals *P* times *Q*, i.e.,

$$TR = P \cdot Q$$

and marginal revenue, (*MR*) is obtained by differentiating $TR = P \cdot Q$ with respect to *P*. Thus,

$$MR = \frac{\partial TR}{\partial P} = P + Q\frac{\partial P}{\partial Q}$$

$$= P\left(1 + \frac{Q}{P}\frac{\partial P}{\partial Q}\right) \quad \ldots (3.1)$$

Note that $\frac{Q}{P}\frac{\partial P}{\partial Q}$ is the reciprocal of the elasticity.

Thus, $\frac{Q}{P}\frac{\partial P}{\partial Q} = -\frac{1}{e}$

By substituting $-\frac{1}{e}$ for $\frac{Q}{P} \cdot \frac{\partial P}{\partial Q}$ in Equation (3.1), we get

$$MR = P\left(1 - \frac{1}{e}\right) \quad \ldots (3.2)$$

NOTES

$$\text{or} \qquad MR = P - \frac{P}{e} \qquad \ldots (3.3)$$

$$\text{Since} \qquad P = AR$$

$$MR = AR - \frac{AR}{e} \qquad \ldots (3.4)$$

This relationship between *MR* and *AR* can be derived geometrically. Consider the *AR* and *MR* curves in Figure 3.13.

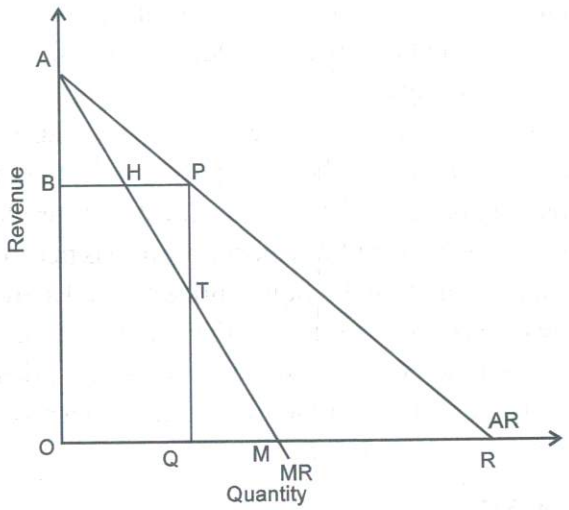


Fig. 3.13 *Relationship between AR and MR*

Let us suppose that price is given at $PQ (= BO)$. The elasticity at point P on the *AR* curve can be expressed as

$$e = \frac{QR}{OQ} = \frac{PR}{AP} = \frac{OB}{AB}$$

where e = elasticity of demand curve

Since $OB = PQ$,

$$\therefore \qquad e = \frac{PQ}{AB} \qquad \ldots (3.5)$$

It can be proved that $AB = PT$. By substituting PT for AB in Equation (3.5), we get

$$e = \frac{PQ}{PT} \qquad \ldots (3.6)$$

Since $PT = PQ - TQ$, Equation (3.6) may be written as:

$$e = \frac{PQ}{PQ - TQ} \qquad \ldots (3.7)$$

It can be seen from (Figure 3.13) that at price OB, $PQ = AR$ and $TQ = MR$. Therefore, Equation (3.7) can be expressed as

$$e = \frac{AR}{AR - MR}$$

NOTES

and $MR = AR - \frac{AR}{e}$ ... (3.8)

Note that Equation (3.4) is the same as Equation (3.8).

Given the Equation (3.7), AR can be easily obtained.

Since $MR = AR - \frac{AR}{e}$

or $MR = AR\left(1 - \frac{1}{e}\right)$ ... (3.9)

$$AR = \frac{MR}{1 - \frac{1}{e}}$$

or $AR = MR\left(\frac{1}{e-1}\right)$ ... (3.10)

The general relationships between AR and MR are given by Equation (3.9) and Equation (3.10). A general pattern of relations between AR and MR can be easily obtained from Equation (3.9) as follows. Given the negative slope of the demand curve,

when

$e = 1, MR = 0,$	$AR > 0$	$\therefore$	$AR > MR$
$e < 1 > 0\ MR < 0,$	$AR > 0$	$\therefore$	$AR > MR$
$e > 1 < \infty\ MR > 0,$	$AR > 0$	but	$AR > MR$
$e = 0, MR < 0,$	$AR = 0$	$\therefore$	$AR > MR$
$e = \infty, MR > 0,$	$AR > 0$	and	$AR = MR$

It must be noted that the slope of the MR curve is twice that of the AR curve.

(ii) Cost and supply curves under monopoly

In the short-run, cost conditions faced by a monopoly firm are, for all practical purposes, identical to those faced by a firm under perfect competitions, particularly when a monopoly firm is a competitive buyer in the input market. But in case a monopoly firm uses specified inputs for which there is no general market and holds the position of a monopolist in the input-market, then the price of the inputs depends on the monopolist's demand for it, given the supply condition. The monopoly firm may then face a positively sloping supply curve in the input market, and its cost curves would be different from those of firms under perfect competition. In fact, the monopoly firm would face a rising supply price and its cost curves would rise rapidly. In general, however, most monopoly firms use unspecified inputs, and they are one among many buyers of the inputs. In the short-run, therefore, a monopoly firm is faced with usual U-shaped AC and MC curves.

It must be noted that under perfect competition, the MC curve forms the basis of a firm's supply curve. It is important to note here that the MC curve is not the monopolist's supply curve. In fact, under monopoly, there is no unique relation

NOTES

between market price and quantity supplied. Therefore, there is no supply curve under monopoly.

3.4.1 Profit Maximization under Monopoly

The objective of a monopoly firm, like all other firms, is assumed to be profit maximization. Profit maximization is, however, not necessarily the sole objective of the firm. The monopoly firm may seek to maximize its utility function, particularly where management of the firm is divorced from its ownership. However, as mentioned earlier, most common objectives of business firm assumed in traditional theory of pricing is profit maximization. We will therefore explain the equilibrium of monopoly firm in short-run and long-run under profit maximization hypothesis.

(a) Monopoly equilibrium in the short-run

Like any other firm, a monopoly firm reaches its equilibrium where it maximizes its total profits. As noted earlier, profits are maximum where following two conditions are fulfilled: (i) that $MC = MR$—the necessary condition, and (ii) that the MC curve must intersect the MR curve from below under increasing cost condition—the supplementary condition. The monopoly firm fixes its price and output in accordance with these conditions.

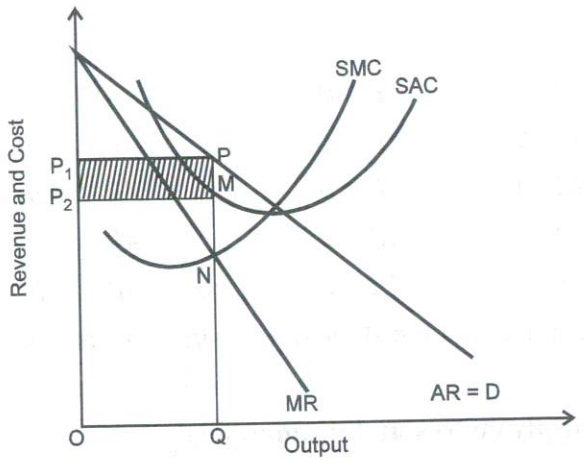


Fig. 3.14 Price Determination under Monopoly: Short-run

The price and output determination under monopoly and the firm's equilibrium are demonstrated in Figure 3.14. The $AR = D$ and MR curves show the revenue conditions, while SMC and SAC present the short-run cost conditions faced by the monopoly firm. Given the revenue and cost curves, the decision rule for selecting profit maximizing output and price is the same as for a firm in the competitive industry, i.e., firm's $MR = MC$ and slope of $MC >$ the slope of MR. Therefore, the monopoly firm chooses a price–output combination for which $MR = SMC$. The MR and SMC curves intersect each other at point N. Thus, the profit maximizing output for the firm is OQ, since at this output firm's $MR = SMC$. Given the demand curve $AR = D$, the output OQ can be sold per unit time at only one price, i.e., PQ $(= OP_1)$. Thus, the determination of equilibrium output simultaneously determines the price for the monopoly firm. Once price and output are determined, the total profits are also simultaneously determined.

At output OQ and price PQ, the monopoly firm maximizes its profit. Its per unit monopoly or super-normal profit (i.e., $AR - SAC$) is $(PQ - MQ) = PM$. Its total profit $p = OQ \times PM$. Since $OQ = P_2M$, $p = P_2M \times PM$, as shown by the shaded area. Since in the short-run cost and revenue conditions are not expected to change, the equilibrium of the monopoly firm will remain stable.

Two common misconceptions

There are two common misconceptions about a monopoly firm exist. One of the misconceptions is that a monopoly firm necessarily makes super-normal profits. There is, however, no guarantee that a monopoly firm will always make profits in the short-run. In fact, whether a monopoly firm makes profits or losses in the short-run depends on its revenue and cost conditions. It is quite likely that its *SAC* lies above its *AR*, as shown in Figure 3.15. The monopoly firm then makes losses to the extent of $PM \times OQ = P_2MPP_1$. The firm may yet continue to produce and sell in the hope of making profits in the long-run. However, the monopoly firm, like a competitive firm, will stick to the maximization rules (i.e., $MR = MC$) in order to minimize its losses.

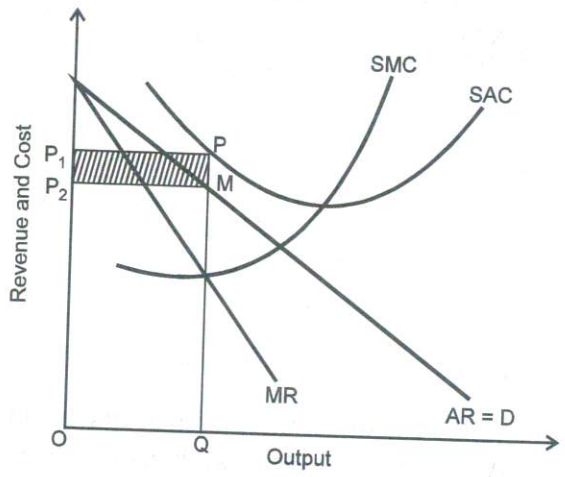


Fig. 3.15 *Monopoly Equilibrium in the Short-run: Losses*

Another common misconception about monopoly is that the demand curve faced by a monopoly firm is perfectly inelastic so that it can charge any price it likes. In fact, the demand curve faced by a monopolist is both firm's and industry's demand curve. In addition, most market demand curves are negatively sloped, being highly elastic towards their upper end and highly inelastic towards their lower ends. The equilibrium output of the monopolist that maximizes his profits will always be within the elastic region of the demand curve if his $MC \neq 0$.

(b) Monopoly equilibrium in the long-run

The long-run conditions faced by a monopolist are different from those faced by competitive firms in an important aspect, i.e., the entry of new firms into the industry. While in a competitive industry, there is free entry of new firms to the industry, a monopoly firm is protected from competition by *barriers to entry*.

NOTES

Protected by barriers to entry, a monopoly firm gets an opportunity to expand the size of its plant with a view to maximizing its long-run profits. The expansion of the plant size may, however, be subject to such conditions as (a) size of the market; (b) expected economic profits; and (c) risk of inviting legal restrictions. Assuming, none of these conditions limits the expansion of monopoly firm, the general case of monopoly equilibrium in the long-run is illustrated in Figure 3.16. The *AR* and *MR* curves show the market demand and marginal revenue conditions faced by the monopoly firm. The *LAC* and *LMC* curves show the long-run cost conditions. The profit maximizing monopoly firm equalizes its *LMC* and *MR* at output OQ_2. The price at which the total output OQ_2 can be sold is P_2Q_2. Thus, in the long-run equilibrium, price is P_2Q_2 and equilibrium output is OQ_2. This output–price combination maximizes the monopolist's long-run profits. The total monopoly profit is shown by the area LP_2SM.

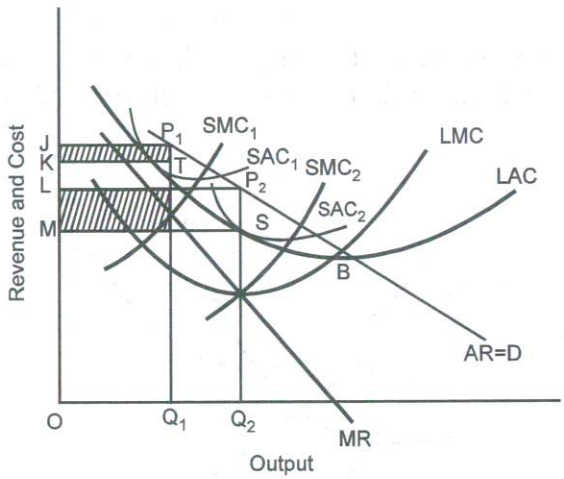


Fig. 3.16 *Monopoly Equilibrium in the Long-run*

It may be noted at the end that if there are barriers to entry, the monopoly firm would neither reach the optimal scale of production in the long-run nor it will make full use of its existing capacity. This case can be verified from Figure 3.16. The optimum size of the plant is given by point *B*, i.e., at the minimum *LAC*. But the monopoly firm settles at less than optimal output because optimum size of the plant will not yield the maximum profit.

Moreover, if the size of the market and the cost conditions permit, a profit maximizing monopoly firm may even exceed the optimum size of the plant and overutilize its long-run capacity. Figure 3.17 depicts the more-than-optimal size of the plant and its overutilization. The optimum size of the plant is given at point *B*—the point of intersection between *LAC* and *LMC*—whereas the monopoly firm chooses output at *M* where profit is maximum. Alternatively, the monopoly firm may find its equilibrium just at the optimum size of the plant. This is possible only when the market size is just large enough to permit optimization and full utilization of the plant size. This possibility has been illustrated in Figure 3.18.

NOTES

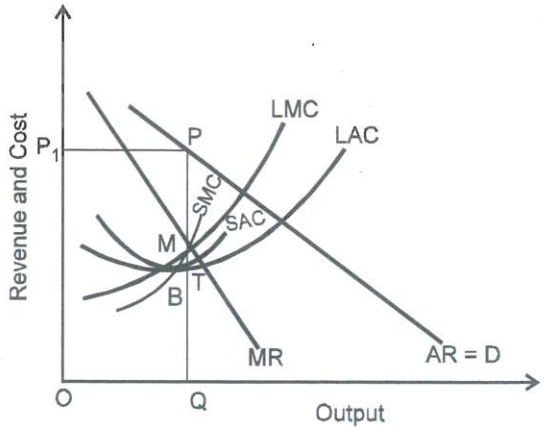


Fig. 3.17 Monopoly Equilibrium: Overutilization of Point Size

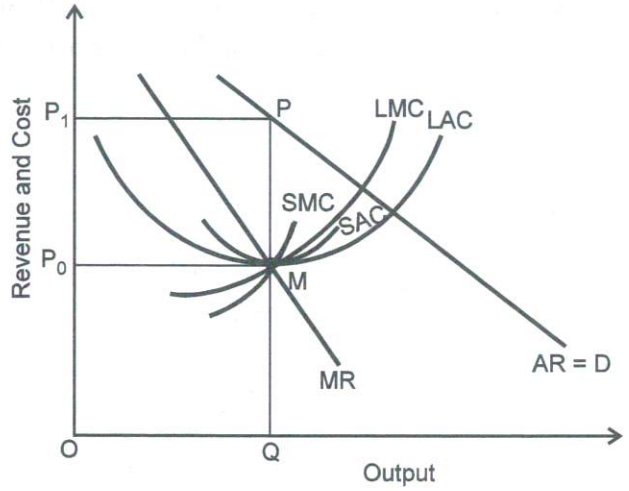


Fig. 3.18 Monopoly Equilibrium at Optimal Size of the Plant

Why is there no supply curve under a monopoly?

As already mentioned, there is no unique or precise supply curve under monopoly. Let us now examine this fact by using the concept of equilibrium output. The supply curve presents a unique relationship between price and quantity demanded. This unique relationship between market price and quantity supplied does not exist under monopoly. The reason is that a profit-maximizing monopoly firm does not determine its output where $P = MC$ or where $AR = MC$. Rather, it determines its equilibrium output where $MR = MC$. Therefore, a unique relationship between price ($AR = P$) and quantity supplied cannot be traced. It is therefore quite possible to trace (i) that given the MC, the same output is supplied at different prices, and (ii) that at a given price, different quantities are supplied if the two downward-sloping demand curves have different elasticities. The two cases are illustrated in Figures 3.19 and 3.20, respectively.

NOTES

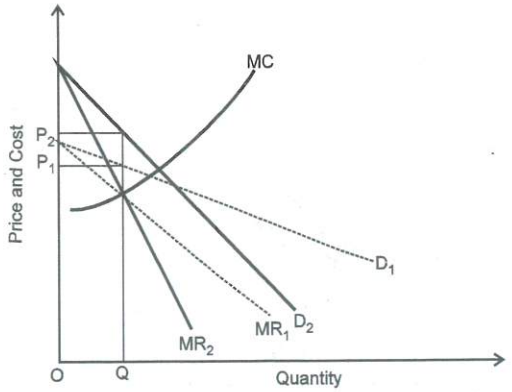


Fig. 3.19 Same Quantity Supplied at Two Different Prices

As Figure 3.19 shows, given the *MC*, the same quantity *OQ* can be supplied at two different prices—OP_1 when demand curve is D_1 and OP_2 when demand curve is D_2. Obviously, there is no unique relationship between price and quantity supplied.

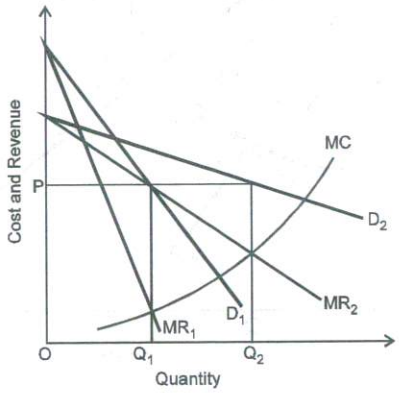


Fig. 3.20 Different Quantities Supplied at the Same Price

Figure 3.20 presents the case of two different quantities supplied at the same price, *OP*. Given the *MC*, quantity OQ_1 is supplied when demand curve is D_1 and quantity OQ_2 is supplied when demand curve is D_2 at the same price *OP*. In this case too, there is no unique relationship between price and quantity supplied. It is thus clear that there is no unique supply curve under monopoly.

3.4.2 Monopoly vs Perfect Competition: Comparison of Long-run Price and Output

Comparison of long-run price and output is easier than that of short-run.

Figure 3.21 presents a comparative analysis of equilibrium price and output under perfect competition and monopoly in the long-run. Let us assume that *LMC* and *LAC* are identical for both a competitive industry and a monopoly.

The equilibrium condition for a competitive industry in the long-run requires that all its firms are in equilibrium. That is, all the firms have their $AR = MR = LAC = LMC$. This condition is satisfied at point *P'* in Figure 3.21. Thus, in a competitive industry, equilibrium price will be OP_1 and equilibrium output will be OQ_2. Now, if

this industry were to be monopolized, the revenue conditions (*AR* and *MR*) and profit maximization rule will be different. The monopoly firm will maximize its profits at the level of output where *MR* = *MC*. The equilibrium condition for the monopoly firm in fulfilled at point *B*. Therefore, the equilibrium output under monopoly will be OQ_1 and the equilibrium price will be OP_2.

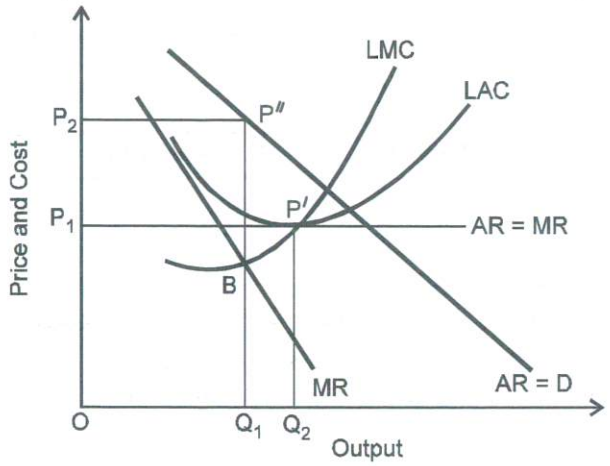


Fig. 3.21 *Comparison of Price and Output: Monopoly vs Competitive Industry*

Two major conclusions

Two important conclusions can be drawn from the comparison of equilibrium price and output of monopoly and competitive industry—(i) monopoly results in an inoptimal output, and (ii) monopoly causes loss of social welfare. These conclusions are illustrated in Fig 3.22, assuming a constant cost industry.

(i) Inoptimal of Output: As Figure 3.22 shows, if both monopoly and competitive industries are faced with identical cost conditions, the equilibrium output under competitive conditions will be higher than under monopoly, and price in the competitive industry will be lower than in monopoly. In other words, output under monopoly is lower and price is higher compared to the competitive industry.

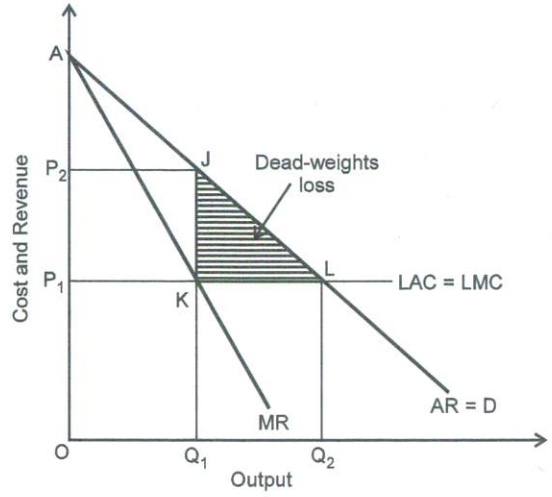


Fig. 3.22 *Price and Output under Monopoly and Perfect Competition*

NOTES

For the purpose of comparison, let us suppose that both monopoly and competitive firms are faced with identical cost and revenue conditions. Given the cost and revenue conditions, the perfectly competitive industry will produce OQ_2 at which its $LAC = LMC = AR$. Its price will be OP_1. On the other hand, the monopoly firm produces and output that equalizes its LMC and MR. Thus, a monopoly firm produces OQ_1 and charges prices OP_2. The comparison of prices and outputs under monopoly and perfect competition stands as follows:

Table 3.2 Comparision of Prices and Outputs

Variable	Monopoly	Competitive	Comparison
Output	OQ_1	OQ_2	$OQ_1 < OQ_2$
Price	OP_2	OP_1	$OP_2 > OP_1$

(ii) Loss of Social Welfare: On the basis of the above conclusion, it is alleged that monopoly firms are less efficient than competitive firms. Monopoly causes loss of social welfare and distortions in resource allocation. The loss of social welfare is measured in terms of loss of consumer's surplus. The total consumer's surplus equals the difference between the total utility which society gains and the total price which it pays for a given quantity of goods. As shown in Figure 3.22, if the industry is perfectly competitive, the total output available to the society will be OQ_2 at price OP_1. The total price which society pays for OQ_2 is given by the area $OP_1LO_2 = OP_1 \times OQ_2$. The total utility which it gains from the output OQ_2 is given by the area $OALQ_2$ which, in Marshalliam terminology, is the value which society would be willing to pay for output OQ_2. Thus, consumer's surplus = area $OALQ_2$ − area OP_1LQ_2 = area AP_1L.

If the industry is monopolised, the consumer's surplus is reduced to AP_2J. Thus, the total loss of consumer's surplus under monopoly is

$$AP_1L - AP_2J = P_2JLP_1$$

Of this total loss of consumer's surplus, P_2JKP_1 goes to the monopolist as monopoly or pure profit. The remainder JKL goes to none, and therefore, it is termed as dead-weight loss to the society caused by monopoly.

3.4.3 Price Discrimination by Monopoly

The theory of pricing under monopoly, as discussed, gives the impression that once a monopoly firm fixes up the price of its product, the same price is charged from all the consumers. This, however, may not be the case. A monopolist, simply by virtue of its monopoly power, is capable of charging different prices from different consumers or groups of consumers. When the same (or slightly differentiated) product is sold at different prices to different consumers, it is called price discrimination. When a monopolist sells the same product at different prices to different buyers, the monopoly is called discriminatory monopoly.

Consumers are discriminated in respect of prices on the basis of their incomes or purchasing powers, geographical location, age, sex, quantity they purchase, their association with the sellers, frequency of visits to the shop, the purpose of the

NOTES

use of the commodity or service, and on other grounds which the seller may find suitable.

A common example of consumers being discriminated on the basis of their incomes is found in medical and legal professions. Consulting physicians and lawyers (having excess capacity) charge different fees from different clients on the basis of their paying capacity. The government charges different rates of tariffs for different grades and purpose of units of electricity consumed. Price discrimination on the basis of age is found in railways, roadways and airways: children below 15 years are charged only half the adult rates. Price discrimination on the basis of quantity purchased is very common. It is generally found that private businessmen charge lower price (or give discount) when bulk-purchase is made. In case of public utility services, however, lower rates are charged when commodity or service is consumed in smaller quantity, for example, lower rates on the first few calls by the telephone owners, and no surcharge on electricity upto certain level of consumption. The most common practice of price discrimination is found in cinema shows, musical concerts, game-shows, etc.

For the purpose of price discrimination, the product or service in question may be identical or slightly modified. For example, services of consulting physicians and lawyers are identical. The services of railways, roadways and entertainment shows may be slightly modified by providing more comfortable seats for the purpose of price discrimination. The modification in service may involve some additional cost. However, price differentials are much more than are justified by cost differentials.

Although price discrimination is the most common practice under monopoly, it should not mean that this practice exists only under monopoly. Price discrimination is also quite common in other kinds of market structures, particularly where market imperfection exists. Most business firms discriminate between their customers on the basis of personal relationship, quantity purchased, duration of their association with the firm as buyers, and so on.

Necessary conditions for price discrimination

The market for different class of consumers must be separable so that buyers of low-price market are not in a position to resell the commodity in the high-price market for such reason as (i) geographical distance involving high cost of transportation, e.g., domestic versus foreign markets; (ii) exclusive use of the commodity, e.g. doctor's services and entertainment shows, and (iii) lack of distribution channels, e.g., transfer of electricity and gas.

If market is divided into submarkets, the elasticity of demand must be different in each submarket. The purpose of price discrimination is to maximize the profit by exploiting the markets with different price elasticities. It is the difference in price elasticities that provides opportunity for price discrimination. If price elasticities of demand in different markets are the same, price discrimination would not serve the objective of profit maximization.

NOTES

The seller must possess some monopoly over the supply of the product to be able to distinguish between different classes of consumers and to charge different prices.

Degrees of price discrimination

The degree of price discrimination refers to the extent to which a seller can divide the market and take advantage of it in extracting the consumer's surplus. According to Pigou, there are three degrees of price-discrimination practised by the monopolists: (i) *first degree* price discrimination; (ii) *second degree* price discrimination; and (iii) *third degree* price discrimination.

(a) **First degree price discrimination:** The discriminatory pricing that attempts to take away the entire consumers' surplus is called *first degree discrimination*. First degree discrimination is possible only when a seller is in a position to know the price each buyer is willing to pay. That is, he knows his buyer's demand curve for his product. Under perfect price discrimination, the seller sets the price at the highest possible level at which all those who are willing to buy the product at that price buy at least one unit each. When the consumer's surplus of this section of consumers is exhausted, he gradually lowers down the prices so that the consumer's surplus of the users of the subsequent units can be extracted. This method of pricing is continued until the whole consumer's surplus available at the price where $MR = MC$ is extracted. Consider the case of services of exclusive use, e.g., medical services. A doctor who knows or can guess the paying capacity of his patients can charge the highest possible fee from presumably the richest patient and the lowest fee from the poorest one. The first degree of price discrimination is the limit of discriminatory pricing.

(b) **Second degree price discrimination:** Under the second degree of discriminatory pricing, the firm charges different prices from different class of consumers—high, middle and low income consumers. The monopolist adopting the *second degree* price discrimination intends to siphon off only the major part of the consumer's surplus rather than the entire of it. The second degree price discrimination is feasible where (i) the number of consumers is large and price rationing can be effective, as in case of utility services like telephones and natural gas; (ii) demand curves of all the consumers are identical; and (iii) a single rate is applicable for a large number of buyers. As shown in Figure 3.23, a monopolist using a second degree price discrimination charges price OP_1 for the first few units, OQ_1, and price OP_2 for the next O_1Q_2 units, and price OP_3 for the next additional purchase of Q_2Q_3 units. Thus, by adopting a block-pricing system, the monopolist maximizes his total revenue (TR) as

$$TR = (OQ_1 \cdot AQ_1) + (Q_1Q_2 \cdot BQ_2) + (Q_2Q_3 \cdot CQ_3)$$

NOTES

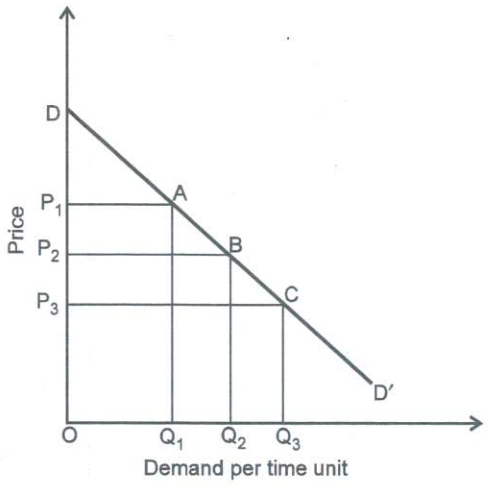


Fig. 3.23 Second Degree Price Discrimination

If a monopolist is restrained from price discrimination and is forced to choose any one of the three prices, OP_1, OP_2, or OP_3, his total revenue will be much less.

(c) **Third degree price discrimination:** When a profit maximizing monopoly sets different prices in different markets having demand curves with different elasticities, it is using third degree price discrimination. When a monopolist is faced with two or more markets, completely separated from each other—each having a demand curve with different elasticity—a uniform price cannot be set for all the markets without losing profits. The monopolist is, therefore, required to allocate total output between the different markets so that profit can be maximized in all the markets. The profit in each market would be the maximum only when $MR = MC$ in each market. The monopolist, therefore, divides total output between the markets so that in all the markets $MR = MC$. The process of allocation of output and determination of price for different markets is illustrated in Figure 3.24. Suppose a monopolist has to sell goods in only two markets, *A* and *B*. The two markets are so separated that resale of commodity is not possible. The demand curve (D_a) and marginal revenue curve (MR_a) given in Figure 3.24(a) represent the *AR* and *MR* curves in market *A*, and curves D_a and MR_b in Figure 3.24(b) represent *AR* and *MR* curves, respectively, in market *B*. The horizontal summation of demand curves D_a and D_b gives the total demand curve for the two markets, as shown by the curve $AR = D$, and horizontal summation of MR_a and MR_b is given by the curve *MR* (Figure 3.24). The firm's marginal cost is shown by the curve *MC*, which intersects *MR* at point *E*. Thus, optimum level of output for the firm is determined at *OQ*. At this level of output, $MR = MC$. Since the whole of *OQ* cannot be profitably sold in any one market because of their limited size, the firm has to allocate the output between the two markets.

The monopolist allocates output *OQ* between the two markets in such proportions that the necessary condition of profit maximization is satisfied in both the markets. That is, in both the markets, $MC = MR$. The profit

NOTES

maximizing output for each market can be obtained by drawing a line from point E and parallel to X-axis, through MR_b and MR_a. The points of intersection on curves MR_a and MR_b at points a and b, respectively, determine the optimum share for each market. As shown in Figure 3.24, the monopoly firm maximizes its revenue in market A by selling OQ_a units at price AQ_a and by selling OQ_b units in market B at price BQ_b.

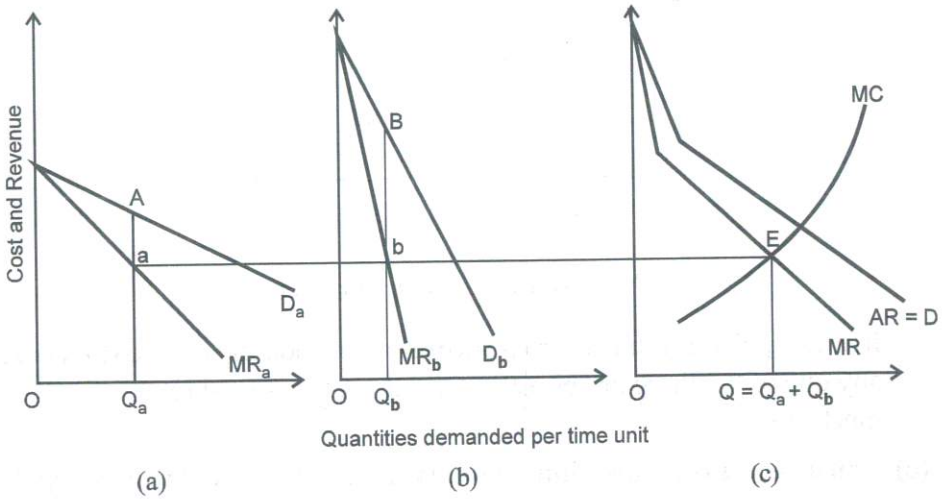


Fig. 3.24 *Third Degree Price Discrimination*

The firm's total equilibrium output $OQ = OQ_a + OQ_b$. Since at OQ_a, $MR_b = MC$ in market A, and at OQ_b, $MR_b = MC$ in market B,

$MC = EQ = MR_a = MR_b$

Thus, the equilibrium condition is satisfied in both market segments, and the monopoly firm adopting the *third degree* method of price discrimination maximizes its profits.

The third degree method of price discrimination is most suitable where the total market is divided between the home and foreign markets. However, it is not limited only to domestic and foreign markets. It may be suitably practised between any two or more markets separated from each other by any or more of such factors as geographical distance, transport barriers or cost of transportation, and legal restrictions on the inter-regional or interstate transportation of commodities by individuals.

Is price discrimination justified?

Price discrimination has been condemned as illegal and immoral. The objection is that why charge higher price from some and lower price from others while there is no extra advantage to those who pay higher price, or why benefit some at the cost of some others? In the United Kingdom and the United States, railways were prohibited from charging discriminatory rates. Discriminatory pricing has also been criticized as a destructive tool in the hands of a monopoly. For example, in the past, large corporations had sought to use price discrimination to prevent the growth of competition. Besides, price discrimination may cause malallocation of resources and hence may be deterrent to social welfare. This is, however, not the

case always. In some cases price discriminations is socially advantageous. In fact as Lipsey has observed, 'whether an individual judges price discrimination to be good or bad is likely to depend upon the details of the case as well as upon his own personal value judgements.' He adds, 'Certainly there is nothing in economic theory to suggest that price discrimination is always in some sense worse than non-discrimination under conditions of monopoly or oligopoly.'

Price discrimination is, however, considered to be desirable in certain specific cases on the following grounds:

(*i*) Goods and services are essential for the society as a whole but their production is uneconomic in the sense that long-run average cost curve (*LAC*) lies much above the aggregated market demand curve, as shown is Figure 3.25. Such goods and services cannot be produced. However, production of such goods and services can be possible if price discrimination is permitted. Price discrimination thus becomes essential for the survival of the industry.

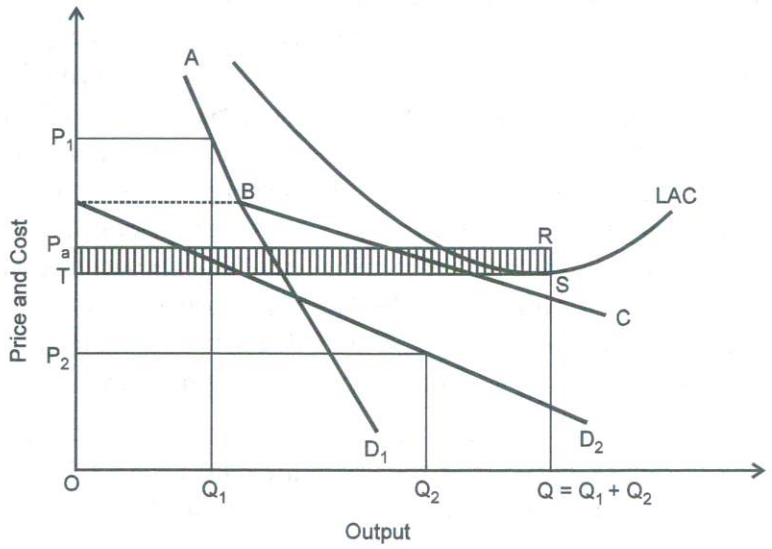


Fig. 3.25 *Price Discrimination for Industry's Survival*

Suppose that, for example, (i) there are two markets I and II, (ii) individual demand curves for the two markets, I and II, are given as D_1 and D_2, (iii) market demand curve is given by *ABC*, and (iv) the long-run average cost curve is given by *LAC* (Figure 3.25). Note that *LAC* lies throughout above the total demand curve *ABC*. Therefore, production is not possible if one price is to be charged. Nevertheless, if price discrimination is adopted and prices are so charged in the two markets that the total revenue exceeds *LAC* at some level of output, then monopoly may profitably survive to the advantage of the society. Let us suppose that the monopolist sets price OP_1 in the market I in which demand is less elastic and OP_2 in market II in which demand is highly elastic. He would sell OQ_1 units at price OP_1 in market I and OQ_2 at price OP_2 in market II. His total output would then be at $OQ = OQ_1 + OQ_2$. His total revenue (*TR*) would be

$$TR = (OP_1 \times OQ_1) + (OP_2 \times OQ_2)$$

NOTES

and suppose

$$AR = (OP_1 \times OQ_1 + OP_2 \times OQ_2)/OQ = OP_a$$

At output OQ, the LAC is OT. Thus his total cost,

$$TC = OQ \times OT = OQST$$

and his total revenue,

$$TR = OQ \times OP_a = OQRP_a$$

Since $OQRP_a > OQST$, the monopoly firms not only covers its cost but also makes excess profit. Its total profit,

$$p = OQRP_a - OQST = P_aRST$$

This kind of situation arises mostly in public utility services, like railways roadways, and post and telegraph services, in which high-paying sector of the market subsidises the low-paying sector. But, if low-paying sector is not subsidised, no production would be possible.

(*ii*) Discriminatory pricing can be adopted with justification where a uniform, single profitable price is likely to restrict the output and deprive many (particularly the people of lower income groups) of the essential goods or service. For example, if doctors in private practice, who often charge discriminatory price for their services, are asked to charge a uniform fee from all the patients, they would charge a fee high enough to maintain the level of their income. The high fee may deprive the poor of the doctor's service and may force them to opt for inferior or inadequate treatment. The result of the uniform high fee will be that the rich patients who can pay a still higher fee gain as they pay a price lower than what they could afford, and on the other hand, poor patients are deprived of proper medical service.

(*iii*) There may be cases where a section of consumers gains more than the people of other sections from the use of the same product. For example, from the use of electricity, factory owners gain more than the households.

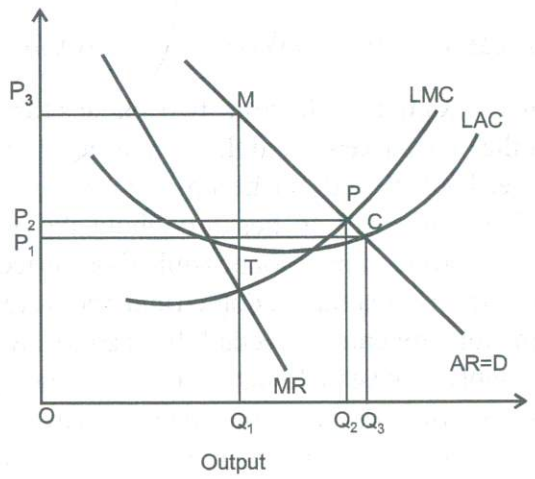


Fig. 3.26 Government-Regulated Monopoly

In such cases, uniform price would be unjustified from a normative point of view, provided the objective is not to restrain the domestic consumption of electricity

and spare it for productive purposes. There is, on the other hand, full justification for discriminatory pricing of electricity.

Government regulation of monopoly prices

The existence of monopolies in a market economy is criticized on the grounds that they restrict production and consumption, widen income and wealth disparities, exploit consumers and employees, cause distortions in allocation of resources, reduce the prospect of employment, and cause loss of social welfare. In most countries, therefore, there is general apathy towards the monopolies. Consequently, governments in the market economies attempt to control and regulate monopolies to the advantage of the society. There are various measures—direct, indirect, legal and otherwise—to control and regulate the monopolies.

Price regulation is a common feature in case of natural monopolies. When the size of the market is small relative to the optimum size of the firm, market size cannot support more than one firm of optimal size. The monopoly in such a market is a natural monopoly. The natural monopoly is thus protected by market size itself. The government may either nationalize such monopolies or regulate their prices so as to eliminate the excess profits. If the government intends to regulate the monopoly price, the question arises: what price should be fixed for the monopolist to charge? The two alternative prices that have been suggested are: (i) that allows some excess profit to the monopolist, and (ii) that allows only normal profit to the monopolist. Both the alternative prices, along with their repercussion on output, are illustrated in Figure 3.26. An unregulated monopoly would produce OQ_1 units, charge price OP_3, and make excess profit of $MT = MQ_1 - TQ_1$ per unit. If monopoly price is regulated, one possible price is given at point P where $LMC = AR$, the price being OP_2 ($= PQ_2$). Alternatively, price may be fixed at point C where $AR = LAC$ and price = OP_1 ($= CQ_3$). When OP_1 is the price set for the monopolist, only a normal profit is allowed to the firm, but output is maximum possible under the given cost and revenue conditions. If price is fixed at OP_2, the monopolist gets some excess profit, but the output is less than that at price OP_1. In both the cases, however, the total output under regulated monopoly is much higher than that under unregulated monopoly. Which of the two alternative prices (OP_1 and OP_2) is more appropriate is a matter of debate.

3.4.4 Measures of Monopoly Power

It is only in rare cases that monopolies have absolute power. Monopoly power varies from industry to industry. The degree of monopoly power matters a great deal in pricing and output decisions of a monopolist. Besides, measuring monopoly power is required also in connection with control and regulation of monopolies. let us discuss here the various measures of monopoly power.

Measuring monopoly power has been a difficult proposition. The efforts to devise a measure of monopoly power have not yielded any universal or non-controversial measure. As Hunter has observed, 'The idea of devising a measure of monopoly power, with reference both to its general incidence and to particular situation has been and probably always will remain an attractive prospect for economists who wish to probe in this field.' If not for any other reason, then for

NOTES

'sheer intellectual curiosity', economy theorists feel compelled to work on this problem, as they 'could not with good conscience go on talking about 'great' or "little" monopoly power or about various degrees of monopoly without trying to ascertain the meaning of these words.'

Therefore, to devise at least a 'conceivable' measure of monopoly, even if 'practical' measurement is impossible, continues to interest the economists, for at least two reasons. **First**, apart from intellectual curiosity, people would like to know about the economy in which they live, its industrial structure, and the industries from which they get their supplies. **Second**, growth of monopolies have forced governments of many countries to formulate policies and devise legislative measures to control and regulate monopolies. If the government is to succeed in its policy of restraining monopoly, it must have at least some practicable measure of monopoly and monopolistic trade practices.

Although economists have devised a number of devices to measure the degree of monopoly power, none of the measures is free from flaws. Yet, the various measures do provide an insight into the monopoly power and its impact on the market structure. Besides, they also help in formulating an appropriate public policy to control and regulate the existing monopolies. let us discuss briefly the various measures of monopoly power suggested by the economists.

(i) **Number-of-firms criterion:** One of the simplest measures of degree of monopoly power is to count the number of firms in an industry. The smaller the number of firms, the greater the degree of monopoly power of each firm in the industry, and conversely, the larger the number of firms, greater the possibility of absence of monopoly power. As a corollary of this, if there is a single firm in an industry, the firm has an absolute monopoly power. This criterion seems to have been derived from the characteristics of the perfect competition in which the number of firms is so large that each firm supplies only an insignificant proportion of the market and no firm has any control on the price.

This criterion, however, has a serious drawback. The number of firms alone does not reveal much about the relative position of the firms within the industry because (i) firms are not of equal size, and (ii) their number does not indicate the degree of control each firm exercises in the industry. Therefore, the numerical criterion of measuring monopoly power is of little practical use.

(ii) **Concentration ratio:** The *concentration ratio* is one of the widely used criteria used for measuring monopoly power. The concentration ratio is obtained by calculating the percentage share of the largest group of the firms in the total output of the industry. The number of firms chosen for calculating the ratio usually depends on some fortuitous element—normally the census of production arrangements of the country concerned. In Britain, the share of the largest three firms of a census industry, and in the USA, the share of the largest four firms is the basis of calculating concentration ratio. Apart from the share of the largest firms in the industry output, 'the size of the firm and the concentration of control in the industry may be measured in terms of production capacity, value of assets, number of employees or some

NOTES

other characteristics.

However, these measures too are, however, not free from drawbacks as they involve statistical and conceptual problems. For example, production capacity may not be straightaway used as it may include 'unused, obsolete or excess capacity'; the value of assets involves valuation problem as accounting method of valuation and market valuation of assets may differ. Employment figures may not be relevant in case of capital intensive industries. The use of such figures may be misleading. The two other convenient measures are 'gross output value' or 'net output' (value added). But the former involves the risk of double counting, and the latter, the omission of inter-establishment transfers.

Another important objection to these measures of degree of monopoly power is that they do not take into account the size of the market. Size of the market may be national or local. A large number of firms supplying the national market may be much less competitive than the small number of firms supplying the local market. Because, it is quite likely that the national market is divided among the thousand sellers so that each seller has status of a monopolist in his own area.

The most serious defect of concentration ratio as an index of monopoly power is that it does not reflect the competition from other industries. The degree of competition is measured by the elasticity of substitution between the products of different industries. The elasticity of substitution may be different under different classification of industries. Therefore, an industry with concentration ratio under one classification of industries may have a very low elasticity of substitution and hence a high degree of monopoly. But, if classification of industries is altered, the same industry with a high concentration ratio may have a very low elasticity of substitution, and hence, may show a low degree of monopoly.

(*iii*) **Excess profitability criterion:** JS Bain and, following him, many other economists have used *excess profit* as a measure of monopoly power. If profit rate of a firm continues to remain sufficiently higher than all opportunity costs required to remain in the industry, it implies that neither competition among sellers nor entry of new firms prevents the firm from making a pure or monopoly profit. While calculating the excess profit, the opportunity cost of owner's capital and margin for the risk must be deducted from the actual profit made by the firm. Assuming no risk, the degree of monopoly may be obtained by calculating the divergence between the opportunity costs (O) and the actual profit, (P), as $(P - O)/P$. If $[(P - O)/P] = O$, there exists no monopoly, and it $[(P - O)/P] > O$, there is monopoly. The higher the value of $(P - O)/P$, the greater the degree of monopoly.

Another measure of degree of monopoly based on excess profitability has been provided by AP Lerner. According to him, the degree of monopoly power (MP) may be measured as

NOTES

$$MP = \frac{P - MC}{P}$$

where P = price, MC = marginal cost. Since for a profit maximizing firm, $MR = MC$, Lerner's measure of monopoly power MP may also be expressed as

$$MP = \frac{P - MR}{P}$$

Since $P/(P - MR) = e$, $(P - MR)/P = 1/e$, i.e., MP equals to the reciprocal of elasticity. Thus, Lerner's measure of monopoly power may be expressed also as $MP = 1/e$. It may thus be inferred that lower the elasticity, the greater the degree of monopoly, and vice versa. According to Lerner's formula, monopoly power may exist even if firm's $AR = AC$ and it earns only normal profit.

Lerner's formula of measuring the degree of monopoly power is considered to be theoretically most sound. Nevertheless, it has been criticized on the following grounds:

First, it is suggested that any formula devised to measure degree of monopoly power should bring out the difference between the monopoly output and competitive output or the 'ideal' output under the optimum allocation of resources. The divergence between P and MC used in Lerner's formula does not indicate the divergence between the monopoly and the 'ideal' output. Lerner has possibly used the divergence between P and MC as the substitute for the divergence between monopoly and 'ideal' output. 'This substitution of a price-cost discrepancy for a difference between actual and ideal output is probably the greatest weakness of formula which is supposed to measure deviation from the optimum allocation of resources.'

Second, price cost discrepancy may arise for reasons other than monopoly, and price and cost may be equal or close to each other in spite of monopoly power.

Third, since data on MC are hardly available, this formula is of little practical use for policy purposes.

(*iv*) **Triffin's cross-elasticity criterion:** Triffin's criterion seems to have been derived from the definition of monopoly itself. According to this criterion, cross-elasticity is taken as the measure of degree of monopoly—the lower the cross-elasticity of the product of a firm, the greater the degree of its monopoly power. However, this criterion is based on the inter-relationships between the individual firms and indicates only the relative power of each firm. It does not furnish a single index of monopoly power.

Check Your Progress

4. State one important feature of pure monopoly.
5. What are patent rights?
6. Define discriminatory monopoly.

NOTES

3.5 PRICE-OUTPUT UNDER MONOPOLISTIC COMPETITION

The model of monopolistic competition developed by Edward H Chamberlin presents a more realistic picture of the actual market structure and the nature of competition. Let us first discuss briefly the nature of the market structure and monopolistic competition among the firms.

Monopolistic Competition is a market structure in which a *large number* of sellers sell *differentiated* products which are close, but not perfect, substitutes for one another. Monopolistic competition combines the characteristics of perfect competition and monopoly.

The assumptions of the monopolistic competition are the same as those of pure competition, with an exception of homogeneity of products. While pure competition model assumes that products are *homogeneous* in every possible dimension, monopolistic competition model assumes that products are *differentiated*. The product of each firm is so differentiated from those of other firms that consumers are able to distinguish the product of a firm from those of others. For example, consumers know for sure the difference between different brands of mobile phones, e.g., Nokia, Sony, Samsung and Reliance. Since each firm produces a product distinguishable from that of other firms, each firm holds a monopoly power over its own products.

Although products are differentiated, they remain a close substitute for one another. Product differentiation, along with other features, creates condition for competition among the firms which are monopolists in case of their own products. This kind of competition is the genesis of monopolistic competition.

Features of monopolistic competition

The main features of monopolistic competition areas follows:

1. The number of firms is fairly large.
2. Firms produce and sell differentiated products—each product being a close substitute for another.
3. There is free entry and free exit of firms to and from the industry.
4. Factors of production enjoy free mobility, at least theoretically.
5. Both buyers and sellers have full knowledge about the market.
6. There is no collision between monopolistic firms.

Foundations of the monopolistic competition model

(i) **Assumptions:** Chamberlin's model of monopolistic competition is based on the following assumptions:

1. There is a large number of buyers and sellers in the market.
2. Each seller sells a product differentiated from that of others.
3. The differentiated products are close, not perfect, substitute for one another.

NOTES

4. There is free entry and free exit of firms.
5. The firms seek to maximize their profits in both short and long-runs.
6. Technology and factor prices are given and the firms are aware of revenue and cost curves.

(ii) **Product differentiation and the demand curve:** Chamberlin has defined *product differentiation* in the following words: 'A general class of product is differentiated if any significant basis exists for distinguishing the goods (or services) of one seller from those of others. Such a basis may be real or fancied, so long as it is of any importance whatever to buyers, and leads to a preference for one variety of the product over another.... Differentiation may be based upon certain characteristics of the product itself, such as exclusive patented features, trademarks, trade names, peculiarities of the package or container, if any, or singularity in quality, design, colour or style. It may also exist with respect to the conditions surrounding its sales. In retail trade, these conditions include such factors as the convenience of the seller's location, the general tone or character of his establishment, his way of doing business, his reputation for fair dealing, courtesy, efficiency, and all the personal links which attach his customers either to himself or to those employed by him.' So far as these and other tangible and intangible factors create consumers' preference for one product over the others, the products are virtually differentiated.

Thus, product differentiation is primarily intended to make consumers distinguish the product of one producer from that of the other producers in the industry. When the consumers are able to distinguish one product from the others, they may develop a preference or brand loyalty for one product over the others. Once preference for a product is created, it alters the course of demand curve for the product. In the ultimate analysis, product differentiation leads to a change in demand curve for the product from a horizontal demand line (as under pure competition) to a downward sloping demand curve. The downward sloping demand curve enables the sellers to exercise some discretion in determining the price of his product.

(iii) **Cost curves and selling cost:** In his model of monopolistic competition, Chamberlin has assumed the traditional U-shaped cost curves, viz., *AC, AVC* and *MC*. In addition, he has introduced a new cost, i.e., *selling cost*. 'Selling costs are defined as costs incurred in order to alter the position or the shape of the demand curve for a product.' Selling costs include all the expenses that are intended to promote the sales, including cost of advertisement, salesmen's salaries, expenses of sales department, margins granted to dealers—wholesalers and retailers—and on window displays and demonstration of new goods. Selling costs affect demand curve in the following two ways:

First, selling costs make the demand curve for the product shift upward by informing consumers about the availability of the product and by increasing consumer's preference for the product.

Second, selling cost makes the demand curve less elastic by strengthening the consumers' preference for the product.

NOTES

Chamberlin assumes average selling cost to be U-shaped, that is, selling cost per unit of sales initially decreases but eventually increases. Thus, the average selling-cost curve has a shape similar to the *AC* curve.

(iv) **concept of industry and product groups:** Under monopolistic competition, products are so differentiated that each product is distinguishable from others, and each firm is, in a sense, an industry in itself, exactly as a monopoly firm is an industry in itself. The heterogeneity of the products, therefore, causes a problem in analytical treatment of the industry. It may be recalled that, in case of homogeneous products, demand curve for an industry can be obtained by adding individual demand curve. But in case of heterogeneous products, the demand for individual products cannot be added to obtain market demand and supply curves.

For this reason, Chamberlin attempted to redefine the industry for his analytical purpose. He defined the monopolistically competitive industry as a 'group' of firms producing a 'closely related' commodity, referred to as *product group*. The product of the 'group' must be *close, technological and economic substitutes*. The two products are *technological substitutes* for each other if they technically satisfy the same want, e.g. personal computers, soaps, toothpastes, automobiles and TV sets. The two products are considered as *economic substitutes* for each other if they satisfy the same want and have *more or less* the same price. For example, all brands of TV sets are economic substitutes for one another. But flat TV sets are not economic substitutes for ordinary ones since their prices are widely different, though they remain technological substitutes. Operationally, the product group may be defined as the group of firms whose products have between themselves high *price* and *cross elasticities*. This definition, although theoretical plausible, involves the problems of measuring cross-elasticities and determining the degree of cross-elasticities that can make a commodity admissible to the group. Determining the product group would therefore involve subjective judgement.

Price and output determination under monopolistic competition

Chamberlin's theory of price and output determination under monopolistic competition is basically the same as that under monopoly with a difference, of course. While under monopoly, demand and cost curves are both assumed to be given, under monopolistic competition, firms are assumed to indulge in competition to change the slope of the demand curve or to make it shift rightward, given the cost curves.

They seek to make these changes in the demand curve by any or all of the following measures:

(i) Change in the price of the product

(ii) Change in the nature of the product

(iii) Change in the advertisement outlays

As to price change, since a monopolistically competitive firm faces an elastic demand curve similar to one faced by a monopoly firm, it has the option to raise

NOTES

the price and sell less or to lower the price and sell more. But, it fixes a price that maximizes its profits. As with change in product, the sales can be promoted by change in the quality of product through technical changes, introduction of a new design, use of better material, use of new package or containers, prompt and courteous services, credit facilities, etc. In addition, the firm may influence its volume of sales by increasing advertisement expenditure so that more consumers are attracted to the product. Increase in advertisement expenditure also increases the selling price. The firm is, therefore, required to so adjust its price and output that its profits is maximum.

While adjustment between price and output for profit maximization is a short-run phenomenon, changes in the quality of the product and advertisement expenses are long-run phenomena.

A. Firm's short-run equilibrium

While monopolistic competition is characteristically closer to perfect competition, it is closer to monopoly in regard to pricing and output determination. Like a monopolist, a monopolistic competitor faces a downward sloping demand curve having a smaller slope. This demand curve is the product of (i) strong preference of a section of consumers for a particular product; and (ii) the quasi-monopoly of the seller over the supply. The strong preference or loyalty of the consumers gives the seller an opportunity to raise the price and yet retain some customers. Moreover, since each product is a close substitute for another, sellers attract the consumers of other products by lowering down their prices.

As mentioned above, short-term analysis of pricing and output determination under monopolistic competition is similar to price and output determination under monopoly. The short-term equilibrium analysis is primarily the adjustment of price and output to the given cost and revenue conditions. The short-run price and output adjustment is illustrated in Figure 3.27. The $AR = D$ and MR curves show the revenue conditions and SAC and SMC curves show the cost conditions faced by the firm in the short-run.

As shown in Figure 3.27, the necessary condition of profit maximization, i.e., MR must be equal to MC, is fulfilled at output OQ. This output can be sold at price PQ, so the price is also determined. At this output and price, the firm earns a maximum economic profit, shown by the rectangle P_1PMP_2.

The economic profit per unit (PM) exists in the short-run because new firms cannot enter the industry. However, the rate of profit would not be the same for all the firms under monopolistic competition because of difference in the elasticity of demand. For the same reason, product price will be different for the different products, though price differentials will be only marginal.

In the short run, the firms may attempt to maximize their profits by changing the nature of the product and by increasing advertisement expenditure. But, since there are many close substitutes, neither of the strategies would be of much avail in the short run. If the firms do adopt these strategies, they would do so only to maximize their profits.

NOTES

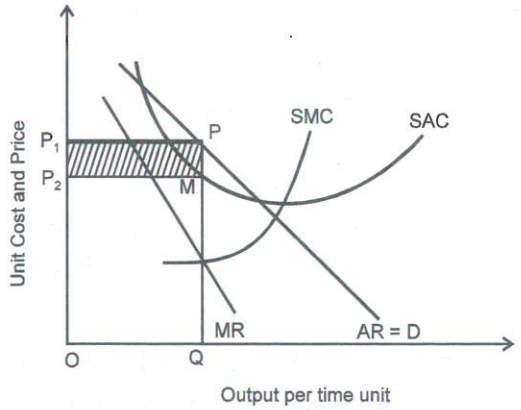


Fig. 3.27 Short-run Equilibrium Under Monopolistic Competition

B. Long-run equilibrium under monopolistic competition

The conditions faced by a firm of the 'product-group' under monopolistic competitions are different from those in the short run. In the long run, a firm can expand its plant size and if there are no barriers to entry, new firms will enter the product group, especially when existing firms are making pure profits. Besides, the firms get the opportunity to change the nature and position of the demand curve for their product by (i) changing the quality of the product, and (ii) incurring a large amount of advertisement expenditure. Since conditions change in the long run, the firms can manoeuvre their price and output in order to maximize their profits in the long run too.

Basic assumptions

Chamberlin has made the following explicit and implicit assumptions in order to develop his theory of monopolistic competition under long-run conditions.

(*i*) The basic assumption is that there are a large number of firms producing differentiated products which are close substitutes for one another.

(*ii*) The number of firms in the product group is so large that each firm expects its manoeuvering of prices and output to go unnoticed by the rival firms.

(*iii*) One of the heroic assumptions of Chamberlin is that both demand and cost curves for all the products are uniform throughout the group. That is, all firms have identical revenue and cost curves.

(*iv*) His second heroic assumption is that consumer's preferences are evenly distributed among the different varieties and that differences between products are not such as to give rise to differences in cost.

Under these assumptions, Chamberlin develops three models of equilibrium:

(i) To analyse equilibrium with free entry of new firms to the industry with non-price competition

(ii) To analyse equilibrium under price competition, assuming no entry or exit

(iii) To present a combined analysis of the first and the second models

Let us study the three models in detail.

NOTES

(i) Long-run equilibrium with free entry of new firms

The long-run equilibrium of firms under the assumption of free entry of new firms is illustrated in Figure 3.28. The *LAC* and *LMC* are the cost curves faced by the firms. The initial *AR* and *MR* curves (i.e., prior to the entry of new firms) are given by $AR = D_2$ and MR_2. Given the cost and revenue curves, the firms will be in short-run equilibrium at point *E* where price is OP_2 and output OQ_2. The firms are making super-normal profits to the extent of *EB* per unit of output.

The existence of super-normal profits attracts new firms to the product group. With the entry of new firms, the sale of each firm in the group decreases. Consequently, the demand curve (or *AR* curve) for the firm shifts leftward and so does the *MR* curve. This forces the firms to adjust their price and output to the new *AR* and *MR* conditions, cost conditions remaining the same. If still there exists super-normal profit more new firms join the product group, and existing firms will be forced to readjust their price and output to another equilibrium position. This process will continue until the demand curve becomes tangent to the *LAC*, and all the firms earn only normal profits. The final equilibrium position of the firms in the long-run under monopolistic competition is shown at point *A*, where price is OP_1 and output OQ_1. At this price and output, all firms make only normal profit. Therefore, there is no incentive for the new firms to enter the industry. The equilibrium will, therefore, be stable at point *A*.

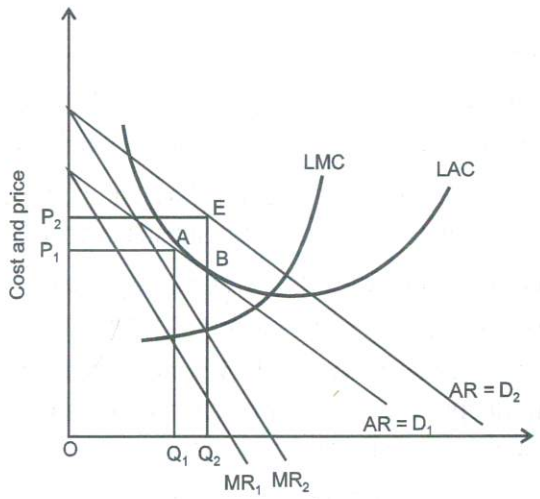


Fig. 3.28 *Introduction of Second Demand Curve*

(ii) Long-run equilibrium with price competition

In his model of long-run equilibrium with price competition, Chamberlin assumes that the number of existing firms in the product group is optimal. That is, the number of firms is compatible with long-run equilibrium of the industry. There is no entry or exit of the firms. In this case, the equilibrium analysis has been accomplished in two stages. In the first stage, a second demand curve is introduced to the model to incorporate the effects of competitive change in prices. In the second stage, the

NOTES

long-run equilibrium under the condition of optimal number of firms has been analysed.

To commence the analysis of stage one, let us assume that demand curve for the product of a firm is given by DD_1, as shown in Figure 3.29, and the firm is in equilibrium at point E, with price OP and output OQ. Now suppose, the firm contemplates a price reduction and assumes that the rival firms will not react to its price cut. By reducing its price, the firm expects to expand its sales on account of two factors: *first*, the demand for its product is elastic, and *second*, the consumers of other products will switch over to the product of this firm if other firms do not simultaneously reduce their prices. Thus, the firm can expect a substantial increase in its total sales.

For example, if the firm *reduces* its price by EM, the demand for its product increases by MB. Of this, MN is due to the elasticity of its original demand DD_1 and NB is due to the fact that some customers of other firms switch over to the product of this firm. On the other hand, if the firm *raises* its price and the rival firms do not raise their prices, it loses its market partly because its demand curve is elastic and partly because its customers switch over to other products which become automatically cheaper. Thus, another demand curve DD_2 emerges. This is called as the *second demand curve*. But, if all the firms change their prices simultaneously, this advantage to an individual firm is lost and all the firms return to the original demand curve DD_1.

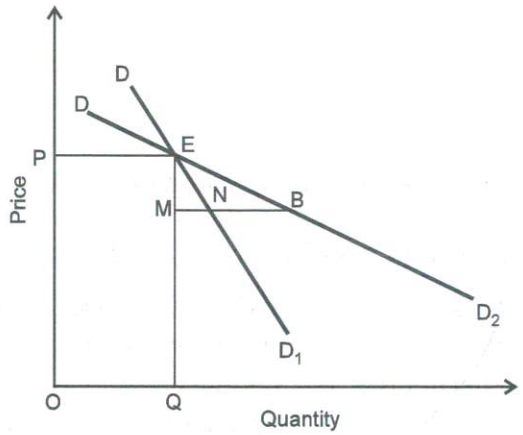


Fig. 3.29 *Introduction of Second Demand Curve*

Having introduced the second demand curve, we can proceed to present Chamberlin's long-run equilibrium with price competition, i.e., *second stage* of his analysis. The long-run equilibrium with price competition is presented in Figure 3.30. The curves DD' and dd_1 are the two demand curves and LAC is the long-run average cost curve of a 'typical' firm of the group. Let the initial short-run equilibrium of the firms of the group be at point P with price OP_2 and output OQ_1. At this price and output, the firm makes the total abnormal profit represented by the area P_2PMC.

Although in equilibrium, each firm regards dd_1 to be its demand curve and believes that it can increase its profits by reducing the price and expanding the

output. The basis of this belief is the elasticity of their assumed demand curve dd_1. In an attempt to increase profits, each firm reduces its price, expecting to move along the demand curve dd_1. However, instead of moving along dd_1, the firms move along the market demand curve DD', because all of them reduce their prices simultaneously. However, according to Chamberlin, the firms do not learn from their past experience and each firm sticks to its own belief that the demand curve (dd_1) for its product is more elastic than the market demand curve (DD'). Therefore, the firms go on reducing their prices independently and their assumed demand curve (dd_1) shifts downward. This process continues until the downward shift in dd_1 makes it tangent to the LAC curve, as shown by dd_2. A further reduction in price will make firms incur loss. Therefore, reduction in price below OP_1 is not desirable. Thus, the long-run equilibrium of firms takes place at E, where each firm produces OQ_2 and fixes its price at OP_1.

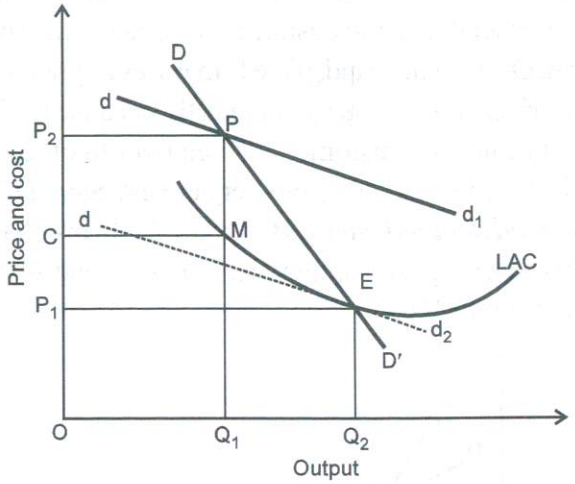


Fig. 3.30 *Long-run Equilibrium Price Competition*

(iii) Long-run Equilibrium with Free Entry and Price Competition

The equilibrium of the firm with free entry and with price competition separately. Now, bring together the two equilibrium analyses and explain Chamberlin's third and the final model of firms' equilibrium under monopolistic competition. According to Chamberlin, the ultimate equilibrium under monopolistic competition can be achieved through an integrated analysis of effects of free entry and price adjustments. The integrated analysis of equilibrium is presented in Figure 3.31.

In Figure 3.31, DD_1 is assumed to represent the initial demand curve and LAC to represent the long-run average cost curve. Let us suppose that the firms are initially in equilibrium at point B, and they make abnormal profits to the extent of vertical distance between DD_1 and the LAC. Since entry to the 'product group' is free, new firms are attracted by the industry. When new firms with slightly differentiated products enter the 'product group', the market share of each existing firm is reduced. Hence, their demand curve DD_1 shifts towards the left. Given the LAC, the leftward shift in the demand curve will continue until it becomes tangent to LAC, as shown by DD_3 in Figure 3.31, because till this point of tangency is

NOTES

reached, firms make abnormal profits and new firms continue to enter the 'product group'.

Thus, it might seem that the long-run equilibrium is attained at point A with output QQ_1 and price OP_3. This is however not the case. This is only half of the story, i.e., the influence of free entry. Let us now consider the competitive manoeuvering of price and its role in determining to the long-run equilibrium.

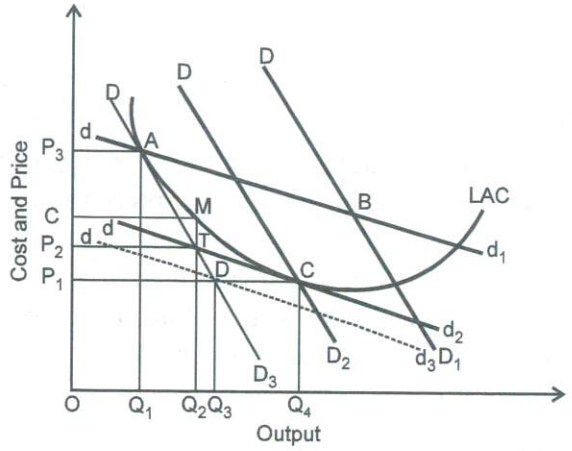


Fig. 3.31 *Long-run Equilibrium with Free Entry and Price Competition*

Once the firms reach point A, each firm thinks that its demand curve is dd_1, not DD_3. Each firm believes that it can increase its profit by reducing the price and thereby increasing the sales. Therefore, in their attempt to increase their profits, they reduce prices of their products simultaneously because each firm has the same incentive to do so. As a result, their subjective demand curve (dd_1) slides downward to dd_2 and they incur losses. For example, if price is reduced to OP_2, the total loss equals the rectangle $CMTP_2$. It might seem that the firms could eliminate their loss by reducing the price to OP_1. But when all the firms reduce their price to OP_1—and they will do so under the assumption—their subjective demand curve dd_2 slides further down to dd_3, the dotted line that lies below the LAC. As a result, the firms make increasing losses. A temporary equilibrium will be attained at point D with output OQ_3, where all firms incur heavy losses. Consequently, the firms that are unable to sustain losses will eventually leave the industry. The remaining firms find their share in the market increasing. Therefore, DD_3 and dd_3 move to the right until DD_3 shifts to DD_2 and dd_3 shifts upward to the position of dd_2. Note that dd_2 intersects DD_2 at point C—the point where dd_2 is tangent to LAC. Thus, the long-run equilibrium is attained at point C, where it is stable because all firms earn only normal profits and, therefore, there is no entry or exit of the firms.

Analysis of selling cost and firm's equilibrium

As noted above, under monopolistic competition, products are differentiated; market is imperfect; and consumers are not fully aware of existence of a particular variety of the product. Therefore, producers always have an opportunity to advertise their product, attract more customers to their product, and create brand loyalties in the minds of consumers and thereby increase their share in the market. In fact,

the basic objective of advertising is to change the position and shape of the demand curve for the product of the advertising firm. Whether a firm succeeds in achieving these objectives depends also on the effectiveness of competitive advertising by the rival firms. But, one thing is obvious that advertising involves additional costs, which pushes the *AC* curve upward.

Apart from advertisement expenses, monopolistically competitive firms incur other costs on competitive promotion of their sales, e.g., salary to sales personnel, allowance to dealers, discounts to customers, expenses on displays, gifts to customers and free samples, and additional costs on attractive packaging of goods. All such expenses plus advertisement expenditure make a firm's *selling cost.*

Incurring selling cost increases sales, but at a varying degree. In the initial stage, the increase in sales may be greater than the proportionate increase in the selling cost. But eventually, it decreases. Consequently, the unit selling cost or the average selling cost (*ASC*) initially decreases (because of economies of scale) but ultimately increases. The *ASC* curve is, therefore, U-shaped, similar to the conventional short-run *AC* curve. It implies that total sales are subject to diminishing returns to increasing selling costs.

Selling costs and group equilibrium

To analyse group equilibrium of firms with selling cost, the main objective of the firm is to maximize its total profits. When they incur selling costs, they do so with the same objective in mind. All earlier assumptions regarding cost and revenue curves remain the same. The analysis of group equilibrium is presented in Figure 3.32. Suppose *APC* represents the average production cost and price is given at OP_3. None of the firms incurs any selling cost.

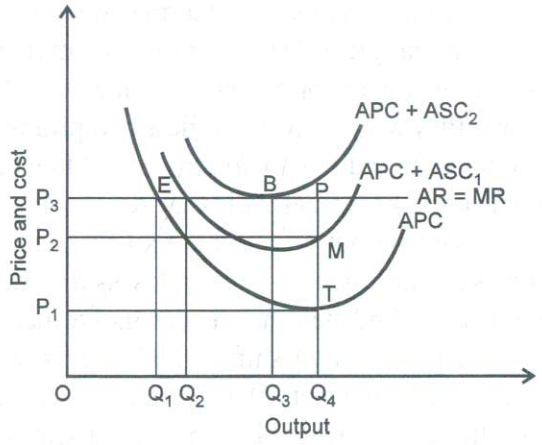


Fig. 3.32 *Selling Costs and Group Equilibrium*

Also, let all the firms be in equilibrium at point *E* where they make only normal profits. Now suppose that one of the firms incurs selling cost so that its *APC* added with average selling costs (*ASC*) rises to the position shown by the curve $APC + ASC_1$ and its total sales increases to OQ_4. At output OQ_4, the firm makes supernormal profits of P_3PMP_2. This profit is, however, possible only so long as other firms do not advertise their own products. If other firms do advertise their products and incur the same amount of selling cost, the initial advantage to

firm advertising first will disappear and its output will reduce to OQ_2. In fact, all the firms produce only OQ_2 units. But their short-sightedness impels them to increase their selling cost because they expect to reduce their *APC* by expanding their output. With increased selling cost, their $APC + ASC$ curve shifts further upward. This process continues until $APC + ASC_2$ becomes tangent to the $AR = MR$ line at point *B*. Beyond point *B*, advertising is of no avail to any firm, even if other firms do not advertise. The equilibrium will be stable at point *B* where each firm produces OQ_3. Note that the equilibrium output OQ_3 is greater than the initial output of OQ_1. In equilibrium, however, firms make only normal profits.

A critical appraisal of chamberlin's theory of monopolistic competition

Chamberlin's theory of monopolistic competition has been criticized on both theoretical and empirical grounds. Its theoretical or methodological weaknesses are as follows:

First, Chamberlin assumes that monopolistic competitors act independently and their price manoeuvering goes unnoticed by the rival firms. This assumption has been questioned on the ground that the sales of other firms are bound to be affected by the decisions of rival firms since their products are close substitutes for one another and, therefore, they are bound to react.

Second, Chamberlin's model implicitly assumes that monopolistically competitive firms do not learn from their past experience. They continue to commit the mistake of reducing their prices even if successive price reductions lead to increase in their losses. Such an assumption can hardly be accepted.

Third, Chamberlin's concept of industry as a 'product group' is ambiguous. It is also incompatible with product differentiation. In fact, each firms is an industry by virtue of their specialized and unique product.

Fourth, Chamberlin's heroic assumptions of identical cost and revenue curves are questionable. Since each firm is an industry in itself, there is a greater possibility of variations in the costs and revenue conditions of the various firms.

Finally, Chamberlin's assumption of free entry is also considered to be incompatible with product differentiation. Even if there are no legal barriers, product differentiation and brand loyalties are in themselves barriers to entry.

Empirical validity

So far as 'empirical validity' of Chamberlin's concept of monopolistic competition is concerned, it is claimed that it is difficult to find any example in the real world to which his model of monopolistic competition is relevant. Most markets frequently available in the real world may be classified under prefect competition, oligopoly or monopoly. It is, therefore, alleged that Chamberlin's model of monopolistic competition analyses an unrealistic market. Some economists, e.g., Cohen and Cyert, hold the position that the model of monopolistic competition is not a useful addition to economic theory because it does not describe any market in the real world.

Despite above criticism, Chamberlin's contribution to the theory of price cannot be denied. Chamberlin is the first to introduce the concept of *differentiated product* and *selling costs* as a decision variable and to offer a systematic analysis

NOTES

of these factors. Another important contribution of Chamberlin is the introduction of the concept of demand curve based on market share as a tool of analysing behaviour of firms, which later became the basis of the *kinked-demand* curve analysis.

3.6 PRICE AND OUTPUT UNDER OLIGOPOLY

Oligopoly is a form of market structure in which a *few* sellers sell differentiated or homogeneous products. 'How *few* are the sellers' is not easy to define numerically in the oligopolistic market structure. The economists are not specified about a definite number of sellers for the market to be oligopolistic in its form. It may be two, three, four, five or more. In fact, the number of sellers depends on the size of the market. Given the size of the market, if number of sellers is such that each seller has command over a sizeable proportion of the total market supply, then there exists oligopoly in the market.

The products traded by the oligopolists may be differentiated or homogeneous. Accordingly, the market may be characterized by heterogeneous oligopoly or homogeneous (or pure) oligopoly. In automobile industry, Maruti Zen, Hyudai's Santro, Daewoo's Matis, Fiat's Palio and Tata's Indica, etc., are the outstanding examples of differentiated oligopoly. Similarly, cooking gas of Indane and of Burshane are the examples of homogeneous oligopoly. Differentiated oligopolies include automobiles, cigarettes, refrigerators, and TV industries. Pure oligopoly includes such industries as cooking gas, cement, baby food, cable wires, and dry batteries. Other examples of oligopolistic industries are aluminium, paints, tractors, steel, tyres, and tubes.

Characteristics of oligopoly

The basic characteristics of oligopolistic market structure are as follows

1. **Intensive competition:** The characteristic fewness of their number brings oligopolist in intensive competition with one another. Let us compare oligopoly with other market structures. Under perfect competition, competition is non-existent because the number of sellers is so large that no seller is strong enough to make any impact on market conditions. Under monopoly, there is a single seller, and therefore, there is absolutely no competition. Under monopolistic competition, number of sellers is so large that degree of competition is considerably reduced. Whereas under oligopoly, the number of sellers is so small that any move by one seller immediately affects the rival sellers. As a result, each firm keeps a close watch on the activities of the rival firms and prepares itself with a number of aggressive and defensive marketing strategies. To an oligopolist, business is a 'life' of constant struggle as market conditions necessitate making moves and counter-moves. This kind of competition is not found in other kinds of market. Oligopoly is the highest form of competition.
2. **Interdependence of business decisions:** The nature and degree of competition among the oligopolists makes them interdependent in respect of decision making. The reason for inter-dependence between the oligopolists

NOTES

is that a major policy change made by one of the firms affects the rival firms seriously and immediately, and forces them to make counter-moves to protect their interest. Therefore, each oligopolist, while making a change in his price, advertisement, product characteristics, etc. takes it for granted that his actions will cause reaction by the rival firms. Thus, interdependence is the source of action and reaction, moves and counter-moves by the competing firms. An illuminating example of strategic manoeuvering by the oligopoly firm has been given by Robert A. Meyer. To quote the example, one of the US automobile companies announces in September an increase of $180 in the list price of its new car model. Following it, a few days later, a second company announces an increase of only $80 and a third announces increase of $91. The first company makes a counter-move: it suddenly reduces the increase in list price to $71 from $180 announced earlier. One can now expect that other firms will follow the first in price cutting. Obviously, there is a good deal of uncertainty in the behaviour of firms.

3. **Barrier to entry:** An oligopolistic market structure is also characterized, in the long run, by strong barriers to entry of new firms to the industry. If entry is free, new firms attracted by the super-normal profits, if it exists, enter the industry and the market eventually becomes competitive. Usually, barriers to entry do exist in an oligopolistic market. Some common barriers to entry are economies of scale, absolute cost advantage to old firms, price-cutting, control over important inputs, patent rights and licencing, preventive price and existence of excess capacity. Such factors prevent the entry of new firms and preserve the oligopoly.

Oligopoly Models: An Overview

The uncertainty in respect of behaviour pattern of a oligopoly firms arising out of their unpredictable action and reaction makes systematic analysis of oligopoly extremely difficult. Under these circumstances, a wide variety of behaviour pattern has been observed: they may come in collusion with each other or may try to fight each other to the death. The agreement may last or may break down soon. Indeterminateness of price and output therefore becomes the basic feature of oligopolistic markets. In accordance with the variety of behaviours, economists have developed a variety of analytical models based on different behavioural assumptions. Among notable models are Cournot's Duopoly model (1838), Bertrand's model (1883), Edgeworth's model (1897), Stackelberg's leadership model (1930), Hotelling's model (1930s), Chamberlin's model (1933), Sweezy's kinked-demand curve model (1939), Neumann and Morgenstern's game theory model (1944), and Baumol's sales maximization model. None of these models, however, provides a universally acceptable analysis of oligopoly, though these models do provide insight into the behavioural pattern of oligopolists. Moreover, these models are studied for their paedagogic importance.

The analytical models of oligopoly, suggested by the economists, may be classified under the following two road categories:

(i) duopoly models, and

(ii) general oligopology models.

NOTES

The oligopoly models may be further sub-classified as (a) non-collusive models, and (b) collusive models.

3.6.1 Duopoly Models

When there are only two sellers of a product, there exists duopoly, a special case of oligopoly. Duopoly is a special case in the sense that it is the limiting case of oligopoly as there must be at least two sellers to make the market oligopolistic in nature. Let us study famous classical models of duopoly.

(i) Cournot's duopoly model

Augustin Cournot, a French economist, was the first to develop a formal duopoly model in 1838. To illustrate his model, Cournot assumed:

(a) two firms, *A* and *B*, each owning an artesian mineral water wells;

(b) both operate their wells at zero marginal cost;

(c) both face a downward sloping straight line demand curve; and

(d) each seller acts on the assumption that his competitor will not react to his decision to change his output and price. This is Cournot's behavioural assumption.

On the basis of this model, Cournot has concluded that each seller ultimately supplies one-third of the market and both the sellers charge the same price. Moreover, one-third of the market remains unsupplied.

Cournot's duopoly model is presented in Figure 3.33. To begin the analysis, suppose that *A* is the only seller of mineral water in the market. In order to maximize his profits or revenue, he sells quantity *OQ* at which his $MC = O = MR$, at price OP_1. His total profit is OP_1PQ.

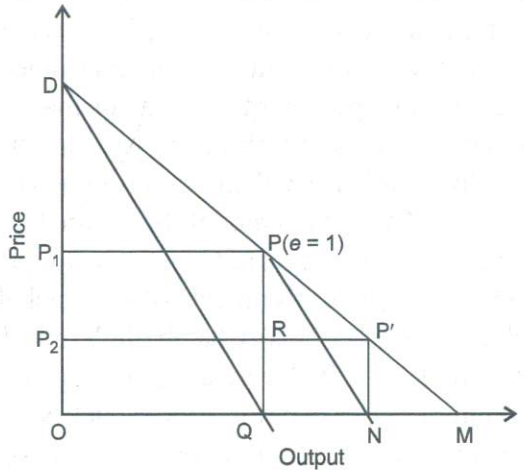


Fig. 3.33 Price and Output Determination under Doupoly: Cournot's Model

Now let *B* enter the market. The part of market open to him equals *QM*, which is half of the total market. Note that *QM* is the part of the market left unsupplied by *A*. It means that *B* can sell his product in the remaining half of the market, *PM* being the relevant part of demand curve for him. *B* assumes that *A*

will not change his price and output because he is making the maximum profit. That is, *B* assumes that *A* will continue to sell *OQ* at price OP_1. Thus, the market available to him is *QM* and the relevant demand curve is *PM*. When he draws his *MR* curve, *PN*, it bisects *QM* at point *N* where *QN* = *NM*. In order to maximize his revenue, *B* sells *QN* at price $OP_2 = P'N$. His total revenue is maximum at *QRP'N*. Note that *B* supplies only $QN = 1/4 = 1/2 \times 1/2$ of the market.

With the entry of *B*, price falls to OP_2. Therefore, *A*'s expected profit falls to OP_2RQ. Faced with this situation, *A* attempts to adjust his price and output to the changed conditions. He assumes that *B* will not change his output *QN* and price OP_2 as he *(B)* is making the maximum profit. Accordingly, *A* assumes that *B* will continue to supply 1/4 of the market and, therefore, he has 3/4 = 1 – 1/4 of the market available to him. To maximize his profit, *A* will supply 1/2 (3/4) = 3/8 of the market. Note that *A*'s market share has fallen from 1/2 to 3/8.

Now it is *B*'s turn to react. Following Cournot's assumption, *B* assumes that *A* will continue to supply only 3/8 of the market and the market open to him equals 1 – (3/8) = 5/8. To maximize his profit under the new conditions, *B* will supply 1/2 (5/8) = 5/16 of the market. It is now for *A* to reappraise the situation and adjust his price and output accordingly.

This process of action and reaction continues in successive periods. In the process, *A* continues to loose his market share and *B* continues to gain. Eventually, a situation is reached when their market share equals at 1/3 each. Any further attempt to adjust output produces the same result. The firms, therefore, reach their equilibrium position with each supplying 1/3 of the market and 1/3 of the market remaining unsupplied.

The process through which firms reach their equilibrium, according to Cournot's model, may be illustrated as presented in the following table.

Table 3.2 *Market Sharing in Cournot's Model*

Period	Firm A	Firm B
I	$\frac{1}{2}(1)=\frac{1}{2}$	$\frac{1}{2}\left(\frac{1}{2}\right)=\frac{1}{4}$
II	$\frac{1}{2}\left(1-\frac{1}{4}\right)=\frac{3}{8}$	$\frac{1}{2}\left(1-\frac{3}{8}\right)=\frac{5}{16}$
III	$\frac{1}{2}\left(1-\frac{5}{16}\right)=\frac{11}{32}$	$\frac{1}{2}\left(1-\frac{11}{32}\right)=\frac{21}{64}$
IV	$\frac{1}{2}\left(1-\frac{21}{64}\right)=\frac{43}{128}$	$\frac{1}{2}\left(1-\frac{43}{128}\right)=\frac{85}{256}$
...	.....	.....
...	.....	.....
...	.....	.....
N	$\frac{1}{2}\left(1-\frac{1}{3}\right)=\frac{1}{3}$	$\frac{1}{2}\left(1-\frac{1}{3}\right)=\frac{1}{3}$

NOTES

Cournot's equilibrium solution is stable. For, given the action and reaction, it is not possible for any of the two sellers to increase their market share. Cournot's model of duopoly can be extended to the case of general oligopoly. For example, suppose there are three sellers, the industry and firms will be in equilibrium when each firm supplies 1/4 of the market. The three sellers together supply 3/4 = 3 (1/4) of the market, 1/4 of the market remaining unsupplied. The formula for determining the share of each seller in an oligopolistic market is $Q \div (n + 1)$, where Q = market size and n = number of sellers.

Algebra of Cournot's model

Cournot's duopoly model may also be presented algebraically. Let us suppose that market demand function is given by a linear function given as

$$Q = 90 - P \quad \ldots (3.11)$$

We have noted above that, under zero cost condition, profit is maximum where $MC = MR = 0$ and profit-maximizing output equals $Q/2$.

Thus, when firm A is a monopolist in the market, his profit-maximizing output (O_A), according to the profit-maximizing rule under zero cost condition, is given by

$$Q_A = 1/2\ (90 - P) \quad \ldots (3.12)$$

When another firm, B, enters the market, its profit-maximizing output equals

$$Q_B = 1/2[1/2(90 - P)] \quad \ldots (3.13)$$

Thus, the respective share of firms, A and B is fixed at Q_A and Q_B. The division of market output may be expressed as

$$Q = Q_A + Q_B = 90 - P \quad \ldots (3.14)$$

The demand function for the firm A may now be expressed as

$$Q_A = (90 - Q_B) - P \quad \ldots (3.15)$$

and for the firm B as

$$Q_B = (90 - Q_A) - P \quad \ldots (3.16)$$

Given the demand function (3.15), the market open to firm A (at $P = 0$) is $90 - Q_B$. The profit-maximizing output for A can be written as

$$Q_A = \frac{90 - Q_B}{2} \quad \ldots (3.17)$$

and for B, as

$$Q_B = \frac{90 - Q_A}{2} \quad \ldots (3.18)$$

The Equations (3.17) and (3.18) represent the reaction functions of firms A and B, respectively. For example, consider Equation (3.17). The profit-maximizing output of firm A depends on the value of Q_B, i.e., the output which firm B is assumed to produce. If firm B chooses to product 30 units, (i.e., $Q_B = 30$), then A's output = 30 [= (90 – 30)1/2]. If firm B chooses to produce 60 units, A's output = [15 (= 90 – 60) 1/2]. Thus, Equation (3.17) is the reaction function of firm A. It can be similarly shown that Equation (3.18) is the reaction function of firm B.

NOTES

Criticism

Although Cournot's model yields a stable equilibrium, it has been criticized on the following grounds:

First, Cournot's behavioural assumption is naive to the extent that it implies that firms continue to make wrong calculations about the competitor's behaviour. That is, each seller continues to assume that his rival will not change his output even though he repeatedly observes that his rival firm does change its output.

Second, his assumption of zero cost of production is unrealistic, though dropping this assumption does not alter his model.

(ii) Bertrand's duopoly model

Betrand, a French mathematician, criticized Cournot's model and developed his own model of duopoly in 1883. Bertrand's model differs from Cournot's model in respect of its behavioural assumption. While under Cournot's model, each seller assumes his rival's output to remain constant, under Bertrand's model, each seller determines his price on the assumption that his rival's price, rather than his output, remains constant.

Bertrand's model concentrates on price competition. His analytical tools are reaction functions of the duopolists. Reaction functions of the duopolists are derived on the basis of iso-profit curves. An iso-profit curve, for a given level of profit, is drawn on the basis of various combinations of prices charged by rival firms. Assuming two firms *A* and *B,* the two axis of the plane on which iso-profit curves are drawn measure the prices of the two firms, one each. Iso-profit curves of the two firms are convex to their respective price axis, as shown in Figures 3.34 and 3.35. Iso-profit curves of firm *A* are convex to its price axis P_A [(Figure 3.34)] and those of firm *B* are convex to P_B [(Figure 3.35).

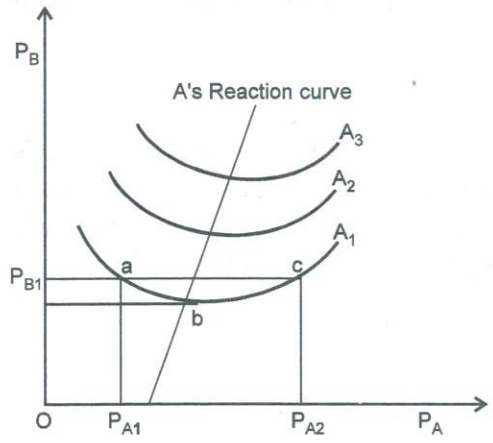


Fig. 3.34 A's Reaction Curve

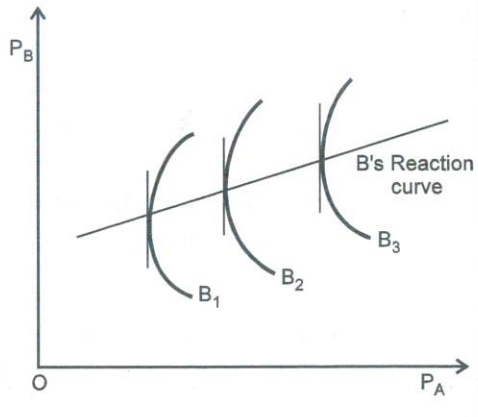


Fig. 3.35 B's Reaction Curve

To explain the implication of an iso-profit curve, consider curve *A* in Figure 3.34. It shows that *A* can earn a given profit from the various combinations of its own and its rival's price. For example, price combinations at points *a, b* and *c* on iso-profit curve A_1 yield the same level of profit. If firm *B* fixes its price P_{B1}, firm *A* has two alternative prices, P_{A1} and P_{A2}, to make the same level of profits. When *B* reduces its price, *A,* may either raise its price or reduce it. *A* will reduce

its price when he is at point *c* and raise its price when he is at point *a*. But there is a limit to which this price adjustment is possible. This point is given by point *b*. So there is a unique price for *A* to maximize its profits. This unique price lies at the lowest point of the iso-profit curve. The same analysis applies to all other iso-profit curves. On joining the lowest points of the iso-profit curves A_1, A_2 and A_3, *A*'s reaction curve if formed. Note that *A*'s reaction curve has a rightward slant. This is so because iso-profit curve tend to shift rightward when *A* gains market from its rival *B*.

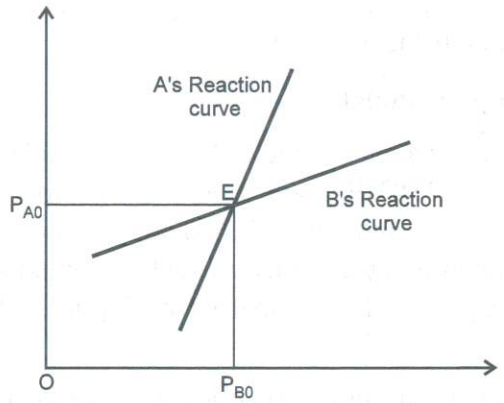


Fig. 3.36 *Duopoly Equilibrium: Bertand's Model*

Following the same process, *B*'s reaction curve may be drawn, as shown in Figure 3.35. The equilibrium of duopolists suggested by Bertrand's model may be obtained by putting together the reaction curves of the firms *A* and *B*, as shown in Figure 3.36. The reaction curves of *A* and *B* intersect at point *E* where their expectations materialize. Point *E* is, therefore, equilibrium point. This equilibrium is stable. For, if any one of the firms deviates from the equilibrium point, it will generate a series of actions and reactions between the firms, which will lead them back to point *E*.

Criticism

Bertrand's model has, however, been criticized on the same grounds as Cournot's model. Bertrand's implicit behavioural assumption that firms never learn from their past experience is naive. Furthermore, if cost is assumed to be zero, price will fluctuate between zero and the upper limit of the price, instead of stabilizing at a point.

(iii) Edgeworth's duopoly model

Edgeworth developed his model of duopoly in 1897. Edgeworth's model follows Bertrand's assumption that each seller assumes his rival's price, instead of his output, to remain constant. His model is illustrated in Figure 3.37.

Let us suppose that there are two sellers, *A* and *B,* in the market. The entire market *M'M* in Figure 3.37 is equally divided between the two sellers who face identical demand curves. *A* has his demand curve as D_A and *B* as D_B. Let us also assume that seller *A* has a maximum capacity of output *OM* and *B* has a maximum output capacity of *OM'*. The ordinate *OD* measures the price.

To begin the analysis of Edgeworth's model, let us suppose that *A* is the only seller in the market. Following the profit-maximizing rule of a monopoly seller, he sells *OQ* and charges a price, OP_2. His monopoly profit, under zero cost, equals OP_2EQ. Now, *B* enters the market and assumes that *A* will not change his price since he is making maximum profit. With this assumption, *B* sets his price slightly below *A*'s price (OP_2) and is able to sell his total output and also to capture a substantial position of *A*'s market.

Seller *A* now realizes the reduction in his sale. In order to regain his market, *A* sets his price slightly below *B*'s price. This leads to price-war between the sellers. The price-war takes the form of price cutting, which continues until price reaches OP_1. At this price, both *A* and *B* are able to sell their entire output *A* sells *OM* and *B* sells *OM'*. The price OP_1 could, therefore, be expected to be stable. But, according to Edgeworth, price OP_1 should not be stable.

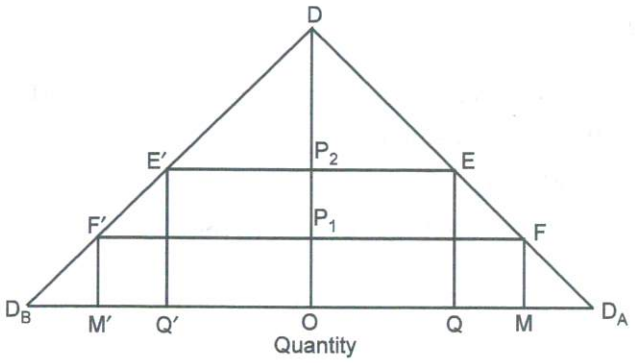


Fig. 3.37 *Edgeworth's Model of Duopoly*

The reason is that, once price OP_1 is set in the market, the sellers observe an interesting fact. That is, each seller realizes that his rival is selling his entire output and he will, therefore, not change his price, and each seller thinks that he can raise his price to OP_2 and can make pure profit. This realization forms the basis of their action and reaction. For example, let seller *A* take the initiative and raise his price to OP_2. Assuming *A* to retain his price OP_2, *B* finds that if he raises his price to a level slightly below OP_2, he can sell his entire output at a higher price and make greater profits. Therefore, *B* raises his price according to his plan.

Now it is *A*'s turn to appraise the situation and react. *A* finds that his price is higher than *B*'s price. His total sale falls. Therefore, assuming *B* to retain his price, *A* reduces his price slightly below *B*'s price. Thus, the price war between *A* and *B* begins once again. This process continues indefinitely and price keeps moving up and down between OP_1 and OP_2. Obviously, according to Edgeworth's model of duopoly, equilibrium is unstable and indeterminate since price and output are never determined. In the words of Edgeworth, there will be an indeterminate tract through which the index of value will oscillate, or, rather will vibrate irregularly for an indefinite length of time.

Edgeworth's model, like Cournot's and Bertrand's model, is based on a naïve assumption, i.e., each seller continues to assume that his rival will never

NOTES

change his price or output even though they are proved repeatedly wrong. But, Hotelling remarked that Edgeworth's model is definitely an improvement upon Cournot's model in that it assumes price, rather than output, to be the relevant decision variable for the sellers.

(iv) Stackelberg's leadership model

Stackelberg, a German economist, developed, his leadership model of duopoly in 1930. His model is an extension of Cournot's model. Stackelberg assumes that one of the duopolists (say *A)* is sophisticated enough to play the role of a leader and the other (say *B)* acts as a follower. The leading duopolist *A* recognizes that his rival firm *B* has a definite reaction function, which *A* uses into his own profit function and maximizes his profits.

Suppose market demand function is given as in Equation (3.11), i.e., $Q = 90 - P$ and *B*'s reaction function is given as in Equation (3.19), i.e.,

$$Q_B = \frac{90 - Q_A}{2} \quad \ldots (3.19)$$

Now, let *A* incorporate *B*'s reaction function into the market function and formulate his own demand function as

$$Q_A = 90 - Q_B - P \quad \ldots (3.20)$$

Since $Q_B = (90 - Q_A)/2$, Equation (3.20) may be written as

$$Q_A = 90 - \frac{90 - Q_A}{2} - P$$

or $$Q_A = 45 + \frac{Q_A}{2} - P$$

or $$2Q_A = 90 + Q_A - 2P \quad \ldots (3.21)$$

$$Q_A = 90 - 2P$$

Thus, by knowing *B*'s reaction function, *A* is able to determine his own demand function. Following the profit-maximization rule, *A* will fix his output at 45 units (= 90/2), i.e., half of the total demand at zero price.

Now, if seller *A* produces 45 units and seller *B* sticks to his own reaction function, he will produce

$$Q_B = \frac{90 - 45}{2} = 22.5 \text{ units} \quad \ldots (3.22)$$

Thus, the industry output will be

$$45 + 22.5 = 67.5.$$

The problem with Stackelberg's model is that it does not decide as to which of the firms will act as leader (or follower). If each firm assumes itself to be the leader and the other to be the follower then Stackelberg's model will be indeterminate with unstable equilibrium.

3.6.2 Oligopoly Models

There are two kinds of oligopoly models: (i) non-collusive models, and (ii) collusive models. The non-collusive models of oligopoly explain the price and

NOTES

output determination in a market structure in which oligopolists recognize their interdependence. Chamberlin's non-collusive model of oligopoly, i.e., 'small group' model, is considered a major contribution to the theory of oligopoly. Another famous model of this category is Sweezy's kinked demand curve model.

I. Non-collusive models of oligopoly

(i) Chamberlin's model of oligopoly: The 'small group' model

The *classical* models of duopoly assumed independent action by the rival firms in their attempt to maximize their profits. Chamberlin rejected the assumption of independent action by the competing firms. He developed his own model of oligopoly assuming *interdependence* between the competitors. He argued that firms do not act independently. They do recognize their mutual interdependence. Firms are not as 'stupid' as assumed in the models of Cournot, Edgeworth and Bertrand. In his own words, 'When a move by one seller evidently forces the other to make a counter-move, he is very stupidly refusing to look further than his nose if he proceeds on the assumption that it will not.' Chamberlin suggests that each seller seeking to maximize his profit reflects well and looks into the consequences of his move. The total consequence of a seller's move consists of both its direct and indirect effects. The *direct effects* are those which results from a seller's own action, rival sellers not reacting to his action. The *indirect effects* are those which result from the reaction of the rival sellers to the moves made by a seller.

Chamberlin suggests in his model that if rival firms are assumed to recognize their interdependence and act accordingly, a stable equilibrium can be reached, where each firm charges monopoly price. When all firms are in equilibrium, industry profit is maximized. Chamberlin's oligopoly model of 'small group' can be best understood if presented in the framework of Cournot's duopoly model since Chamberlin follows Cournot to develop his own model.

Cournot's model is reproduced in Figure 3.38, except the ordinate *JK*. Assuming that there are two firms, *A* and *B,* let *A* first enter the market as a monopolist. Following the profit maximization rule, firm *A* will produce *OQ* and charge monopoly price OP_2 *(= PQ)*. When firm *B* enters the market, it considers that *PM* is its demand curve. Under Cournot's assumption, firm *B* will sell output *QN* at price OP_1. As a result, market price falls from OP_2 to OP_1. It is now *A*'s turn to appraise the situation. At this point, Chamberlin deviates from Cournot's model. According to Cournot's model, firm *A* does not recognize its interdependence and acts independently. Chamberlin, however, assumes that firm *A* recognizes the interdependence between them and it recognizes the fact that *B* will react to its decisions. Therefore, firm *A* decides to compromise with the existence of firm *B,* and decides to reduce its output to *OK* which is half of the monopoly output, *OQ*. Its output *OK* equals *B*'s output *QN* (= *KQ*). In its turn, firm *B* also recognizes their interdependence. It realizes that *KQ* is the most profitable output for it. Thus, the industry output is *OQ* which is the same as monopoly output, and market price is OP_2 (= *PQ*) which equals monopoly price. Thus, according to Chamberlin, by recognizing their interdependence, the firms

NOTES

reach an equilibrium which is the same as monopoly equilibrium, and share the market equally. One of the firms supplies *OK* and the other supplies *KQ* where $OK + KQ = OQ$, the profit maximizing monopoly output. This equilibrium is stable because under the condition of interdependence, firms do not gain by changing their price and output.

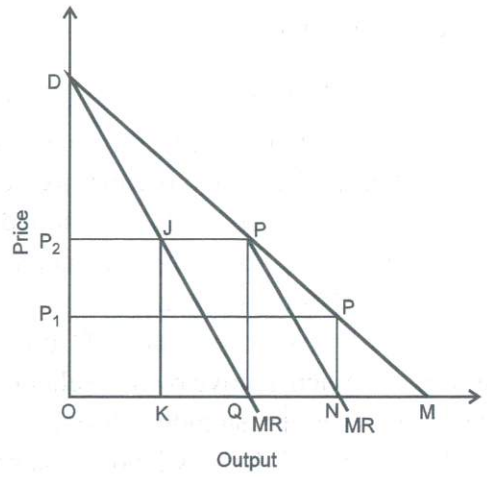


Fig. 3.38 *Chamberlin's Model of Stable Oligopoly Equilibrium*

Chamberlin's model is regarded as an improvement over the earlier models, at least in respect of its behavioural assumption of interdependence. His model has, however, been criticized on the grounds that his idea of joint profit maximization is beset with problems of estimating demand and cost functions. Unless demand and cost functions are fully known to the competitors, joint profit maximization is doubtful.

(ii) Sweezy's kinked-demand curve model of oligopoly

The origin of kinked-demand curve can be traced in Chamberlin's theory of monopolistic competition. Later, Hall and Hitch used the *kinked-demand curve* to explain rigidity of prices in oligopolistic market. However, neither Chamberlin nor Hall and Hitch used the kinked-demand curve as a tool of analysis in their respective theories. It was Paul M Sweezy who used the kinked-demand curve in his model of price stability in oligopolistic market.

The kinked-demand curve model developed by Paul M Sweezy has features common to most oligopoly pricing models. This is the best known model to explain, relatively more satisfactorily, the behaviour of the oligopolistic firms. The *kinked-demand curve* analysis does not deal with price and output determination. Instead, it seeks to establish that once a price–quantity combination is determined, an oligopoly firm will not find it profitable to change its price even in response to the small changes in the cost of production. The logic behind this proposition is as follows. An oligopoly firm believes that if it reduces the price of its product, the rival firms would follow and neutralize the expected gain from price reduction. But if it raises the price, the firms would either maintain their prices or even indulge in price cutting, so that the price-raising firms stand to lose, at least, a part of its market share. This behaviour is true for all the firms. The oligopoly firms would, therefore, find it more desirable to maintain the prevailing price and output.

NOTES

To look more closely at the kinked-demand curve analysis, let us look into the possible actions and reactions of the rival firms to the price changes made by one of the firms.

There are three possible ways in which rival firms may react to change in price by one of the firms: (i) the rival firms follow the price changes, both cut and hike; (ii) the rival firms do not follow the price changes; and (iii) rival firms do not react to price hikes but they do follow the price cutting. If rival firms react in manners (i), an oligopoly firm taking lead in changing prices will face demand curve *dd'* in Figure 3.39. If rival firms react in manner (ii), the firm faces demand curve *DD'*. The demand curve *dd'* which is based on reaction (i), is less elastic than the demand curve *DD'* which is based on reaction (ii). Demand curve *dd'* is less elastic because changes in demand in response to changes in price are restrained by the counter-moves by the rival firms.

Given the two demand curves, let point *P* represent the equilibrium price–quantity combination of an oligopolist. Let us now introduce reaction (iii), i.e., rival firms follow the oligopolist leading in price cutting when he reduces his price but do not follow him when he increases his price. This asymmetrical behaviour of the rival firms makes only a part of each of the two demand curves relevant for the oligopolist. This can be established by allowing an oligopolist to alternatively increase and decrease his price. If an oligopolist increases his price and his rivals do not follow him, he loses a part of his market to his rivals. The demand for his product decreases considerably, indicating a greater elasticity. The oligopolist is, therefore, forced down from demand curve *dP* to *DP*. Thus, the relevant segment of demand curve for the oligopolist is *DP*.

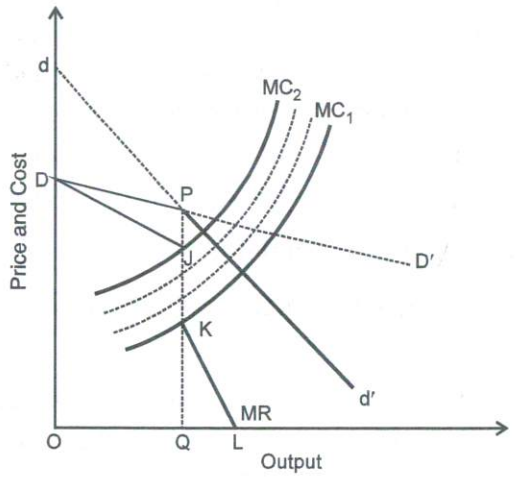


Fig. 3.39 *Kinked-demand Analysis*

On the other hand, if an oligopolist decreases his price, the rival firms react by cutting down their prices by an equal amount or even more. This counter-move by the competitors prevents the oligopolist from taking full advantage of price cut along the demand curve *DD'*. Therefore, his demand curve below point *P* rotates down. Thus, the relevant segment of demand curve for the oligopolist (below point *P)* is *Pd'*. If the two relevant segments of the two demand curves are put together, the relevant demand curve for the oligopolist is *DPd'* which has a *kink* at point *P*. Therefore, it is called a 'kinked-demand curve'.

NOTES

Now, consider now the relationship between *AR* (=*D*) and *MR*. We know that *MR* = *AR*–*AR/e*. The *MR* curve, drawn on the basis of this relationship, will take a shape as shown by *DJKL* in Figure 3.39. It is discontinuous between point *J* and *K*, at output *OQ*. Suppose that the original marginal cost curve resembles MC_1 which intersects *MR* at point *K*. Since at output *OQ*, the necessary condition of maximum profit (*MR* = *MC*) is satisfied, the oligopolist is earning maximum profit. Now, if marginal cost curve shifts upwards to MC_2 or to any level between points *J* and *K*, his profit would not be affected. Therefore, he has no motivation for increasing or decreasing his price. It is always beneficial to stick to the price and output. Thus, both price and output are stable.

The oligopolists will think of changing their price and output only if *MC* rises beyond point *J* or decreases below point *K* (Figure 3.39). But, even if it so happens, price and output would tend to stabilize. Suppose the general level of costs rises for the industry so that *MC* moves above point *J*. The oligopolists will ultimately find it profitable to raise the price. When one of the oligopolists raises his price, his competitors match the price increase. As a result, the kinked-demand curve shifts upward to a new position and the point of kink shifts rightward and horizontally. Again, at the new price there is no incentive for any oligopolist to raise his price. Therefore, price tends to stabilise.

Alternatively, if *MC* moves down below point *K*, firms get incentive to reduce their price. When one firm cuts its price, others follow with matching price-reduction. There is a possibility of competitors reducing their prices by a greater margin. The only way to prevent this situation is that the oligopolist must keep his costs as low as possible, at least lower than that of his competitors. This is the reason why there is keen technological competition in an oligopolistic market. In other words, there is incentive for oligopoly finds to use new and efficient technique of production, introduce new products, make innovations, increase their productivity or reduce their cost of production to the possible minimum. They find it safe to concentrate on efficiency rather than to indulge in price-war.

Some implications of Sweezy's model

According to Sweezy, his model for price stability in an oligopolistic market has the following implications:

First, since elasticity of the demand curve below point *P* is assumed to be less than unity and *MR* beyond a point is negative, the conditions of short-run equilibrium are not precise. That is, profit maximization rule, *MC*= *MR*, cannot be applied to the short-run conditions.

Second, since *MC* can shift up and down between the finite points *J* and *K* (Figure 3.39), *MR* remaining the same, his model deviates from the marginal productivity theory, i.e., factor prices do not equal their marginal revenue productivity.

Third, any short-term disturbance in *MC* will not affect the equilibrium price or output and the total profits. Thus, the general belief that a successful strike by the trade unions reduces profits gets little theoretical support from Sweezy's model.

NOTES

Criticism

The major criticism against this model is that it explains only the stabilization of output and price. It does not tell why and how the initial price is fixed at a certain level. The Sweezy's thesis must, therefore, be regarded as an *ex-post* rationalization rather than as an *ex-ante* explanation of market equilibrium.

Besides, Sweezy's claim of price stability does not stand the test of empirical verification: there is a surprising lack of price rigidity. Monopoly prices have been found more stable than oligopoly prices. However, economists are divided on the issue of price rigidity. While Stigler doubts the existence of kinked-demand curve and price rigidity, Liebhafsky finds considerable evidence of price rigidity in the US. Cohen and Cyert argue that kink in the demand curve and price rigidity may exist for a brief period for lack of inter-firm information, particularly when new and unknown rivals enter the market. They are of the opinion that kink is clearly not a stable long-run equilibrium.

II. Collusive models of oligopoly

In the non-collusive models, oligopoly firms are assumed to act independently. In the collusive models, however, firms are assumed to act in unison, i.e., in collusion with one another. This assumption is based on empirical facts, rather than being conjectural.

Why Collusion?

There are at least three major factors that bring collusion between the oligopolistic firms. **First,** collusion reduces the degree of competition between the firms and helps them act monopolistically in their effort of profit maximization. **Second,** collusion reduces the oligopolistic uncertainty surrounding the market since cartel members are not supposed to act independently and in the manner that is detrimental to the interest of other firms. **Third,** collusion forms a kind of barrier to the entry of new firms.

Collusion between oligopoly firms may take many forms, depending on their relative strength, their objective and legal status of collusion. There are, however, two main types of collusion: (i) cartels and (ii) price leadership.

(i) Cartels under oligopoly

A cartel is a formal organization of the oligopoly firms in an industry. Cartels are the perfect form of collusion. A general purpose of cartels is to centralize certain managerial decisions and functions of individual firm in the industry with a view to promoting common benefits. Cartels may be in the form of *open collusion* or *secret collusion.* Whether open or secret, cartel agreements are *explicit* and *formal* in the sense that agreements are enforceable on member firms trying to pursue an independent pricing policy. Cartels are, therefore, regarded as the perfect form of collusion. Cartels and cartel-type agreements between the firms in manufacturing and trade are illegal in most countries. Yet, cartels in the broader sense of the term exist in the form of trade associations, professional organizations and the like.

NOTES

A cartel performs a variety of services for its members. The two typical services of central importance are (i) fixing price for joint maximization of industry profits; and (ii) market sharing between it members.

Cartels and profit maximization

Let us suppose that a group of firms producing a homogeneous commodity forms a cartel aiming at joint profit maximization. The firms appoint a central management board with powers to decide the following aspects:

- Total quantity to be produced;
- Price at which the product has to be sold
- Share or each firm in the total output

The central management board is provided with cost figures of individual firms. Besides, it is supposed to obtain the necessary data required to formulate the market demand (*AR*) curve. The management board calculates the marginal cost (*MC*) and marginal revenue (*MR*) for the industry. Furthermore, the management board holds the position of a multiplant monopoly. It determines the price and output for each firm in the manner a multiplant monopoly determines the price and output for each plant.

The model of price and output determination for each is presented in Figure 3.40. It is assumed for the sake of convenience that there are only two firms, *A* and *B*, in the cartel. Their respective cost curves are given in the first two panels of Figure 3.40. In the third panel, the *AR* and *MR* curves represent the revenue conditions of the *industry*. The *MC* curve is the summation of *MC* curves of the individual firms. The *MC* and *MR* intersect at point *C* determining the industry output at *OQ*. The market price is determined at *PQ*. The industry output *OQ* is so allocated between firms *A* and *B* that for each of them *MC* = *MR*. The share of each firm in the industry output, *OQ*, can be determined by drawing a line from point *C* and parallel to *X*-axis through MC_2 and MC_1. The points of intersection C_1 and C_2 determine the level of output for firms *A* and *B*, respectively. Thus, the share of each of the two firms *A* and *B*, is determined at Oq_1 and Oq_2, respectively, where $Oq_1 + Oq_2 = OQ$. Their respective profit can be computed as (P_m – firm's *ac*) × firm's output, which is maximum. The total profit of each firm may be different. But there is no motivation for changing price quantity combination, since their individual profit is maximum.

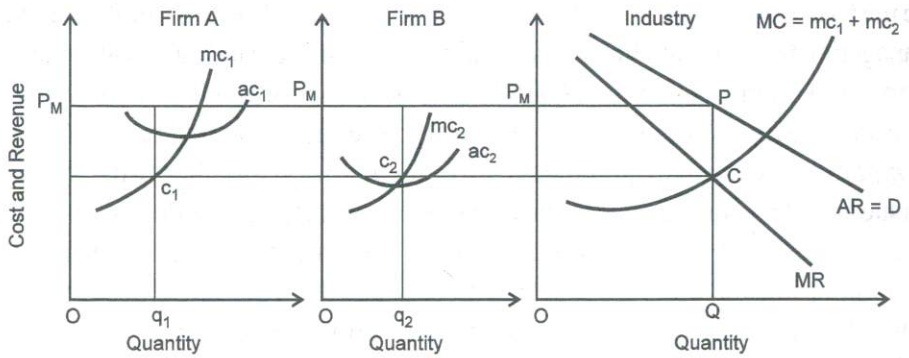


Fig. 3.40 Price and Output Determination under Cartel

NOTES

Critical appraisal

Although monopoly solution to joint profit maximization by cartels look theoretically sound, William Fellner gives the following reasons why joint profits may not be maximized.

First, it is difficult to estimate market demand curve accurately since each firm thinks that the demand for its own product is more elastic than the market demand curve because its product is a perfect substitute for the product of other firms.

Second, similarly an accurate estimation of industry's *MC* curve is highly improbable for lack of adequate and correct cost data. If industry's *MC* is incorrectly estimated, industry output can be only incorrectly determined. Hence, joint profit maximization is doubtful.

Third, cartel negotiations take a long time. During the period of negotiation, the composition of the industry and its cost structure may change. This may render the estimates irrelevant, even if they are correct. Besides, if the number of firms increase beyond 20 or so, cartel formation becomes difficult, or even if it is formed, it soon breaks down.

Fourth, there are 'chiselers' who have a strong temptation to give secret concessions to their customers. This tendency in the members reduces the prospect of joint profit maximization.

Fifth, if cartel price, like monopoly price, is very high, it may invite government attention and interference. For the fear of government interference, members may not charge the cartel price.

Sixth, another reason for not charging the cartel price is the fear of entry of new firms. The high cartel price which yields monopoly profit may attract new firms to the industry. To prevent the entry of new firms, some firms may decide on their own not to charge the cartel price.

Finally, another reason for not charging the cartel price is the desire to build a public image or good reputation. Some firms may, to this end, decide to charge only a fair price and realize only a fair profit.

Cartel and market sharing

The market-sharing cartels are more common because this kind of collusion permits a considerable degree of freedom in respect of style of the product, advertising and other selling activities. There are two main methods of market allocations: (i) non-price competition, and (ii) quota system.

(a) **Non-price competition:** The non-price competition agreements are usually associated with loose cartels. Under this kind of arrangement between the firms, a uniform price is fixed and each firm is allowed to sell as much as it can at the cartel price. The only requirement is that firms are not allowed to reduce the price below the cartel price.

The cartel price is a bargain price. While low-cost firms press for a low price, the high-cost firms press for a higher price. But the cartel price is so fixed by mutual consent that all member firms are able to make some profits.

NOTES

But the firms are allowed to compete with one another in the market on a non-price basis. That is, they are allowed to change the style of their product, innovate new designs, and to promote their sales by advertising.

Whether this arrangement works or breaks down depends on the cost conditions of the individual firms. If some firms expect to increase their profits by violating the price agreements, they will indulge in cheating by charging lower price. This may lead to a price-war and cartel may break down.

(b) **Quota system:** The second method of market sharing is quota system. Under this system, the cartel fixes a quota of market share for each firm. There is no uniform principle for fixing quota. In practice, however, the main considerations are (i) bargaining ability of a firm and its relative importance in the industry, (ii) the relative sales of the firms in pre-cartel period, and (iii) production capacity of the firm. The choice of base period depends on the bargaining ability of the firm.

Another popular basis of market-sharing is the geographical division of market. Examples of this kind of market-sharing are mostly found in the case of international markets.

Unequal quota for unequal firms

Fixation of quota is a difficult proposition. Nevertheless, some theoretical guidelines for market-sharing have been suggested by the economists: (i) unequal quota for unequal firms, i.e., firms with different cost curves, and (ii) equal quota for equal firms—firms with identical cost and revenue curves.

A reasonable criterion of ideal market-sharing can be to share the total market between the cartel members in such proportions that the industry's marginal cost equals the marginal cost of individual firms. This criterion is illustrated in Figure 3.41. The profit maximizing output of the industry is OQ. The industry output OQ is shared between the two firms A and B, as Oq_1 and Oq_2 respectively. Note that $OQ = Oq_1 + Oq_2$. At output Oq_1, mc of firm A equals industry's marginal cost, MC, and at output Oq_2, mc of firm B equals MC. Thus, under quota system, the quota for firms A and B may be fixed as Oq_1 and Oq_2, respectively. Given the quota allocation, the firm may set different prices for their product depending on the position and elasticity of their individual demand curves. This criterion is identical to the one adopted by a multiplant monopolist in the short-run, to allocate the total output between the plants.

Equal quota for equal firms

Another reasonable criterion for market-sharing under quota system is equal market share for equal firms. This criterion is applicable where all have identical cost and revenue curves. This criterion also leads to a monopoly solution. It also resembles Chamberlin's duopoly model.

To illustrate the quota allocation, let us assume that there are only two firms, A and B. Their AR, MR and MC curves are given as shown in Figure 3.41(a) and (b). The market revenue and cost curves, which are obtained by adding up individual revenue and cost curves, respectively, are presented in part (c) of the figure. The industry output is determined at OQ_M. The quota for each firm, which

NOTES

maximizes their profits, is so determined that $OQ_M = OQ_A + OQ_B$. Given the identical cost and revenue conditions, $OQ_A = OQ_B$. That is, market is divided equally between firms *A* and *B*. This result can be obtained also by drawing an ordinate from point *R* where price line (P_M) intersects the *MR*.

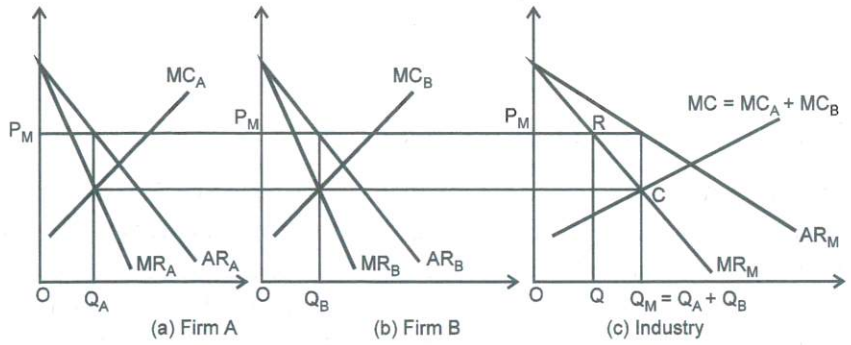


Fig. 3.41 *Quota Allocation under Cartel Agreements*

It may be mentioned at the end that cartels do not necessarily create the conditions for price stability in an oligopolistic market. Most cartels are loose. Cartel agreements are generally not binding on the members. Cartels do not prevent the possibility of entry of new firms. On the contrary, by ensuring monopoly profits, cartels in fact create conditions which attract new firms to the industry. Besides, chiselers and free riders create conditions for instability in price and output.

(ii) Price leadership models of oligopoly

Collusion through *price leadership* is another form of collusion between oligopoly firms. Price leadership is an informal position of a firm in an oligopolistic setting to lead other firm in fixing price of their product. This leadership may emerge spontaneously due to technical reasons or out of tacit or explicit agreements between the firms to assign leadership role to one of them.

The *spontaneous* price leadership may be the result of such technical reasons as size, efficiency, economies of scale or firm's ability to forecast market conditions accurately or a combination of these factors. The most typical case of price leadership is the leading role played by the dominant firm, the largest firm in the industry. The dominant firm takes lead in price changes and the smaller ones follow. Sometimes price leadership is *barometric*. In the barometric price leadership, one of the firms, not necessary the dominant one, takes lead in announcing change in price, particularly when such a change is due but is not affected due to uncertainty in the market.

The price leadership is possible under both *product homogeneity* and *product differentiation* or *heterogeneity*. There may be however price differentials commensurating with product differentiation. Price differentials may also exist on account of cost differentials.

Another important aspect of price leadership is that it often serves as a means to price discipline and price stabilisation. Achievement of this objective establishes an effective price leadership. Such price leadership can however exist only when (a) number of firms is small; (b) entry to the industry is restricted;

NOTES

(c) products are, by and large, homogeneous; (d) demand for industry is inelastic or has a very low elasticity; and (e) firms have almost similar cost curves.

The three common types of price leaderships are as follows:

(a) Price leadership by a low-cost firm

(b) Price leadership by a dominant firm

(c) Barometric price leadership

(a) Price leadership by a low-cost firm

How price and output decisions are taken under price leadership of a low-cost firm is illustrated in Figure 3.42. Suppose all the firms face identical revenue curves as shown by $AR = D$ and MR. But the largest firm or the low-cost firm, has its cost curves as shown by AC_1 and MC_1 whereas all other rival firms, smaller in size have their cost curves as shown by AC_2 and MC_2. The largest firm has the economies of scale and its cost of production is lower than that of other firms. Given the cost and revenue conditions, the low-cost firm would find it most profitable to fix its price at OP_2 ($= LQ_2$) and sell quantity OQ_2. Since at this level of output its $MC = MR$, its profit will be maximum. On the other hand, the high-cost firms would be in a position to maximize their profit at price OP_3 and quantity OQ_1. However, if low-cost firms charge profit maximizing price OP_3, they would lose their customers to the low-cost firm charging a lower price OP_2. The high-cost firms are therefore forced to accept the price OP_2 and recognize the price leadership of the low-cost firm. Note that the low-cost firm can eliminate other firms and become a monopolist, by cutting its price down to OP_1. At price OP_1, the low-cost firm can sell the same quantity OQ_2 and make, of course, only normal profit as its AC = price OP_1. But, it may not do so for the fear of anti-monopoly laws.

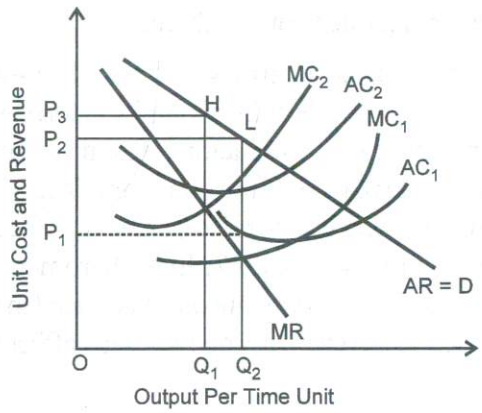


Fig. 3.42 Price Leadership

(b) Price leadership by the dominant firm

Price leadership by the dominant firm is more common than by a low-cost firm. In the analysis of price leadership by a dominant firm, it is assumed that there exists a large-sized firm in the industry, which supplies a large proportion of the total market. The dominance of the large firm is indicated by the fact that it could possibly eliminate all its rival firms by price cutting. But then the large firm gains the

NOTES

status of a monopoly which may create legal problems. The dominant firm, therefore, compromises with the existence of rival firms in the market. It uses its dominance to set its price so as to maximize its price. The smaller firms have no alternative but to accept the price set by the dominant firm. The smaller firms recognize their position and behave just like a firm in a perfectly competitive market. That is, smaller firms assume that their demand curve is a straight horizontal line.

The price leadership and market sharing between the dominant firm and the other firms as a group is illustrated in Figure 3.43. Suppose that the market demand curve is given by DD_M in part (a) of the figure. The problem confronting the dominant firm is to determine its price and output that will maximize its profits, leaving the rest of the market to be jointly supplied by the small firms. Now, the dominant firm has to find its own demand curve. Given the market demand curve (DD_m) and joint supply curve of small firms (SS_s), the dominant firm finds its demand curve by deducting from the market demand the quantity supplied jointly by the small firms below the equilibrium price. The part of the market demand not supplied by the small firms will be its own share. Thus, the market share of the dominant firm equals the market demand less the share of small firms.

For example, suppose equilibrium price is set at OP_3, the total supply by the smaller firms is P_3E which equals the market demand. Therefore, at price OP_3, the market left for the dominant firm is zero. When price is $O\overline{P}$, market demand is $\overline{P}B$ out of which $\overline{P}A$ is supplied by smaller firms. The market unsupplied by the smaller firms is AB. Thus, at price $O\overline{P}$, the demand for dominant firm's product equals

$$\overline{P}B - \overline{P}A = AB$$

Similarly, when price is reduced to OP_2, the demand for dominant firm's product is CF. Following this process, the market share of the dominant firm at other prices can be easily obtained.

The information so derived and plotted graphically gives P_3D_L as the demand curve for the dominant form [Figure 3.43(b)]. Since the relation between AR and MR is known, the MR curve for the dominant firm can be derived as MR_L [Figure 3.43(b)]. If the MC curve of the dominant firm is assumed to be given as MC_L, its profit maximizing output will be OQ_L and price $O\overline{P} = PQ_L$.

Once the dominant firm sets its price at $O\overline{P}$, the market demand curve for the small firms is the horizontal straight line $\overline{P}B$, because they can sell, at this price, as much as they can produce. But, in order to maximize their joint profits, small firms will produce only $\overline{P}A$. Recall that given the price, the line $\overline{P}B$ is the same as their $AR = MR$ line and their supply curve P_1S_s intersects $AR = MR$ at point A. For small firms, therefore, profit-maximizing output is $\overline{P}A$.

NOTES

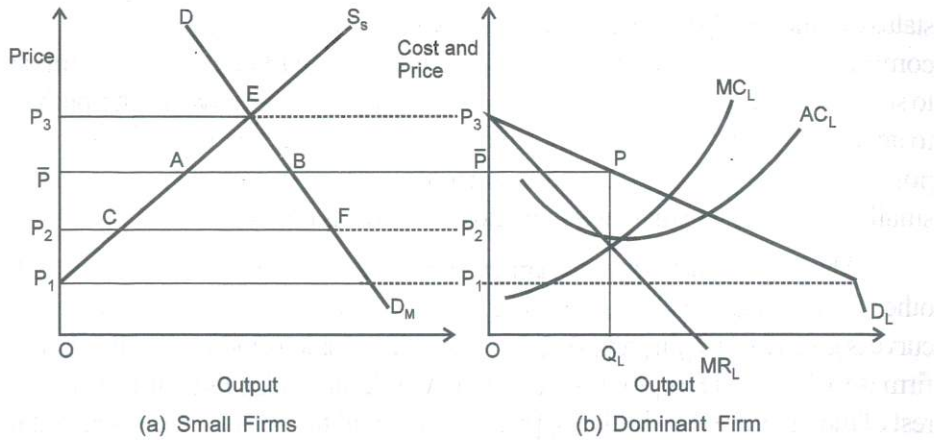


Fig. 3.43 Price Determination by the Dominant Firm

Finally, the dominant firm sets its price at $O\bar{P}$, which is accepted by the small firms. Thus, the dominant firm plays the role of a price leader. If it wants to eliminate the small firm it may set its price at OP_1 (though at a loss in the short run) at which small firms would not be able to survive. But, for the legal reason mentioned above, the dominant oligopoly firm would not do so. It would prefer, and be content, with its position of a price leader.

Critical appraisal of price leadership model

The price leadership model, as presented above, yields a stable solution to the problem of oligopoly pricing and output determination, only if small firms faithfully follow the leader. That is, small firms produce a right quantity and charge the price set by the dominant firm. Besides, the model requires that the dominant firm should be both a large-and low-cost firm. For, if a firm does not enjoy the advantages of being large enough and, consequent upon it, the advantages of its low cost, it cannot act as a price leader.

In practice, however, one finds many cases of price leadership by a firm which is neither a large nor a low-cost firm. But such cases are found mostly under recessionary conditions when a relatively smaller firm reduces its price to survive in the market.

Furthermore, if a leading firm loses its cost advantages, it loses also its leadership. Such cases are frequent in the real business world. Leadership also changes following the innovations of products and techniques of production by the smaller firms.

Besides, where there are many large firms of equal size and have some cost advantage, price leadership of any firm or group of firms becomes less probable, particularly when number of small firms is smaller than that of large firms. Under such conditions, barometric leadership emerges.

Lastly, it is assumed that entry of new firms is prevented either by low cost or by initial high cost. In practice, however, many firms having the capacity to diversify their products enter the industry with relatively initial low cost.

For these reasons, leadership model is not a realistic one as it is based on unrealistic assumptions. For the same reasons, the solution given by leadership model may not be stable.

NOTES

(c) Barometric price leadership

Another form of price leadership is *barometric* price leadership. In this form of price leadership, a firm initiates well-publicized changes in price that are generally followed by the rival firms in the industry. The price leader may not necessarily be the largest firm of the industry. The barometric firm is, however, supposed to have a better knowledge of prevailing market conditions and has an ability to predict the market conditions more precisely than any of its competitors. This qualification of the barometric firm should have been established in the past. Price decisions by a firm having the qualifications of price leadership is regarded as a barometer, which reflects the changes in business conditions and environment of the industry. The price changes announced by the barometric firm serves as a barometer of changes in demand and supply conditions in the market.

The barometric leadership evolves for various reasons of which the major ones are following:

First, the rivalry between the larger firms may lead to cut-throat competition to the disadvantage of all the firms. On the other hand, rivalry between the larger firms may make them unacceptable as a leader. So, a firm which has better predictive ability emerges as price leader.

Second, most firms in the industry may have neither the capacity nor the desire to make continuous calculations of cost, demand and supply conditions. Therefore, they find it advantageous to accept the price changes made by a firm which has a proven ability to make reasonably good forecasts.

Third, Kaplan state that barometric price leadership often develops as a reaction to a long economic warfare in which all the firms are losers.

Concluding remarks on oligopoly models

Most oligopoly models concentrate on price competition. In reality, oligopolists may be reluctant to wage price war and encroach upon each other's market share. It means that there is an absence of price competition in the oligopolistic market structure. The absence of price competition should not mean the absence of competition among oligopoly firms. In fact, the competition among oligopoly firms takes the form of *non-price competition.* The forms of non-price competition are diverse. Yet, there are two most important methods of non-price competition.

First, non-price competition involves product differentiation which is intended to attract new customers by creating preference for the new design and variety of product.

Second, perhaps the most important technique of non-price competition is advertisement. The primary objective of advertising is to make the demand curve for the product shift upward. The sellers try to encroach on the markets of other sellers through advertising. Advertising is also necessary to retain the market-share if there is tough competition between the firms.

3.6.3 The Game Theory Approach to Oligopoly

The classical theories show, in fact, only the beginning of the effort to analyse firms' behaviour in an oligopoly market. In recent years, economists have attempted

NOTES

to use a mathematical technique called game theory to explain the collusion among the oligopoly firms. The *game theory approach* was developed in 1944 by a mathematician John von Neumann (1903–57) and an economist Oscar Margenstern (1902–77). In recent times, the game theory approach is regarded as 'economists' most widely used approach to analyse oligopoly behaviour'.

In this section, we discuss the *game theory approach* to explain the strategic interaction among the oligopoly firms. This approach uses the apparatus of *game theory*—a mathematical technique—to show how oligopoly firms play their game of business. The game theory approach uses two fundamental tools—*strategic actions* by the firms and *pay-off matrix* of their actions—to find the optimum solution. As already mentioned, the first systematic attempt was made in this field by von Neumann and Margenstern. Though their work was followed by many others, Martin Shubik is regarded as the 'most prominent proponent of the game theory approach' who seems to believe that the only hope for the development of a general theory of oligopoly is the games theory. Though his hope does not seem to be borne out by further attempts in this area, the usefulness of game theory in revealing the intricate behavioural pattern of the oligopoly firms cannot be denied.

The nature of the problem: Prisoners' dilemma

The nature of the problem faced by the oligopoly firms is best explained by the *prisoners' dilemma game.* To illustrate *prisoners' dilemma,* let us suppose that there are two persons, *A* and *B* who are partners in a case of kidnapping for ransom. On a tip-off, the CBI arrests *A* and *B* on suspicion of their involvement in kidnapping a person. They are arrested and lodged in separate jails with no possibility of communication between them. They are being interrogated separately by the CBI officials with following conditions disclosed to them in isolation.

1. If you confess your involvement in kidnapping, you will get a 5-year imprisonment.
2. If you deny your involvement and your partner denies too, you will be set free for lack of evidence.
3. If one of you confesses and turns approver, and other does not, then one who confesses gets a 2-year imprisonment, and one who does not confess gets 10-year imprisonment.

Given these conditions, each suspect has two options open to him: (i) to confess, and (ii) not to confess. Now, both *A* and *B* face a dilemma on how to decide whether or not to confess. While taking a decision, both have a common objective, i.e., to minimize the period of imprisonment. Given this objective, the option is quite simple that both of them deny their involvement in kidnapping. But there is no certainty that if one denies, the other will also deny: the other may confess and turn approver with this uncertainty, the dilemma in making a choice still remains. For example, if *A* denies his involvement, and *B* confesses (settles for a 2-year imprisonment), then *A* gets a 10-year jail term. So is the case with *B*. If they both confess, then they get a 5-year jail term each. Then what to do? That is the dilemma. The nature of their problem of decision making is illustrated in the Table 3.4 in the

form of a 'pay-off matrix'. The pay-off matrix shows the pay-offs of their different options in terms of the number of years in jail.

Table 3.4 *Prisoner's Dilemma*

		B's Option			
		Confess		Deny	
		A	*B*	*A*	*B*
	Confess	5	5	2	10
A's Options		*A*	*B*	*A*	*B*
	Deny	10	2	0	0

Given the conditions, it is quite likely that both the suspects may opt for 'confession', because neither *A* knows what *R* will do nor *B* knows what *A* will do. When they both confess, each gets a 5-year jail term. This is the second best option. For his decision to confess, *A* might formulate his strategy in the following manner. He argues to himself: If I confess (though I am innocent), I will get a maximum of 5 years' imprisonment. But, if I deny (which I must) and *B* confesses and turns approver, I will get 10 years' imprisonment, which will be the worst of the worst. It is quite likely that suspect *B* also reasons out in the same manner, even if he too is innocent.

If they both confess, they would avoid 10 year's imprisonment, the maximum possible jail sentence under the law. This is the best they could achieve under the given conditions.

Relevance of prisoners' dilemma to oligopoly

The prisoners' dilemma illustrates the nature of problems oligopoly firm are confronted with in the formulation of their business strategy with respect to strategic advertising, price cutting and cheating in case of a cartel. Look at the nature of problems an oligopoly firm is confronted with when it plans to increase its *advertisement* expenditure (ad-expenditure for short). The basic issue is whether or not to increase the ad-expenditure. If the answer is 'do not increase', then the questions are: will the rival firms increase ad-expenditure or will they not? And if they do, what will be the consequences for the firm under consideration? And, if the answer is 'increase', then the following questions arise: what will be the reaction of the rival firms? Will they increase or will they not increase their ad-expenditure? What will be the pay-off if they do not and what if they do? If the rival firms do increase their advertising, what will be the pay-off to the firm? Will the firm be a net gainer or a net loser? The firm will have to find the answer to these queries under the conditions of uncertainty. It will have to anticipate actions, reactions and counteraction by the rival firms and chalk out to own strategy. It is in case of such problems that the case of prisoners' dilemma becomes an illustrative example.

Application of game theory to oligopoly

Let us now apply the game theory to an example of 'whether or not to increase ad-expenditure', assuming that there are only two firms, *A* and *B,* i.e., the case of

NOTES

a duopoly. We know that in all the games, the players have to anticipate the move made by the opposite player(s) and formulate their own strategy to counter the different possible moves by the rival. To apply the game theory to the case of 'whether or not to increase ad-expenditure' the firm needs to know or anticipate the following:

(i) Counter moves by the rival firm in response to increase in ad-expenditure by this firm, and

(ii) The *pay-offs* of this strategy when (a) the rival firm does not react, and (b) the rival firm does make a counter-move by increasing its ad-expenditure.

After these data are obtained, the firm will have to decide on the best possible strategy for playing the game and achieving its objective of, say, increasing sales and capturing a larger share of the market. The best possible strategy in game theory is called the 'dominant strategy'. A *dominant strategy is one that gives optimum pay-off, no matter what the opponent does.* Thus, the basic objective of applying the game theory is to arrive at the dominant strategy.

Suppose the possible outcomes of the ad-game under the alternative moves are given in the pay-off matrix presented in Table 3.5.

Table 3.5 *Pay-off Matrix of the Ad-Game*
(Increase in sales in milion `)

	B's Option			
	Increase Ad		Don't increase	
	A	*B*	*A*	*B*
Increase Ad 20	10	30	0	
A's Options	*A*	*B*	*A*	*B*
Don't increase	10	5	15	5

As the matrix shows, if Firm *A* decides to increase its ad-expenditure, and Firm *B* counteracts *A's* move by increasing its own ad-expenditure, Firm *A's* sales go up by ₹20 million and that of Firm *B* by ₹10 million. And, if Firm *A* increases its advertisement and *B* does not, then Firm *A*'s sales gain is ₹30 million and no gain to Firm *B*. One can similarly find the pay-offs of the stategy 'Don't increase' in case of both of firms.

Given the pay-off matrix, the question arises, what strategy should Firm *A* choose to optimize its gain from extra ad-expenditure, irrespective of counteraction by the rival Firm *B*. It is clear from the pay-off matrix that Firm *A* will choose the strategy of increasing the ad-expenditure because, no matter what Firm *B* does, its sales increase by at least ₹20 million. This is, therefore, the *dominant strategy* for Firm *A*. A better situation could be that when Firm *A* increases its expenditure on advertisement, Firm *B* does not. In that case, Firm *A*'s sales could increase by ₹30 million and sales of Firm *B* do not increase. However, there is a greater possibility that Firm *B* will go for counter-advertising in anticipation of losing a part of its market to Firm *A* in future. Therefore, a strategy based on the assumption that Firm *B* will not increase its ad-expenditure involves a great of uncertainty.

NOTES

Nash equilibrium

In the preceding section, a very simple example to illustrate the application of game theory to an oligopolistic market setting, with the simplifying assumptions: (i) that strategy formulation is a one-time affair, (ii) that one firm initiates the competitive warfare and other firms only react; and (iii) that there exists a *dominant strategy*—a strategy which gives an optimum solution. The real-life situation is, however, much more complex. There is a continuous one-to-one and tit-for-tat kind of warfare. Actions, reactions and counteractions are regular phenomena. Under these conditions, a *dominant strategy* is often non-existent. To analyse this kind of situation, John Nash, an American mathematician, developed a technique, known as *Nash equilibrium. Nash equilibrium technique* seeks to establish that each firm does the best it can, given the strategy of its competitors and a *Nash equilibrium is one in which none of the players can improve their pay-off given the strategy of the other players.* In case of our example, Nash equilibrium can be defined as one in which none of the firms can increase its pay-off (sales) given the strategy of the rival firm.

The Nash equilibrium can be illustrated by making some modifications in the pay-off matrix given in Table 3.5. Now we assume that action and counter action between Firms *A* and *B* is a regular phenomenon and the pay-off matrix that appears finally, as given in Table 3.6. The only change in the modified pay-off matrix is that if neither Firm *A* nor Firm *B* increases its ad-expenditure, then pay-offs change from (15, 5) to (25, 5).

Table 3.6 *Pay-off Matrix of the Ad-Game*
(Increase in sales in milion `)

		B's Option			
		Increase Ad		Don't increase	
		A	*B*	*A*	*B*
	Increase Ad	20	10	30	0
A's Strategy		*A*	*B*	*A*	*B*
	Don't increase	10	15	25	5

It can be seen from the pay-off matrix (Table 3.5) that Firm *A* has no more a *dominant strategy.* Its optimum decision depends now on what Firm *B* does. If Firm *B* increases its ad-expenditure, Firm *A* has no option but to increase its advertisement expenditure. And, if Firm *A* reinforces its advertisement, Firm *B* will have to follow the suit. On the other hand, if Firm *B* does not increase its ad-expenditure, Firm *A* does the best by increasing its ad-expenditure. Under these conditions, the conclusion that both the firms arrive at is to increase ad-expenditure if the other firm does so, and 'don't increase', if the competitor 'does not increase'. In the ultimate analysis, however, both the firms will decide to increase the ad-expenditure. The reason is that if none of the firms increases advertisement, Firm *A* gains more in terms of increase in its sales (₹25 million) and the gain of Firm *B* is much less (₹5 million only). And, if Firm *B* increases advertisement expenditure, its sales increase by ₹10 million. Therefore, Firm *B*

would do best to increase its ad-expenditure. In that case, Firm *A* will have no option but to increase its ad-expenditure, Thus, the *final conclusion* that emerges is that both the firms will go for advertisement war. In that case, each firm finds that it is doing the best given what the rival firm is doing. This is the Nash equilibrium.

However, there are situations in which there can be more than one Nash equilibrium. For example, if we change the pay-off in the south-east corner from (25, 5) to (22, 8), each firm may find it worthless to wage advertisement war and may settle for 'don't increase' situation. Thus, there are two possible Nash equilibria.

Conclusion

The game theory can be used to find an equilibrium solution to the problems of an oligopolistic market setting under different assumptions regarding the behaviour of the oligopoly firms and market conditions. However, despite its merit of revealing the nature and pattern of oligopolistic warfare, the game theory often fails to provide a determinate solution.

Check Your Progress

7. What is the difference in the assumptions of monopolistic competition and pure competition?
8. Why is their inter-dependence between the oligopolists in business decisions?
9. How does Bertrand's model differ from Cournot's model?
10. What are the two methods of market allocation?

3.7 PRICING STRATEGIES

We have discussed the conventional theories of price determination under the conditions of different market structures, under the postulate that firms' objective is to maximize profit. The alternative theories of firm offered by the next generation of economists have built their theories assuming different objectives of firms. A section of economists has built game theory and have shown its application to business decision-making. As noted in previous chapters, all these theories and strategic models have their own deficiencies and problems in application because of increasing complexity of the business world. The other economists have recognized the complexity of business world and have explained how firms formulate their strategic pricing policy and determine the price of their product. In this section, you will learn about some of the major pricing strategies.

3.7.1 Pricing in Life Cycle of a Product

The life cycle of a product is generally divided into five stages: (i) introduction or initial stage, (ii) growth, (iii) maturity, (iv) saturation, and (v) decline. Figure 3.44 presents the five stages of a product's life cycle through a curve showing the behaviour of the total sales over the life cycle. The *introduction* phase is the period taken to introduce the product to the market. The total sale during this

period is limited to the quantity put on the market for trial with considerable advertisement. The sales during this period remain almost constant. *Growth* is the stage, after a successful trial, during which the product gains popularity among the consumers and sales increase at an increasing rate as a result of cumulative effect of advertisement over the initial stage. *Maturity* is the stage in which sales continue to increase but at a lower rate and the total sale eventually becomes constant. During the *saturation period* the total sale saturates—there is neither increase nor decrease in the sales volume. After the saturation stage, comes the stage of *decline* in which the total sales register a declining trend for such reasons as (i) increase in the availability of substitutes, and (ii) loss of distinctiveness of the product.

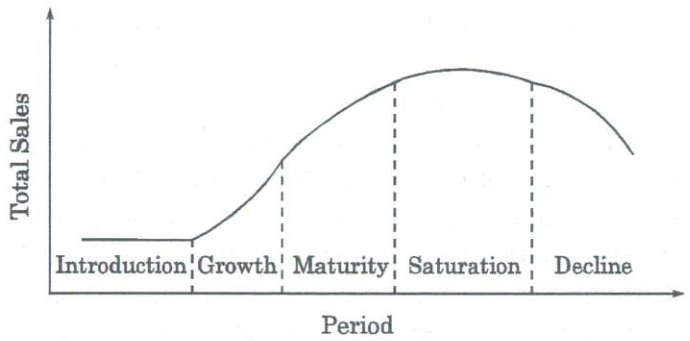


Fig. 3.44 *Life Cycle of a Product*

The pricing strategy varies from stage to stage over the life cycle of a product, depending on the market conditions. From the pricing strategy point of view, growth and maturity stages may be treated likewise. Let us discuss the pricing of a product in its initial stage as pricing of a new product and then in the 'maturity' and 'decline' stage.

- **Pricing in growth stage**

A new product may simply be either another brand name added to the existing ones or an altogether new product. Pricing a new brand for which there are many substitutes in the market is not as big a problem as pricing a new product for which close substitutes are not available. For, in case of the former, the market provides adequate information regarding cost, demand, and availability of market. Pricing in this case depends on the nature of the market. In pricing a new product without close substitutes, however, problems arise because, for lack of information, there is some degree of uncertainty.

Thus, pricing policy in respect of a new product depends on whether or not close substitutes are available. Depending on whether or not close substitutes are available, in pricing a new product, generally two types of pricing strategies are suggested, viz. (i) skimming price policy, and (ii) penetration price policy.

(i) **Skimming price policy:** This *policy* is adopted where close substitutes of a new product are not available. This pricing strategy is intended to skim the cream of the market, i.e., consumer's surplus, by setting a high initial price, three or four times the ex-factory price, and a subsequent lowering of prices in a series of reductions, especially in case of consumer durables.

NOTES

The initial high price would generally be accompanied by heavy sales promotion on expenditure.

(ii) **Penetration price policy:** In contrast to skimming price policy, the penetration price policy involves a reverse strategy. This pricing policy is generally adopted in the case of new products for which substitutes are available. This policy requires fixing a lower initial price designed to penetrate the market as quickly as possible and intended to maximize profits in the long run. Therefore, the firms pursuing the penetration price policy set a low price for the product in the initial stage. As the product catches the market, price is gradually raised up.

The choice between the two strategic price policies depends on: (i) the rate of market growth, (ii) the rate of erosion of distinctiveness, and (iii) the cost-structure of the producers. If the rate of market growth is slow for such reasons as lack of information, consumers' hesitation, etc., the penetration price policy would be unsuitable. The reason is that a low price will not mean a large sale. If the pioneer product is likely to lose its distinctiveness at a faster rate, skimming price policy would be unsuitable— it should be followed when lead time, i.e., the period of distinctiveness is fairly long. If cost-structure shows increasing returns over time, the penetration price policy would be more suitable, since it enables the producer to reduce his cost and prevents potential competitors from entering the market in the short-run.

- **Pricing in maturity period**

Maturing period is the second stage in the life-cycle of a product. It is a stage between the growth period and the decline period of sales. Sometimes, maturity period is bracketed with saturation period. Maturity period may also be defined as the period of decline in the growth rate of sales (not the total sales) and the period of zero growth rate. The concept of maturity period is useful to the extent it gives out signals for taking precaution with regard to pricing policy. However, the concept itself does not provide guidelines for the pricing policy. Joel Dean suggests that the 'first step for the manufacturer whose speciality is about to slip into the commodity category is to reduce real prices as soon as the system of deterioration appears'. But he warns that this does not mean that the manufacturer should declare open price war in the industry. But he should rather move in the direction of 'product improvement and market segmentation.'

- **Pricing a product in decline**

The product in decline is one that enters the post-maturity stage. In this stage, the total sale of the product starts declining. The first step in pricing strategy in this stage is obviously to reduce the price. The product should be re-formulated and re-modelled to suit the consumers' preferences. It is a common practice in the book trade. When the sale of a hard-bound edition reaches saturation, its paper-back edition is brought into the market. This facility is, however, limited to only a few commodities. As a final step in the strategy, the advertisement expenditure may be reduced drastically or withdrawn completely, and rely on the residual market. This, however, requires a strong will of the producer.

NOTES

Pricing in relation to established products

Many producers enter the market often with a new brand of a commodity for which a number of substitutes are available. For example, the cold drinks like Coke and Spot, were quite popular in the market when new cold drinks brands like Limca, Thums Up, Double Seven, Mirinda, Pepsi, Teem, Campa, etc., were introduced in the market over time. So has been the case with many other consumer goods. Many other models of motor cars appeared in the market despite the popularity of Maruti cars. A new entrant to the market faces the problem of pricing his product because of strong competition with established products. This problem of pricing of a new brand is known as pricing in relation to the established products.

In pricing a product in relation to its well-established substitutes, generally three types of pricing strategies are adopted, viz., (i) pricing below the ongoing price, (ii) pricing at par with the prevailing market price, and (iii) pricing above the existing market price. Let us now see which of these strategies are adopted under what conditions.

(i) Pricing below the market price

Pricing below the prevailing market price of the substitutes is generally preferred under two conditions. First, if a firm wants to expand its product-mix with a view to utilizing its unused capacity in the face of tough competition with the established brands, the strategy of pricing below the market price is generally adopted. This strategy gives the new brand an opportunity to gain popularity and establish itself. For this, however, a high cross-elasticity of demand between the substitute brands is necessary. This strategy may, however, not work if existing brands have earned a strong brand loyalty of the consumers. If so, the price incentive from the new producers must, therefore, overweigh the brand loyalty of the consumers of the established products, and must also be high enough to attract new consumers. This strategy is similar to the penetration pricing. Second, this technique has been found to be more successful in the case of innovative products. When the innovative product gains popularity, the price may be gradually raised to the level of market price.

(ii) Pricing at market price

Pricing at par with the market price of the existing brands is considered to be the most reasonable pricing strategy for a product which is being sold in a strongly competitive market. In such a market, keeping the price below the market price is not of much avail because the product can be sold in any quantity at the existing market rate. The strategy is also adopted when the seller is not a 'price leader'. It is rather a 'price-taker' in an oligopolistic market. This is, in fact, a very common pricing strategy, rather the most common practice.

(iii) Pricing above the existing market price

This strategy is adopted when a seller intends to achieve a prestigious position among the sellers in the locality. This is a more common practice in case of products considered to be a commodity of conspicuous consumption or prestige goods or deemed to be of much superior quality. Consumers of such goods prefer shopping

in a gorgeous shop of a posh locality of the city. This is known as the 'Veblen Effect'. Sellers of such goods rely on their customers' high propensity to consume a prestigious commodity. After the seller achieves the distinction of selling high quality goods, though at a high price, they may even sell ordinary goods at a price much higher than the market price. This practice is common among sellers of readymade garments.

Besides, a firm may sets a high price for its product if it pursues the skimming price strategy. This pricing strategy is more suitable for innovative products when the firm can be sure of the distinctiveness of its product. The demand for the commodity must have a low cross-elasticity in respect of competing goods.

3.7.2 Competitive Bidding of Price

A kind of pricing in which a firm is required to quote its price under uncertain cost and price conditions, with a view to winning a contract or a tender is known as competitive bidding or contract pricing. Competitive bidding is a process of quoting a contract price for supplying goods or services under specified terms and conditions.

In contractual business, there are different types of transactions. For example, the purchaser specifies the quantity and quality of goods (or services) and the time and place of delivery, along with other terms and conditions. The offer to buy certain goods or services or for construction of a building, road, dam, etc., are made known to the potential sellers or contractors through public notification inviting 'tenders'. The suppliers and contractors interested in the tender evaluate the offer in terms of its cost and profitability. If they find the offer profitable, they submit their tender. But the problem here is not simply to quote a suitable supply price while submitting the tender. In fact, the foremost problem is to quote a supply price which can win the contract without unduly reducing the profit margin. For this, the quoted supply price should be lower than that of the rival contractors.

To bid a contract winning supply price is an extremely difficult task, mainly because the prices of the rival firms or contractors are unknown. The problem becomes much more complicated if there is uncertainty about the future prices of the inputs, particularly when input prices are subject to fluctuation. Uncertainty about future cost conditions increases the degree of risk because bidding takes place at current prices, and the contracted goods and services are to be produced and supplied at future input prices which may be uncertain. Despite these difficulties in bidding competitive prices contractors do quote a supply price which wins them a contract and yields profits. Let us briefly discuss the process of competitive price bidding, assuming that there are no 'bribes' and 'kickbacks'.

Major factors in competitive bidding

There are three major factors which contractors analyse in the process of competitive bidding. These are as follows:

(i) Bidder's current and projected capacity to handle the contract

(ii) Overall objective of the bidder

(iii) Expected bid of the rival contractors

A bidder's present and projected capacity to handle the contract matters a great deal in competitive bidding. Given his present capacity, the contractor may find himself with (*a*) excess capacity, (*b*) full utilization of capacity, and (*c*) undertaking activity in excess of capacity. The three kinds of different capacity positions put the bidder in three different positions to bid the supply price. If the bidder has an excess capacity, he may bid a price at par with his cost or even below the cost if his circumstances so demand. This kind of bidding may be motivated by the contractor's desire (*i*) to popularize his business, (*ii*) to retain the contract-giving parties in face of tough competition, and (*iii*) to retain the labour and other services. Even if contractors are engaged to their capacity, they may submit their tender with a view to maintaining their reputation, and to give the potential contracting bodies a feeling that they can still handle more business, though they may not really want to win the contract. If they do not want to win the contract, they usually bid a high price and lose the contract to their competitors, but not necessarily. If a contractor has a high reputation in comparison with his rivals, there is always a chance that the contract is awarded to him. Therefore, they keep their bid high enough to cover the additional cost resulting from their overcapacity operation and the extra cost in case they go for sub-contracting a part or whole of their new contracts.

3.7.3 Transfer Pricing

Large-sizes firms, very often, divide their operation into product divisions or subsidiaries. Growing firms add new divisions or departments to the existing ones. These firms then transfer some of their activities to other divisions. The goods and services produced by the new division are used by the parent organization. In other words, the parent division buys the product of its subsidiaries. Such firms face the problem of determining an appropriate price for the product transferred from one division or subsidiary to the other. Specifically, the problem is of determining the price of a product produced by one division of the same firm. This problem becomes much more difficult when each division has a separate profit function to maximize. Pricing of intra-firm 'transfer product' is referred to as 'transfer pricing'. One of the most systematic treatments of the transfer pricing technique has been provided by Hirshleifer.

Let us suppose that a refrigerator company, established a decade ago, used to produce and sell refrigerators fitted with compressors bought from a compressor manufacturing company. Now the refrigerator company decides to set up its own subsidiary to manufacture compressors. Let us also assume that:

(i) Both parent and subsidiary companies have their own profit functions to maximize.

(ii) The refrigeration company sells its product in a competitive market and its demand is given by a straight horizontal line.

(iii) The refrigerator company uses all the compressors produced by its subsidiary.

In addition, assume that there is no external market for the compressors. Later drop this assumption and alternatively assume that there is an external market for the compressors and discuss the technique of transfer pricing under both the

NOTES

alternative conditions. Let us first discuss transfer pricing with no external market.

(a) Transfer pricing without external market

Given the foregoing assumptions, the company has to set an appropriate price for the compressors so that the profit of its subsidiary too is maximum. To deal with the 'transfer pricing' problem, let us first look into the pricing and output determination of the final product, i.e., refrigerators. Since the company sells its refrigerators presumably in a competitive market, the demand for its product is given by a straight horizontal line as shown by the line $AR_r = MR_r$ in Figure 3.45.

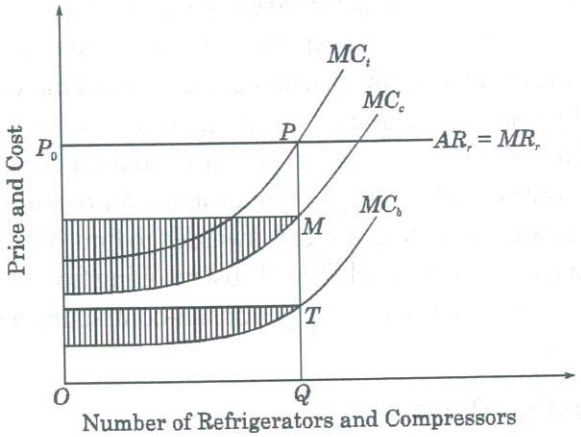


Fig. 3.45 Price Determination of the Final Product (Refrigerators)

The marginal cost of intermediate good, i.e., compressor, is shown by MC_c curve and that of the refrigerator body by MC_b. The MC_c and MC_b added vertically give the combined marginal cost curve, the MC_t. At output OQ, for example, $TQ + MQ = PQ$. The MC_t intersects line $AR_r = MR_r$ at point P. An ordinate drawn from point P down to the horizontal axis determines the most profitable outputs of refrigerator body and compressors each at OQ. Thus, the output of both refrigerator body and compressors is simultaneously determined. Since at OQ level of output, the firm's $MC_t = MR_r$, the refrigerator company maximizes its profits from the final product, the refrigerators.

Now, let us find the price of the compressors. The question that arises is: what should be the price of the compressors so that the compressor manufacturing division, too, maximizes its profit? The answer to this question can be obtained by applying the marginality principle which requires equalizing MC and MR in respect of compressors. The marginal cost curve for the compressors is given by MC_c in Figure 3.46. The firm, therefore, has to obtain the marginal revenue for its compressors. The marginal revenue of the compressors (MR_c) can be obtained by subtracting the non-compressor marginal cost of the final good from the MR_r. Thus,

$$MR_c = MR_r - (MC_t - MC_c) \quad \ldots (3.21)$$

For example, in Figure 3.45, at output OQ, $MR_r = PQ$, $MC_t = PQ$, and $MC_c = MQ$. By substituting these values in Equation 3.21, we get

$$MR_c = PQ - (PQ - MQ)$$

$$PQ - PM = MQ$$

or, since in Figure 3.46, $PQ - MQ = PM$, and $PM = TQ$, therefore,

$$MR_c = PQ - TQ = PT$$

and

$$PT = MQ$$

NOTES

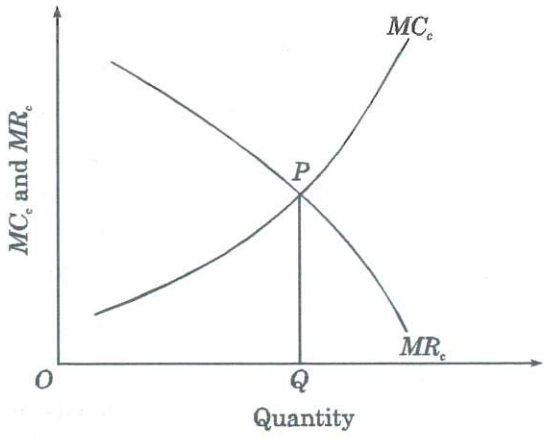


Fig. 3.46 *Determination of Transfer Price*

We may recall that $AR_r = MR_r$, i.e., MR_r is constant, and that MC_t is a rising function. Thus, $MR_r - MC_c$ will be a decreasing function. Notice the vertical distance between $AR_r = MR_r$ line, and MC_c curve is decreasing as shown in Figure 3.46. When MR_c (which equals $MR_r - MC_c$) is obtained for different levels of output and graphed, it yields a curve like MR_c curve shown in Figure 3.47. The MC_c curve (which is the same as MC_c curve in Figure3.46) intersects the MR_c at point P. At point P, $MR_c = MC_c$ and output is OQ. Thus, the price of compressors is determined at PQ in Figure 3.46. This price enables the compressor division to maximize its own profit.

(b) Transfer pricing with external competitive market

The transfer pricing under the assumption that there is no external market for the compressors. It implies that the refrigerator company was the sole purchaser of its own compressors and that the compressor division had no external market for its product. Let us now discuss the transfer pricing technique assuming that there is an external market for the compressors. The existence of the external market implies that the compressor division has the opportunity to sell its surplus production to other buyers and the refrigeration company can buy compressors from other sellers if the compressor division fails to meet its total demand. Also, assume that the external market is perfectly competitive. Determination of transfer price under these conditions is a little more complicated task.

The method of transfer pricing with external market is illustrated in Figure 3.47. Since the compressor market is perfectly competitive, the demand for compressor is given by a straight horizontal line as shown by P_2D in which case $AR = MR$. The marginal cost curve of the compressors is shown by MC_c. The MR_c curve shows the marginal net revenue from the compressor, (See Figure 3.47). Note that in the absence of the external market, the transfer price of compressor would have been fixed at $OP_1 = P'Q_2$, i.e., the price where $MR_c = MC_c$. At this price the parent company would have bought compressors only from its subsidiary. But, since compressors are to be produced and sold under

NOTES

competitive conditions, the effective marginal cost of the compressor to the refrigeration company is the market price of the compressor, i.e., OP_2. Besides, the price OP_2 is also the potential *MR* for the compressor division. Therefore, in order to maximize the profit, compressor's price will be set at point *P* where $MR_c > MC_c$. Thus, the transfer price of compressor will be fixed at PQ_1 and the refrigeration company would buy OQ_1 compressors from the compressor division.

The total output of compressors is determined at a level where MC_c intersects the demand line, *D* (*AR* = *MR*), i.e., at point *R*. At point *R*, the total output of compressors is OQ_3. Of this, OQ_1 is bought by the refrigeration company itself and the remaining output, Q_1Q_3 will be sold in the external market, both at price OP_2. At this level of output and price, the compressor division maximizes its profit.

Shift in MR_c and transfer price

Let us now consider how transfer price is determined when MR_c shifts upward to the right. The MR_c may shift upward because of an increase in demand causing an upward shift in *AR* = *MR*. Let the MR_c in Figure 3.47 shift to MR'_c which intersects the MC_c at point *B*. In the absence of an external market, the refrigeration company would have set transfer price of compressors at OP_3—a price higher than the free market price OP_2. But, since an external market does exist in which a price, OP_2, is given, the transfer price cannot exceed the market price or else the refrigerator company would not be in a position to maximize its profit. Moreover, the transfer price cannot be less than the market price, otherwise the compressor division would not be able to maximize its profit. Thus, if there is an external market in which market price of an intermediary product, produced by a subsidiary company is given, then the problem is to determine the quantity to be produced by the subsidiary and the quantity to be purchased from the external market.

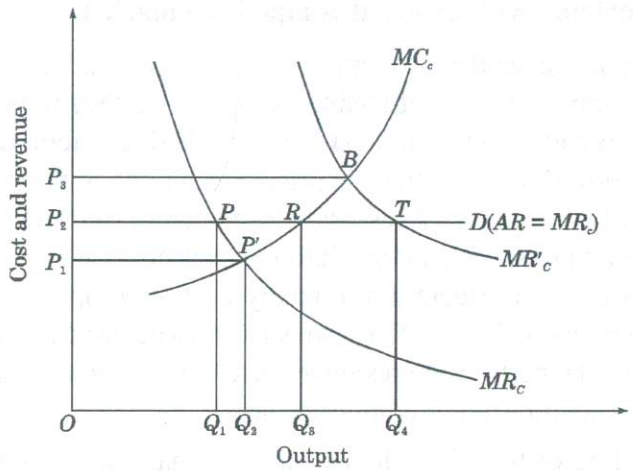


Fig. 3.47 Determination of Transfer Price with External Market

Figure 3.47 shows that after the shift in MR_c curve to MR'_c, the demand for compressors by the refrigeration company increases to OQ_4 where $AR = MR = MR'_c$. But the subsidiary company cannot produce OQ_4 units of compressors, given its MC_c and the market price. It will, therefore, produce OQ_3 number of compressors, which equalizes MC_c with *MR* at point *R*. Given the market price,

OQ_3 is the most profitable output of compressors. Therefore, the difference between the total demand and the total internal supply from the subsidiary, i.e., $OQ_4 - OQ_3 = Q_3Q_4$, will be bought in the external market, at price $OP_2 = TQ_4$. Thus, the refrigeration company will buy OQ_3 compressors from its compressor division and buy Q_3Q_4 in the external market.

(c) Transfer pricing under imperfect external market

When the refrigerator market is imperfect, the compressor division faces a demand curve with a negative slope in the external market, instead of a straight horizontal demand line. The downward sloping demand curve makes transfer pricing a much more complicated task. To illustrate the transfer pricing technique under imperfect market conditions in the external market, let us suppose, (*i*) that the average and marginal revenue curves for the compressors are given by AR_x and MR_x, respectively, in Figure 3.48, and (*ii*) that the 'marginal net revenue' from the internal use of compressors and the marginal cost of producing compressors are represented by MR_c and MC_c, respectively. With a view to maximizing the overall profit, the refrigerator company will determine the output of compressors where $MC_c = MR_c + MR_x$, i.e., where marginal cost of compressors equals the composite marginal revenue. The composite marginal revenue is obtained through horizontal summation of the MR_c and MR_x curves as shown by MR_t in Figure 3.48.

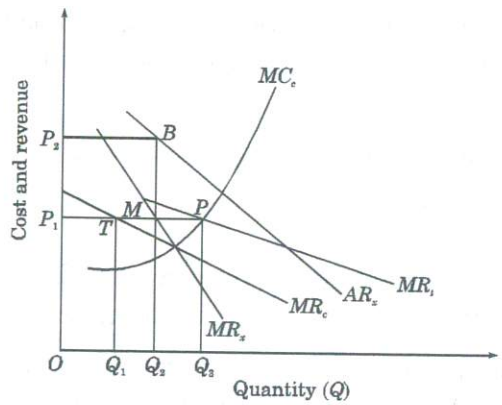


Fig. 3.48 *Transfer Pricing and Imperfect External Market*

As shown in Figure 3.48, MC_c intersects MR_t at point P which determines the output of compressors at OQ_3. The compressor division can maximize its profit by dividing its output between the refrigerator company and the external market so as to equalize its *MC* and *MR* in both the markets—internal and external. If a line (PP_1) is drawn from point *P* parallel to the horizontal axis to the vertical axis, it intersects MR_x at point *M* and MR_c at point *T*. The points of intersection (*T* and *M*) determine the share of refrigerator company and the external market in the total output OQ_3. At point *M*, $MC_c = MR_x$ and at point *T*, $MC_c = MR_c$. Thus, the refrigeration company, the parent body, will buy OQ_1 and sell OQ_2 in the open market. Note that $OQ_1 + OQ_2 = OQ_3$. The profit maximizing price in the external market is OP_2 (= BQ_2) and the profit maximizing transfer price is set at OP_1. With these prices and output, both refrigerator and compressor companies maximize their respective profits.

NOTES

Check Your Progress

11. How many stages are there in the life cycle of a product?
12. Which is the most suitable pricing strategy in a strongly competitive market?
13. What do you understand by competitive bidding?

3.8 ANSWERS TO 'CHECK YOUR PROGRESS'

1. On the basis of the degree of competition, the market structure is generally classified as (i) perfect competition (ii) imperfect competition (a) monopolistic competition (b) oligopoly with and without product differentiation (c) duopoly (iii) monopoly.
2. Two characteristics of perfect competition are the following:
   - Large number of buyers and sellers
   - Perfect mobility of factors of production
3. The supply curve of an individual firm is derived on the basis of its equilibrium output.
4. An important feature of a pure monopoly is that a monopolized industry is a single-firm industry, i.e., there is no distinction between the firm and the industry.
5. Patent rights are granted by the government to a firm to produce a commodity of specified quality and character or to use a specified technique of production.
6. When a monopolist sells the same product at different prices to different buyers, the monopoly is called discriminatory monopoly.
7. The assumptions of the monopolistic competition are the same as those of pure competition, with the exception of homogeneity of products. While pure competition model assumes that products are homogeneous in every possible dimension, monopolistic competition model assumes that products are differentiated.
8. The nature and degree of competition among the oligopolists makes them interdependent in respect of decision making. The reason for inter-dependence between the oligopolists is that a major policy change made by one of the firms affects the rival firms seriously and immediately and forces them to make counter-moves to protect their interest.
9. Bertrand's model differs from Cournot's model in respect of its behavioural assumption. While under Cournot's model, each seller assumes his rival's output to remain constant, under Bertrand's model, each seller determines his price on the assumption that his rival's price, rather than his output, remains constant.
10. The two main methods of market allocations are non-price competition and quota system.

NOTES

11. There are two main methods of market allocations: (i) non-price competition, and (ii) quota system.
12. The life cycle of a product is generally divided into five stages: (i) introduction or initial stage, (ii) growth, (iii) maturity, (iv) saturation, and (v) decline.
13. Pricing at par with the market price of the existing brands is considered to be the most reasonable pricing strategy for a product which is being sold in a strongly competitive market.
14. Competitive bidding is a process of quoting a contract price for supplying goods or services under specified terms and conditions

3.9 SUMMARY

- The theory of value propounded by Marshall on the assumptions of perfect competition and a static equilibrium system was regarded to provide answers to all questions regarding price and output determination.
- A monopoly firm has the sole power to influence the market price.
- The term 'market structure' refers to the organizational features of an industry that influence a firm's behaviour in its choice of price and output.
- Organizational features include the number of firms, distinctiveness of their products, elasticity of demand and the degree of control over the price of the product.
- The market structure is generally classified on the basis of the degree of competition—perfect competition, imperfect competition and monopoly.
- Perfect competition is a market situation in which a large number of producers offer a homogeneous product to a very large number of buyers of the product.
- Imperfect competition is said to exist when a number of firms sell identical or differentiated products with some control over the price of their product.
- Monopolistic competition is a kind of market in which a large number of firms supply differentiated products.
- Oligopoly is an organizational structure of an industry in which a small number of firms supply the entire market, each seller having a considerable market share and control over the price.
- Monopoly is the market of a single seller with control over his price and output.
- A perfectly competitive market is characterized by a complete absence of rivalry among the individual firms.
- According to the traditional theory of the firm, a firm is in equilibrium when its profit is maximum. Maximization of profits depends on the revenue and cost conditions.
- In the short-run, it may earn just a normal profit or even make losses.
- In order to minimize losses in short-run, a firm must cover its short-run average variable cost (SAVC).

NOTES

- The supply curve of an individual firm is derived on the basis of its equilibrium output at different levels of the price. The equilibrium output is determined by the intersection of MR and MC curves.
- Under perfect competition, market price in a perfectly competitive market is determined by the market forces, viz., demand and supply.
- The supply curve in the long-run is supposed to be more elastic than in the short-run.
- An industry is in equilibrium when its market demand equals its market supply.
- An industry in which factor prices are independent of the rate of increase in factor demand is referred to as the constant cost industry.
- An industry is referred to as an increasing cost industry if factor prices increase due to increase in demand for inputs.
- The term pure monopoly signifies an absolute power to produce and sell a product which has no close substitute.
- Monopoly results in an optimal output, and causes loss of social welfare.
- According to Pigou, there are three degrees of price-discrimination practised by the monopolists: (i) first degree price discrimination; (ii) second degree price discrimination; and (iii) third degree price discrimination.
- Short-term analysis of pricing and output determination under monopolistic competition is similar to price and output determination under monopoly.
- Under Cournot's model, each seller assumes his rival's output to remain constant, whereas under Bertrand's model, each seller determines his price on the assumption that his rival's price, rather than his output, remains constant.
- Edgeworth's model follows Bertrand's assumption that each seller assumes his rival's price, instead of his output, to remain constant.
- The kinked-demand curve analysis does not deal with price and output determination. Instead, it seeks to establish that once a price–quantity combination is determined, an oligopoly firm will not find it profitable to change its price even in response to the small changes in the cost of production.
- Collusion between oligopoly firms may take many forms, depending on their relative strength, their objective and legal status of collusion. There are, however, two main types of collusion: (i) cartels and (ii) price leadership.
- The life cycle of a product is generally divided into five stages: (i) introduction or initial stage, (ii) growth, (iii) maturity, (iv) saturation, and (v) decline.
- Economists have recognized the complexity of business world and have explained how firms formulate their strategic pricing policy and determine the price of their product.
- The pricing strategy varies from stage to stage over the life cycle of a product, depending on the market conditions. From the pricing strategy point of view, growth and maturity stages may be treated likewise.

- A kind of pricing strategy in which a firm is required to quote its price under uncertain cost and price conditions, with a view to winning a contract or a tender is known as competitive bidding or contract pricing.
- Large-size firms, very often divide their operation into product divisions or subsidiaries. Growing firms add new divisions or departments to the existing ones. Such firms face the problems of determining an appropriate price for the product transferred from one division or subsidiary to the other. Pricing of intra-firm transfer product is referred to as transfer pricing.

3.10 KEY TERMS

- **Market Structure:** It refers to the characteristics of market organizations that determine the behaviour of companies in an industry.
- **Perfect Competition:** A market situation wherein several producers offer a homogeneous product to several buyers, and each firm competes with so many other firms that competition is virtually absent.
- **Imperfect Competition:** A market situation wherein several firms sell identical or differentiated products with some control over the price of their product.
- **Monopoly:** It is a market situation where there is a single seller who has control over his price and output.
- **Cartel:** An association of manufacturers or suppliers aimed at maintaining prices at a high level and restricting competition.
- **Competitive bidding:** It is a process of quoting a contract price for supplying goods or services under specified terms and conditions.

3.11 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Differentiate between monopoly and oligopoly.
2. When does a firm achieve the state of equilibrium? Show with the help of a graph.
3. Write the characteristic features of monopolistic competition.
4. What are the conditions when the price of a product is below the prevailing market price of the substitutes?
5. List the major factors in competitive bidding.

Long-Answer Questions

1. Discuss various pricing methods under competitive environment.
2. 'A firm even can make losses in short-run.' Justify the statement.
3. Analyse the major sources of barriers to entry to a monopolised market.

NOTES

4. What is the role of the relationship between AR and MR in price and output determination under monopoly?
5. Discuss the structure of the market under oligopoly. What are the problems posed by such a structure in the theory of oligopoly?
6. Explain the process of transfer pricing with the help of an example.
7. Describe Cournot's duopoly model and apply his solution to oligopoly. Does it provide a stable solution?

3.12 FURTHER READING

Dwivedi, D.N. 2015. *Managerial Economics,* Eighth Edition. New Delhi: Vikas Publishing.

Keat, Paul G. and Philip, K.Y. Young. 2003. *Managerial Economics: Economics Tools for Today's Decision Makers,* Fourth Edition. Singapore: Pearson Education.

Keating, B. and J.H. Wilson. 2003. *Managerial Economics: An Economic Foundation for Business Decisions,* Second Edition. New Delhi: Biztantra.

Peterson, H.C. and W.C. Lewis. 1999. *Managerial Economics*, Fourth Edition. Singapore: Pearson Education.

Salvantore, Dominick. 2001. *Managerial Economics in a Global Economy*, Fourth Edition. Australia: Thomson-South Western.

Thomas, Christopher R. and Maurice S. Charles. 2005. *Managerial Economics: Concepts and Applications*, Eight Edition. New Delhi: Tata McGraw-Hill.

UNIT 4 INVESTMENT DECISIONS AND FIRM OBJECTIVES

NOTES

Structure

4.0 INTRODUCTION

Nowadays, every organization focuses on optimal investment allocation in the current environment. These investment decisions have a substantial impact on the company's value since they influence profitability, growth, and risk. The investment decision refers to the decision taken by investors or senior management regarding the amount of money to be invested in various investment opportunities. These decisions involve allocating cash to various investment options in order to obtain the best potential return. It simply aids businesses in selecting the appropriate assets to deploy their capital. These choices are made by the investor or top-level managers, who thoroughly examine each prospect before investing any funds. Investment decisions are critical for any business since they affect its profitability. It should be assured that a thorough risk and return analysis is carried out.

Such investment decisions involve a high amount of risk as they are taken on the basis of estimated return. Large funds are invested for earning income in future which is completely unpredictable. These returns fluctuate with the changes in fashion, taste, research and technological advancement thereby leading to a greater risk. The main and primary objective of an organisation is profit maximisation. A firm prefers to produce at the point where it can earn maximum profits. However, in practice it is not possible for a firm to make maximum of the profit in the long run due to various constraints, such as time, assets, resources, knowledge, risk tolerance and interests of shareholders. This unit will discuss various investment decisions taken by a firm. In addition, it will explain the objectives of a firm along with the constraints it faces.

NOTES

4.1 OBJECTIVES

After going through this unit, you will be able to:

- Discuss the methods of making investment decisions under condition of risk and uncertainty
- Explain different objectives and constraints of a firm

4.2 INVESTMENT DECISIONS

Closely related to concept of investment decisions are the elements of risk and uncertainty

Our discussion so far has been concerned with the managerial problems of decision-making pertaining to demand and supply, production, cost of production, market structure, nature of competition, and price determination under the assumption that the stock of capital is given. In this part of the book, we will discuss the issue related to investment and enhancing the stock of capital. *Investment is an activity of spending resources (money, labour and time) on creating assets that can generate income over a long period of time or which enhances the returns on the existing assets.*

Investment decisions can be studies under three conditions, under certainty, under risk and under uncertainty. In real life, a large area of investment decisions falls in the realm of *risk and uncertainty*. It is important to note here that risk and uncertainty go hand in hand. Wherever there is uncertainty, there is risk. The probability of some kinds of risk is calculable whereas that of some other kinds of risk is not. The calculable risk like accident, fire, theft, etc. are insurable. Therefore, decision-making in case of insurable risks is a relatively easier task. But, incalculable risks are not insurable. Therefore, investment decision is greatly complicated under the condition of uncertainty, i.e., in case the probability of an outcome is not estimable and predictable. However, some useful techniques have been devised and developed by the economists, statisticians and management experts to facilitate investment decision-making under the conditions of risk and uncertainty. Also, there are several techniques and methods that are applied under different business conditions and for evaluating investment projects. In this section, we will concentrate only on the widely used methods of investment decision-making under (*i*) under risk, and (*ii*) under uncertainty.

The concept of risk and uncertainty can be better explained and understood in contrast to the concept of certainty. Here, let us have a closer look at the concept of certainty and then proceed to explain the concepts of risk and uncertainty. *Certainty* is the state of perfect knowledge about the market conditions. In the state of certainty, there is only one rate of return on the investment and that rate is known to the investors. That is, in the state of certainty, the investors are fully aware of the outcome of their investment decisions. For example, if you deposit your savings in 'fixed deposit' bearing 10 per cent interest, you know for certain that the return on your investment in time deposit is 10 per cent, and FDR can be converted into cash any day at a lower rate of interest, of course. Or, if you

buy government bonds, treasury bills, etc. bearing an interest of 11 per cent, you know for sure that the return on your investment is 11 per cent per annum, your principal remaining safe. In either case your are sure that there is little or no possibility of the bank or the government defaulting on interest payment or on refunding the money. This is called the *state of certainty*.

However, there is a vast area of investment avenues in which the outcome of investment decisions is not precisely known. The investors do not know precisely or cannot predict accurately the possible return on their investment. Some examples will make the point clear. Suppose a firm invests in R&D to innovate a new product, spends money on its production and sale. The success of the product in a competitive market and the return on investment in R&D and in production and sale of the product can hardly be predicted accurately. There is, therefore, an element of uncertainty. Consider another example. Suppose a company doubles its expenditure on advertisement of its product with a view to increasing its sales. Whether sales will definitely increase proportionately can hardly be forecast with a high degree of certainty, for it depends on a number of unpredictable conditions. Consider yet another example. Maruti Udyog Limited (*MUL*) decided in July 2000 to invest money in financing the sale of its own cars with a view to preventing the downslide in its sales which it had experienced over the past two years. However, the managers of *MUL* could hardly claim knowledge of or predict the outcome of this decision accurately. So this decision involves risk and uncertainty. In real life situations, in fact, a large number of business decisions are taken under the conditions of risk and uncertainty, i.e., the lack of precise knowledge about the outcome of the business decisions. Let us now look into the precise meaning of the terms risk and uncertainty in business decisions.

- **Risk:** In common parlance, risk means a low probability of an expected outcome. From business decision-making point of view, risk refers to a situation in which a business decision is expected to yield more than one outcome and the probability of each outcome is known to the decision makers or can be reliably estimated. For example, if a company doubles its advertisement expenditure, there are three probable outcomes: (*i*) its sales may more-than-double, (*ii*) it may just double, or (*iii*) may less than double. The company has the knowledge of these probabilities or has estimated the probabilities of the three outcomes on the basis of its past experience as (*i*) more-than double – 10 per cent (or 0.1), (*ii*) almost double – 40 per cent (or 0.4), and (*iii*) less-than double – 50 per cent (or 0.5). It means that there is 90 per cent risk in expecting more-than-doubling the sales, and there is 60 per cent risk in doubling the sale, the risk is 60 per cent, and so on.

  There are two approaches to estimate probabilities of outcomes of a business decision, viz., (*i*) *a priori approach*, i.e., the approach based on deductive logic or intuition, and (*ii*) *posteriori approach*, i.e., estimating the probability statistically in the basis of the past data. In case of *a priori probability*, we know that when a coin is tossed, the probabilities of 'head' and 'tail' are 50:50, and when a dice is

NOTES

thrown, each side has 1/6 chance to be on the top. The *posteriori* assumes that the probability of an event in the past will hold in future also. The probability of outcomes of a decision can be estimated statistically by way of 'standard deviation' and 'coefficient of variation'.

- **Uncertainty:** *Uncertainty refers to a situation in which there is more than one outcome of a business decision and the probability of no outcome is known or can be meaningfully estimated.* The unpredictability of outcome may be due to lack of reliable market information, inadequate past experience, and high volatility of the market conditions. For example, if an Indian firm, highly concerned with population burden on the country, invents an irreversible sterility drug, the outcome regarding its success is completely unpredictable. Consider the case of insurance companies. It is possible for them to predict fairly accurately the probability of death rate of insured people, accident rate of cars and other automobiles, fire accident rate of houses, and so on, but it is not possible to predict the death of insured individual, a car meeting an accident or a house catching fire, etc.

The long-term investment decisions involve a great deal of uncertainty with unpredictable outcome. But, in reality, investment decisions involving uncertainty have to be taken on the basis of whatever information can be collected, generated and guesstimated. For the purpose of decision making, the uncertainty is classified as follows:

(*a*) Complete ignorance

(*b*) Partial ignorance

In case of *complete ignorance*, investment decisions are taken by the investors using their own judgement or using any of the rational criteria. What criterion he chooses depends on his attitude towards risk. The investor's attitude towards risk may be that of (*i*) a risk averter, (*ii*) a risk neutral, or (*iii*) a risk seeker or risk lover. In simple words, a risk averter avoids investment in high-risk business. A risk-neutral investor takes the best possible decision on the basis of his judgement, understanding of the situation and his past experience. He does his best and leaves the rest to the market. A risk lover is one who goes by the dictum that the higher the risk, the higher the gain. Unlike other categories of investors, he prefers investment in risky business with high expected gains.

In case of *partial ignorance*, on the other hand, there is some knowledge about the future market conditions; some information can be obtained from the experts in the field, and some probability estimates can be made. The available information may be incomplete and unreliable. Under this condition the decision makers use their subjective judgement to assign an *a priori probability* to the outcome or the pay-off of each possible action such that *the sum of such probability distribution is always equal to one*. This is called subjective probability distribution. The investment decisions are taken in this case on the basis of the *subjective probability distribution*. Decision-making process under partial ignorance is described in the following section beginning with the pay-off matrix.

4.2.1 Investment Decisions under Risk

NOTES

The methods of making investment decisions under the condition of risk are as follows:

(i) The pay-off matrix method

The pay-off matrix is also known as decision matrix used in game theory. The technique of pay-off matrix was originally developed as a tool of game theory applied to decision-making under oligopoly. However, it is widely used in general business decision-making and also in investment decisions. For our purpose, here, pay-off matrix can be defined as a tabular array of strategic actions and their corresponding pay-offs under different states of nautre. It is a simple technique of investment decision-making under the condition of uncertainty. The process of making pay-off matrix and deriving conclusion is described here briefly. The decision makers using this method are required (*i*) to state the objective of decision-making, (*ii*) to decide on the possible *strategic actions* under different *states of nature* of the economy, and (*iii*) to assign a pay-off to each strategy under each state of nature. When this set of information is arranged in a tabular form, it gives a matrix, called the *pay-off matrix*. Before we proceed, a brief explanation of the terms—*strategy, states of nature* and *pay-offs*—is in order.

Strategy means one of several alternative actions that can be taken to achieve a certain goal. For example, given the profit maximization objective, the firm may conceive of such actions as (*i*) reducing cost of production through scale expansion, (*ii*) shifting upward the demand curve for its product through aggressive advertisement, (*iii*) product diversification, (*iv*) product innovation, (*v*) cutting down the price with a view to expanding demand for its product, and (*vi*) change in production technology. Each of these strategic actions is expected to affect the total profit to a different extent depending on the 'states of nature'.

The *states of nature* refer to the future market conditions in the long run on which the firm has no control. These conditions might arise on their own or may be the result of actions taken by the firm. The conditions that arise on their own are like depression, recession, boom, or stable growth. The conditions that might arise due to the actions taken by the firm include all possible reactions of the rival firms. A firm is not supposed to have control over the *states of nature* but the states of nature do affect the outcome or the pay-off of each action taken by the firm. The states of nature affect the pay-off of each action in different ways and to different extents. The decision makers are, therefore, required to estimate the pay-off of the strategic actions under different states of nature.

Estimating objectively the pay-off of different strategies even in the state of partial ignorance is an impossible task. Therefore, the decision makers use their subjective judgment, intuition and experience to assign a certain value to the pay-offs of each strategy associated with each state of nature of the economy. The arrangement of such pay-offs in a tabular form produces a matrix, called 'Pay-off Matrix' or 'decision matrix'. A hypothetical pay-off matrix is given in Table 4.1. It assumes that an investor conceives of four strategic investment projects, S_1, S_2, S_3 and S_4 (e.g., investment in plant and machinery, equity, real estate, and government bonds, respectively) under four different states of nature

NOTES

of the economy, N_1, N_2, N_3 and N_4 (e.g., high growth, low growth, stagnation and recession, respectively). The expected pay-offs of these strategies under the different states of nature are given in Table 4.1.

Table 4.1 *The Pay-off Matrix*

	States of Nature			
Strategy	N_1	N_2	N_3	N_4
S_1	20	12	6	5
S_2	15	16	4	–2
S_3	16	8	6	–1
S_4	5	12	3	2

The *second step* in decision making is to assign a probability to each of the strategies under each state of nature. As already noted, under the condition of uncertainty, even with partial ignorance, the probability of any of these possibilities is not known and nor can be meaningfully estimated. It is here that the decision-makers use their subjective judgment in assigning a numerical value to the probability of each expected pay-off. In practice, the decision makers follow the Bayesian rule. This rule says that where meaningful estimate of probabilities is not available, the outcome of each strategy under each state of nature must be assigned the same probability and that the sum of probabilities of outcome of each strategy must be equal to 1. For example, if probabilities P_1, P_2, P_3 and P_4 are assigned to four possible pay-offs under four states of nature, then $P_1 + P_2 + P_3 + P_4 = 1$

If the number of states of nature and outcomes, very very large, there will be a large number of probabilities. Then, according to the Bayesian rule, $P_1 + P_2 + P_3 + \ldots + P_n = 1$

The *third step* is to find the *expected value*. The *expected value* is the weighted mean of the pay-offs or returns (R). In computing the expected value, probabilities are used as the weights. The formula for expected value $E(R)$ is given below.

$$E(R) = R_1 P_1 + R_2 P_2 + R_3 P_3 + \ldots + R_n P_n$$

$$= \sum_{i=1}^{n} R_i P_i$$

where $E(R)$ = expected value or weighted mean of pay-offs; R_1, R_2 ... R_n are expected pay-offs; P_1, P_2 ... P_n are the probabilities.

Let us suppose that subjective probability distribution corresponding to the pay-offs given in Table 4.1 is given in Table 4.2.

Table 4.2 *Subjective Probability Distribution*

		States of Nature			
(a)	*Strategy*	N_1	N_2	N_3	N_4
	S_1	0.50	0.30	0.20	0.00
	S_2	0.60	0.20	0.10	0.10
	S_3	0.40	0.40	0.15	0.05
	S_4	0.55	0.35	0.10	0.00

Given the subjective probability distribution, the expected value of each strategy under different states of nature can be obtained, as shown in Table 4.2(b):

(b) Strategy	Pay-off multiplied by probability	= Expected value
S_1	20 (0.50) + 12 (0.30) + 6 (0.20) + 5 (0.00)	= 14.80
S_2	15 (0.60) + 19 (0.30) + 4 (0.10) – 2 (0.10)	= 12.20
S_3	16 (0.40) + 8 (0.20) + 6 (0.10) – 1 (0.05)	= 10.79
S_4	5 (0.55) + 12 (0.35) + 3 (0.10) + 2 (0.00)	= 7.25

Investment decision: It can be seen from this calculation that strategy S_1 gives the highest possible expected value of the alternative strategic investment options available to the decision makers. The investment decision falls clearly in favour of strategy S_1. However, in case two or more strategies yield the same expected value, then the decision makers may use their discretion or use some other method of making investment decision.

(ii) Risk-adjusted discount rate method

The use of *risk-adjusted discount* rate is one of the methods used for *adjusting the valuation models for risk.* A common, though crude, method of accounting for the risk factor in investment decisions is to use a *risk-adjusted discount rate* in the formula for calculating the *net present value* (*NPV*). The risk-adjusted discount rate (*d*) is defined as

$$d = \frac{1}{1+r+\mu}$$

where r is risk-free discount rate, and μ denotes the risk probability.

The formula for risk-adjusted discount rate can be used to obtain a risk-adjusted or risk-free present value (*PV*) of a return (*R*) expected in future by multiplying the expected return by *d*. For example, risk-adjusted present value of a return (R_5) expected 5 years hence can be obtained as

$$PV = \frac{1}{(1+r+\mu)^5} \; R_5$$

In the *risk-adjusted discount rate approach* to investment decision, the risk-adjusted discount rate (*d*) as defined above, is used to calculated the *NPV*. The formula for calculating the *risk adjusted NPV* for *nth* year is given as

$$\text{Risk-adjusted } NPV = \sum_{t=1}^{n} \frac{R_n}{d^n} - C_0$$

where R_n = return in the *nth* year and C_0 = *original* cost of capital.

Or, by substitution,

$$\text{Risk-adjusted } NPV = \sum_{t=1}^{n} \frac{R_n}{(1+r+\mu)^n} - C_0 \qquad ...(4.1)$$

Given the formula for computing risk-adjusted *NPV* is Eq. (4.1), a risk-adjusted *NPV* can be worked out, if the parameters of the formula are known. For example, suppose an investment project costing ₹ 100 million is expected to yield a cash flow or return of ₹ 132 million after one year. Assuming a

NOTES

discount rate of 8 per cent (or 0.08) and risk probability 0.12, the risk-adjusted *NPV* can be easily calculated as shown below.

$$\text{Risk-adjusted } NPV = \frac{132}{1+0.08+0.12} - 100$$

$$= ₹ 10 \text{ million}$$

This calculation shows that the project gives a risk-free *NPV* = ₹ 10 million, which is, incidentally, a 10 per cent risk-free return. If we work out *NPV* without the risk-factor of 0.12, we find that the *NPV* comes to ₹ 22.22 million. Note that *NPV* > 0 by both the criteria—the *NPV* and the *risk-adjusted discount rate criteria*. The project is, therefore, acceptable on both the criteria.

Properties of the risk-Adjusted discount rate approach

The risk-adjusted discount rate approach has the following useful properties:

First, the discount rate can be adjusted for the varying degree of risk in different years simply by changing the risk factor *u* in the formula.

Second, a higher risk factor in remote future is automatically accounted for since the denominator in the formula for the risk-adjusted rate is raised automatically to a high power. For example, the risk-adjusted discounting rate for the 100th year will be

$$d = \frac{d}{(1+r+\mu)^{100}}$$

This method reduces the *NPV* of a sum expected after 100 years to zero.

Third, this method is extremely straight forward and easy to handle.

However, a major weakness of this method is that it does not include any method to estimate the risk factor. Therefore, this method has to be supplement with a technique of estimating the risk-factor. Or else, the decision-makers will have to use their own intuitive discretion.

(iii) Certainty-equivalent approach

The certainty-equivalent approach is another method of *adjusting valuation models for risk*. This method is similar to the risk-adjusted discount rate approach. Both the methods use the *NPV* formula for evaluating a project. There is, however, a procedural difference between the two methods. In the *risk-adjusted discount rate*, it is the rate of discount (*r*) which is adjusted for accounting for the risk. In the *certainty-equivalent approach*, on the other hand, it is the expected return (*R*) which is adjusted for the risk factor. Procedurally, while the *risk-adjusted discount rate approach* makes risk-adjustment in the denominator of the *NPV* formula, the *certainty-equivalent approach* makes risk-adjustment in its numerator. The formula for *certainty-equivalent approach* is given as

$$\text{Certainty-equivalent } NPV = \frac{\alpha R_n}{(1+r)^n} - C_0 \qquad ...(4.2)$$

where *a* is the risk-equivalent factor.

In fact, α in Eq. (4.2) is called *certainty–equivalent coefficient*. The use of the coefficient (α) makes an expected risky-return (R_n^*) equivalent to an expected certain return (R_n). The *certainty-equivalent coefficient* (α) is worked out as the ratio of R_n to R_n^*. That is,

$$\alpha = \frac{\text{Expected certain return}}{\text{Expected risky return}} = \frac{R_n}{R_n^*} \qquad ...(4.3)$$

For example, suppose an investor has two projects, A and B to choose from. Project A is risky and is expected to yield a sum of ₹ 50 million after a year. Project B is risk-free and is expected to yield a sum of ₹ 40 million after one year. And the investor treats both the projects equally good. It means that he treats a risk-free ₹ 40 million as an equivalent of risky ₹ 50 million. In that case,

$$\alpha = \frac{\text{Rs 40 mm}}{\text{Rs 50 mm}} = 0.8$$

The value of α varies between 0 and 1, and degree of risk moves in the reverse direction, *i.e.*, the smaller the value of α, the greater the degree of risk. For example, if $\alpha = 0$, it means that the expected sum is totally risky or the degree of risk is 100 per cent. And, if $\alpha = 1$, it means that the degree of risk is 0 and expected return is 100 per cent certain.

The use of *certainty-equivalent approach* can be illustrated by a hypothetical example. Suppose (*i*) an investor is considering two projects A and B, costing ₹ 30 million and ₹ 28 million, respectively, (*ii*) Project A is expected to yield a risk-free sum of ₹ 44 million after two years, and B a risky sum of ₹ 55 million, (*iii*) the risk-free discount rate is 10 per cent, and (*iv*) certainty-equivalent coefficient (α) = 0.8. By certainty-equivalent approach, the choice between the two projects can be made by comparing their certainty-equivalent *NPV*.

Project A

$$\text{Certainty-equivalent } NPV = \frac{R_n}{(1+r)} - C_0$$

$$= \frac{44}{(1+0.1)} - 30 = ₹\ 10 \text{ million}$$

Project B

$$\text{Certainty-equivalent } NPV = \frac{\alpha R_n}{(1+r)^n} - C_0$$

$$= \frac{(0.8)55}{(1+0.1)} - 28 = ₹\ 12 \text{ million}$$

Our calculations show that certainty-equivalent *NPV* of Project A equals ₹ 10 million and that of Project B equals ₹ 12 million. Obviously, Project B is preferable to project A by *certainty-equivalent approach.*

(iv) The probability theory approach

In contrast with the methods discussed above, the *probability theory method* considers a whole range of possible risk-return combinations which represent adequately the full range of alternative outcomes of a risky undertaking.

NOTES

Therefore, it is necessary to consider a large number of alternative cash flows and the associated probabilities, for each time period under consideration. The process is illustrated in Table. 4.3.

Table 4.3 *Alternative Cash-flows and Associated Probabilities*

Cash flows	*1st Year*		*2nd Year*		*3rd Year*	
(₹)	*Probability*	*Expected Returns*	*Probability*	*Expected Returns*	*Probability*	*Expected Returns*
1000	0.50	500	0.50	500	0.70	700
2000	0.25	500	0.40	800	0.30	600
3000	0.15	450	0.10	300	0.00	—
4000	0.10	400	0.00	—	0.00	—
Total	1.00	1850	1.00	1600	1.00	1300

As the table shows, given the alternative cash-flows and the associated probabilities, the expected returns with probability 1 are ₹ 1850 in the 1st year, ₹ 1600 in the 2nd year, and ₹ 1300 in the 3rd year. These expected returns may now be discounted to their present value. The rest of the decision procedure is the same as in case of discounted present value criterion.

Limitations

The probability theory method has a limited application. It can be applied only if a large number of similar investments are to be undertaken and if one project fails to yield the expected return, others faring so well that the loss is more than compensated. This method is more appropriate for insurance policies. Among the insurance policy-holders only a few die before the terminal date of insurance policy. Only in these cases, the insurance company suffers a loss. But, since most policy-holders survive till the maturity date of the policy, the insurance company makes profits which more than compensate for the loss caused by the death of some policy-holders before the policy matures.

Under portfolio selection probability theory approach

Despite its limitations, the probability theory method has a great deal of application to the problem of portfolio selection of securities, i.e., in determining the optimum combination of stocks, bonds, government securities and other financial instruments. The application of this method to the problem of portfolio selection is an exemplary case. In his approach, Morkowitz has utilised two focal measures: (*i*) an index of expected returns; and (*ii*) an index of risk. The index of expected returns can be constructed by computing the average expected earning. The index of risk is constructed on the basis of standard deviation of the expected earning. Then a series of combinations of risk and return can be formed for different kinds of financial instruments. These combinations when graphed give a risk-return probability curve, showing the various rates of return and the associated risk, as shown by the curve *RR′* in Figure 4.1. This curve shows the various combinations of return rates and the associated risk.

NOTES

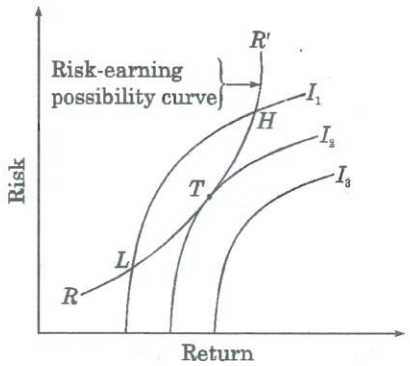


Fig. 4.1 Risk-Earning Combination

The curve *RR'* called the *risk-earning possibility curve*. The various points on the curve *RR'* represent, in a sense, the optimal combinations of risk and return : the lower the risk, the lower the return and the higher the risk, the higher the return. As such, *a priori* choice of a particular point will be arbitrary. Those who prefer lower risk and lower income would choose a point on the lower part of the curve, e.g., point *L*, and those who prefer higher risk and higher return would choose point *H*. This pattern of portfolio selection does not provide an optimal combination of stocks and securities.

The optimal combination of portfolio can, however, be determined by superimposing a *risk-return indifference map* on the same figure, as shown by curves I_1, I_2 and I_3. It is not impossible, at least in principle, to draw risk-return indifference curves for the prospective investors. The risk-return indifference curves would, of course, be inverted as compared with standard indifference curves used in consumer's analysis or isoquant curves used in the analysis of optimal input-combination. Contrary to the indifference curves and isoquants, the risk return indifference curves (I_1, I_2 and I_3) have a positive slope because as the risk increases, a relatively higher rate of return must be associated with it to keep the investor indifferent between the lower and higher risk-return combinations.

Given the properties of the risk-earning possibility curve and the risk-return indifference curves, one of the latter is bound to be tangent to the former. Such as it is, the *optimal combination of portfolio* is decided by the point of tangency between the risk-return possibility curve and the risk-return indifference curves. The optimal combination is shown by point *T* in Figure 4.1. This proposition is, however, only a theoretical solution to the problem of portfolio selection. In practice, it depends to a great extent on the attitude of the investors towards the risk.

(v) Decision tree method

Under the conditions of risk, the decision makers often confront a situation in which they visualize several options available to them, each leading to different probable outcomes under the different states of nature. There are two specific problems in this kind of decision-making.

NOTES

First, the decision makers are required to make a choice (or series of choices) from the alternative investment avenues available to them. They are not supposed to leave the matter undecided.

Second, the decision makers know for sure that all the decisions will yield a positive outcome, but they cannot tell in advance the exact outcome of a decision. They might be knowing that a particular decision will yield a higher return than another but they do not know for sure how high or low the outcome will be.

The question that decision makers face under these conditions is how to find the most profitable or gainful solution. The method that is used to find an acceptable solution under these conditions is called decision tree. *A decision tree is a graphical device to map all possible managerial decisions in a sequence and their expected outcomes under different states of the economy.* Since all possible strategic decisions and their possible outcomes are arranged graphically in the form of branches of a tree, the technique is called "*decision tree*". The decision tree, as shown in Figure 4.2, presents the entire decision options and possible outcomes in the form of a diagram and, thereby, guides the decision makers to a rational decision.

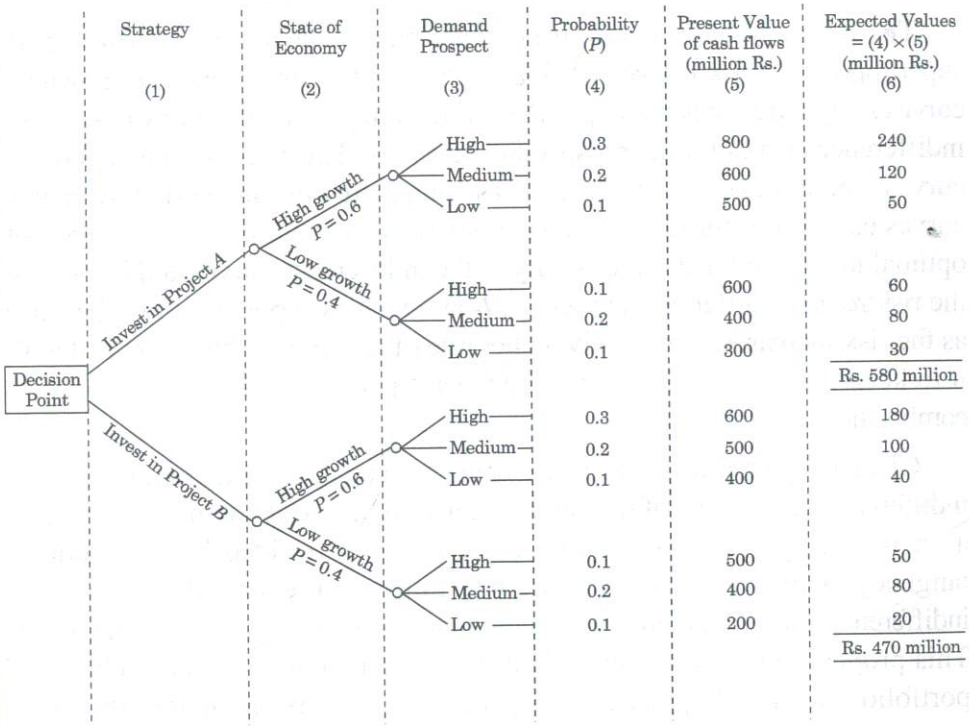


Fig. 4.2 *Decision Tree for Investment Decision*

The decision tree method is illustrated by a simple hypothetical example in Figure 4.2. Suppose an investor visualizes two viable Projects A and B. Project A costs ₹ 500 million and Project B costs ₹ 400 million. The prospective yields (cash flows) of the two projects depend on the states of the economy, i.e., whether the economy has a high or a low growth. The probable growth rates of the economy determine the demand prospects for the product of each project. Demand prospect may be high, medium and low, depending on the

market conditions and consumers perception and preference. The demand conditions determine the expected returns, the cash flows from both the projects. The investor has the information about the expected cash flows under the different states of economy and the demand prospects. All the possibilities are presented graphically in Figure 4.2.

The decision shows that the process of decision-making begins at the 'decision point'. The decision maker has to make a strategic choice between the two Projects A and B. Column (2) shows the probabilities of the 'states of economy'—the economy may grow at a high or at a low growth rate. High growth has a probability of 0.6 and low growth has a lower probability of 0.4 Under both these growth probabilities, the prospect for the product demand again has three probabilities—high, medium and low—under both high and low growth of the economy. Note that probability distribution in respect of demand prospects—high, medium and low—under high growth add up to 0.6 and in case of low growth they add up to 0.4.

Let us now suppose that the investor has the information on the *present value* of cash flows under each probability as presented in Col. (5). When the present value of cash flow is multiplied by the corresponding probability in Col. (4), it gives the 'expected value' which is the same as the 'certainty-equivalent' of the present value of cash flows. Col. (6) gives the 'expected value' of the two projects under all the stipulated conditions. The investor has now the full information for decision making.

Investment Decision

Given the information in Figure 4.2 decision making becomes an easy task. The investor can easily find out the *net expected value* of each project and decide in favour of the project having a higher net expected value, as shown below.

Project A Total expected value = ₹ 580 million
Less project cost = ₹ 500 million
Net expected value = ₹ 80 million

Project B Total expected value = ₹ 470 million
Less project cost = ₹ 400 million
Net expected value = ₹ 70 million

This calculation shows that the net expected value of Project A is higher than that of Project B. A rational investor would, therefore, decide to invest in Project A and shelve Project B.

(vi) Simulation

You have already studied so far the methods of making investment decisions under a limited probability of risk. In real life situations, business decision-making is much more complex. A large number of unpredictable factors has to be taken into account, especially where investment involves a huge sum and the project has very long life. Suppose Daewoo, a car-manufacturing company,

NOTES

encouraged by the sale of its small car model Matiz, plans to set up another production plant. The number of considerations that figure in its decision making process are numerous including size and cost of the plant, life of the plant, future demand prospects, market saturation point, competition by the rival firms, optimum level of production, probable price, revenue, input prices (labour and material costs), expected rate of return, labour problem, selling costs, prospects for increasing price, price elasticity, profitable market share, government policy, sales tax, entry of new competing brands of car, entry of foreign cars in the Indian market, export prospects, and so on. The decision-makers are supposed to be equipped with all the requisite data and related information. The technique that is used to facilitate decision making under this kind of situation is called simulation.

Simulation is a mathematical technique used to produce alternative target variable under certain stipulated conditions. The first step in *simulation technique* is to build a mathematical model, called *simulation model*. The second step is to find the *probability distributions* of each factor involved in decision making. Probability distributions can be generated by company's own market research department, economic research centres and published sources. The necessary data for constructing probability distribution can also be obtained from government publications, media sources and private research institutions. Once these tasks are completed, a computer programme is used to simulate the target variable(s), e.g., the rate of return, profit or cash flows, from a combination of randomly selected variables and their probabilities. In this process, the computer picks up a variable randomly and one element from its probability distribution and determines the rate of return. This process is repeated until all the selected variables and their probabilities are exhausted. Each time a different probable target variable (the rate of return or profit) is produced. This process is called iteration. It tells, if this, this, this,..., then what will be the return. In this process, the computer builds a frequency distribution of the rates of return on different set of conditions. The decision-makers can choose one or many from the options generated by the simulation technique.

Although simulation is a versatile and very powerful technique, a full-scale simulation is very expensive and time-consuming technique. Therefore, this technique can be used only by large corporations where a huge investment is involved and division making is beset with complexities. For a general use point of view, this technique is not feasible.

(vii) Sensitivity Analysis

Sensitivity analysis is a simple version of a full-fledged simulation technique and also inexpensive. While simulation technique uses the whole range of probability distributions of each decision variable, sensitivity analysis uses probabilities in the high range only. Sensitivity analysis begins with the best estimate of the target variable (rate of return, profit or cash flow) associated with the highest probability

of each variable. Then changes in the decision variables are introduced to find their effect on the target variable, say the rate of return. In fact, analysts attempt to find the variables to which the target variable is highly sensitive or responsive. This technique helps the decision makers in eliminating unimportant variables and in concentrating on the most important ones, i.e., the variables to which the rate of return is highly sensitive. This is called *sensitivity analysis*.

4.2.2 Investment Decisions under Uncertainty

In the preceding sections, you have already studied the techniques that are used in investment decision-making under the condition of *risk*. Let us discuss the techniques of investment decisions under *uncertainty*. Uncertainty refers to a situation in which a decision is expected to yield more than one outcome and the probability of none of the possible outcomes is known. Therefore, decisions taken under uncertainty are necessarily *subjective*. However, analysts have devised some decision rules to impart some objectivity to the subjective decisions, provided *decision-makers are able to identify the possible 'states of nature' and can estimate the outcome of each strategy*. Some such important decision rules are discussed below.

(i) Wald's maximin decision criterion

Wald's *maximin decision criterion* tells that the decision-makers should specify first the worst possible outcome of each strategy and accept a strategy that gives best out of the worst outcomes. The application of maximin criterion can be illustrated by applying it to the example given in Table 4.1 reproduced below. To apply the maximin criterion, the decision makers need to find the worst (minimum) outcome of each strategy. This can be done by reading Table 4.4. row-wise. The Maximin column presents the worst outcome of each strategy. The best or the highest outcome of the worst outcome is 5 of strategy S_1. Going by the maximin criterion, the decision-makers would accept strategy S_1.

***Table 4.4** Application of Maximin Criterion*

	States of Nature				
Strategy	N_1	N_2	N_3	N_4	*Maximin*
S_1	20	12	6	5	5
S_2	15	10	4	– 2	– 2
S_3	16	8	6	– 1	– 1
S_4	5	12	3	2	2

If you look closely at the maximin decision rule, it implies a pessimistic approach to investment decision-making. It gives a conservative decision rule for risk avoidance. However, this decision rule can be applied by those investors who fall in the category of risk averters. This investment rule can also be applied by firms whose very survival depends on avoiding losses.

(ii) Minimax regret criterion or savage decision criterion

Minimax regret criterion is another decision rule under uncertainty. This criterion suggests that the decision-makers should select a strategy that minimizes the

maximum regret of a wrong decision. What is regret? 'Regret' is the difference between the pay-off of a given strategy and the pay-off of the best strategy under the same state of nature. Thus, regret is the opportunity cost of a decision. Suppose an investor has three strategies for investment, S_1, S_2 and S_3, giving returns of ₹ 10,000, ₹ 8000 and ₹ 6000, respectively. If the investor opts for strategy S_1, he gets the maximum possible return. He has no regret. But, if he opts for S_2 by way of an incorrect decision, then his regret or opportunity cost equals ₹ 10,000 – ₹ 8000 = ₹ 2000. Similarly, if he opts for S_3, his regret equals ₹ 10,000 – ₹ 6000 = ₹ 4000. Going by the minimax regret criterion, the investor should opt for strategy S_2 because it minimizes the regret.

The application of minimax regret criterion can be illustrated with the help of the example Table 4.1. By using the pay-off matrix, we can construct a *regret matrix*. The method is simple. Select a column (the state of nature), find the maximum pay-off and subtract from it the pay-offs of all strategies. This process gives a pay-off column. For example, under column N_1, strategy S_1 has the maximum pay-off (20). When we subtract 20 from 20, we get 0. It means that if S_1 is chosen under the state of nature N_1, the regret is zero. Next, strategy S_2 has a pay-off 15. When we subtract 15 from 20, we get regret which equals 5. By repeating this process for all the strategies (S_1, S_2, ... S_n) and all the states of nature (N_1, N_2, ... N_n), we get a regret matrix as shown in Table 4.5. From the regret matrix, 'maximin regret' can be determined by listing the maximum regret for each strategy, as shown in the last column. The column 'maximin regret' shows that maximum regret is minimum (3) in case of strategy S_4. According to maximin creterion, therefore, strategy S_4 should be selected for investment.

Table 4.5 *Pay-off Matrix and Regret Matrix*

	States of Nature				*Regret Matrix*				*Maximin*
Strategy	N_1	N_2	N_3	N_4	N_1	N_2	N_3	N_4	*Regret*
S_1	20	12	6	5	0	0	0	0	0
S_2	15	10	4	– 2	5	2	2	8	8
S_3	16	8	6	– 1	4	4	0	6	6
S_4	5	12	3	2	15	0	3	3	3*

(iii) Hurwicz decision criterion

Hurwicz has suggested another criterion for investment decisions under uncertainty. In his opinion, full realization of optimistic pay-off or full realization of most pessimistic pay-off is a rare phenomenon. The actual pay-off of a strategy lies somewhere between the two extreme situations. According to Hurwicz criterion, therefore, the decision makers need to construct a *decision index* of most optimistic and most pessimistic pay-offs of each alternative strategy. The *decision index* is, in fact, a weighted average of maximum possible and minimum possible pay-offs, weight being their subjective probability such that sum of probabilities of maximum (Max) and minimum (Min) pay-offs equals one. Hurwicz formula for *decision index* (D) is given below.

$D_i = \alpha \text{Max}_i + (1 - \alpha)\ \text{Min}_i$ where D_i = decision index of the ith strategy; and α = is probability of maximum pay-off.

The construction of Hurwicz *decision index* is illustrated in Table 4.6. Column (2) presents the maximum possible pay-offs of investment strategies, S_1, S_2, S_3 and S_4 listed in column (1). Column (3) shows the probability of maximum pay-offs. Column (4) gives the weighted pay-offs of the maximum pay-offs of the four strategies. *Weighted pay-off* equals the maximum pay-off multiplied by α (where α is subjective probability a pay-off). Note that the same probability applies to all the strategies. Columns (5), (6) and (7) give similar value of minimum pay-offs of the four strategies. The last colume (8) gives the *decision index.*

Table 4.6 *Hurwicz Decision Index*

State	*Max*	α	α *Max*	*Min*	$(1 - \alpha)$	$(1 - \alpha)$ *min*	*D*
(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*	*(8)*
S_1	10	0.8	8	6	0.2	1.2	9.2
S_2	20	0.8	16	10	0.2	2.0	18.0
S_3	15	0.8	12	5	0.2	1.0	13.0
S_4	12	0.8	9	– 10	0.2	– 1.0	8.0

As regards the *investment decision*, as the table (Col 8) shows, strategy S_2 has the highest decision index (18.0). Therefore, strategy S_2 is preferable to all other strategies.

(iv) Laplace decision criterion

The Laplace criterion uses the Bayesian rule to calculate the *expected value* of each strategy. As mentioned earlier, the Bayesian rule says that where meaningful estimate of probabilities is not available, the outcome of each strategy under each state of nature must be assigned the same probability and that the sum of probabilities of outcome of each strategy must add up to one. For this reason, the Laplace criterion is also called the 'Bayesian criterion'. By assuming equal probability for all events, the environment of 'uncertainty' is converted into an environment of 'risk'.

Once this decision rule is accepted, then decision-makers can apply the decision criteria that are applied under the condition of risk. The most common method used for the purpose is to calculate the 'expected value'. Once expected value of each strategy is worked out, then the strategy with the highest expected value is selected. This decision rule avoids the problem that arises due to subjectivity in assuming a probability of pay-offs. This criterion is, therefore, regarded as the criterion of rationality because it is free from a decision maker's attitude towards risk.

Conclusion

To sum up, uncertainty is an important factor in investment decisions but there is no unique method of dealing with uncertainty. There are several ways of making investment decisions under the condition of uncertainty. None of the methods, as

described above, lead to flawless decision. But they do add some degree of certainty to decision making. The choice of method depends on the availability of necessary data and reliability of a method under different conditions.

NOTES

Decision-making under Certainty

In situations of certainty, all the information is available and at hand. Such data is also easy to attain and not very expensive to gather. As a result, the person in charge of making the decision knows for sure the consequence of each alternative, strategy or course of action to be taken. In these circumstances, it is possible to foresee (if not control) the facts and the results.

Let us see the following example of decision making under certainty. A farmer wants to decide which crop should plant among three crops, on his 10 hectare farm. The payoff from each is dependent on the rainfall during the monsoon seasons. The farmer has categorized the amount of rainfall into high, medium and low. Based on the categorization, the farmer can take a decision on which crop to plant according to the amount of rainfall. Here, the farmer has complete knowledge about outcome therefore he is able to take an effective decision with maximum payoff.

Let us see another example. The manager of a company has kept aside ₹ 1 lakh to cover the renovation of all executive offices. This money is kept in a savings account at a government bank that pays 7.50 percent interest. Half of the money will be drawn out next month and the rest when the job is completed in 90 days.

Can the manager determine today how much interest will be earned on the money over the next 90 days? Given the fact that he knows how much is being invested, the length of investment time, and the interest rate, the answer is yes. Therefore, the investment of money in a government bank is a decision made under conditions of certainty. The outcome in terms of interest is known today. In this condition, the decision-maker knows with reasonable certainty what the alternatives are, what conditions are associated with each alternative? So under this condition, the manager has enough information to known the outcome of the decision before it is made.

Check Your Progress

1. Mention the two approaches to estimate probabilities of outcomes of a business decision.
2. What is a decision tree?
3. What is the first step in the simulation technique?

4.3 FIRM OBJECTIVES AND CONSTRAINTS

Economists argue that objectives of business firms are various and vary in accordance with the nature of the market. The objectives which figure in economic literature prominently are as follows:

(i) Maximization of profit

(ii) Maximization of sales revenue

(iii) Maximization of firms growth rate

(iv) Maximization of managerial utility function

(v) Maximization of firm's net worth

(vi) Making a satisfactory or standard profit

(vii) Long-run survival and a larger market share.

The conventional theory of pricing is based on profit maximization.

4.3.1 Profit Maximization

Conventional theory of firm assumes profit maximization as the sole objective of business firms. Baumol has, however, argued, 'There is no reason to believe that all businessmen pursue the same objective'. Recent researches on this issue reveal that the objectives that business firm pursue are more than one.

All business firms have undoubtedly some organizational goals to pursue. What is the most common objective of business firms? There is no definitive answer to this question. Perhaps the best way to find out the common objective of business firms would be to ask the business executives. However, Baumol, a well-known authority on business economics, has remarked that firms and their executives are often not clear about their objectives. 'In fact, it is common experience when interviewing executives to find that they will agree to every plausible goal about which they are asked.' However, profit maximization is regarded as the most common and theoretically most plausible objective of business firms. In this section, we describe the conditions of profit maximization and the controversy on profit maximization as the objective of business firms.

Profit maximizing conditions

Total profit (Π) is defined as

$$P = TR - TC \qquad ...(4.4)$$

where TR = total revenue and TC = total cost

There are two conditions that must be fulfilled for $TR - TC$ to be maximum. These conditions are called (i) necessary condition, and (ii) secondary or supplementary condition. The two conditions are also called the 'first order' and the 'second order' conditions.

- The necessary or the first order condition requires that for profit to be maximum marginal revenue (MR) must be equal to marginal cost (MC), i.e., $MR = MC$. By definition, marginal revenue is the revenue obtained from the production and sale of one additional unit of output and marginal cost is the cost arising due to the production of one additional unit of output.
- The secondary or the second order condition requires that the necessary condition must be satisfied under the condition of decreasing MR and rising MC. The fulfillment of the two conditions makes the *sufficient condition.* This condition is illustrated by P_2 in Figure 4.3.

The profit maximizing conditions can be presented in technical terms as follows. As noted above, profit maximizing firm seeks to maximize $\Pi = TR -$

NOTES

TC. Let us now look at the mathematical logic of the first and second order conditions of profit maximization.

(a) **First-order condition**

The first-order condition of maximizing a function is that its first derivative must be equal to zero. Thus, the first-order condition of profit maximization is that the first derivative of the profit function Eq. (4.4) must be equal to zero. Differentiating the total profit function and setting it equal to zero, we get

$$\frac{\partial \Pi}{\partial Q} = \frac{\partial TR}{\partial Q} - \frac{\partial TC}{\partial Q} = 0 \qquad ...(4.5)$$

This condition holds only when

$$\text{or} \quad \frac{\partial TR}{\partial Q} = \frac{\partial TC}{\partial Q}$$

In Eq. (4.5), the term $\partial TR/\partial Q$ gives the slope of the total revenue curve, which is the same as marginal revenue *(MR)*. Similarly, the term $\partial TC/\partial Q$ gives the slope of the total cost curve or what is the same as marginal cost *(MC)*. Thus, the first-order condition for profit maximization can be stated as:

$$MR = MC$$

The first-order condition is generally known as *necessary condition.* A condition is said to be necessary if its non-fulfillment results in non-occurrence of an event.

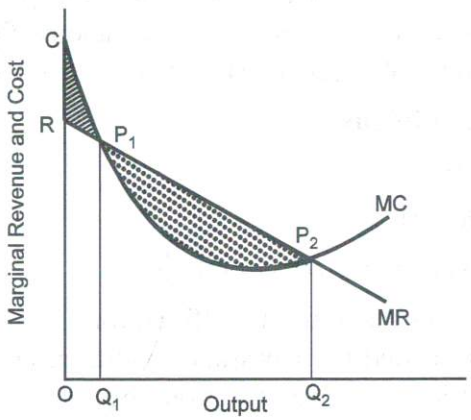


Fig. 4.3 *Marginal Conditions of Profit Maximization*

(b) **Second-order condition**

As already mentioned, in non-technical terms, the second-order condition of profit maximization requires that the first order condition is satisfied under rising *MC* and decreasing *MR*. This condition is illustrated in Figure 4.3. The *MC* and *MR* curves are the usual marginal cost and marginal revenue curves, respectively. *MC* and *MR* curves intersect at two points, P_1 and P_2. Thus, the first order condition is satisfied at both the points, but the second order condition of profit maximization is satisfied only at point P_2. Technically, the second order condition requires that its second derivative of the profit function is negative. When second derivative of profit function is negative, it implies that the total profit curve has turned downward after reaching the peak, i.e., the highest point on profit scale. The second derivative of the total profit function is given as

$$\frac{\partial \Pi}{\partial Q} = \frac{\partial^2 TR}{\partial Q^2} - \frac{\partial^2 TC}{\partial Q^2} \quad ...(4.6)$$

The second-order condition requires that in Eq. (4.6),

$$\frac{\partial^2 TR}{\partial Q^2} - \frac{\partial^2 TC}{\partial Q^2} < 0$$

or $$\frac{\partial^2 TR}{\partial Q^2} < \frac{\partial^2 TC}{\partial Q^2} \quad ...(4.7)$$

Since $\partial^2 TR/\partial Q^2$ gives the slope of *MR* and $\partial^2 TC/\partial Q^2$ gives the slope of *MC*, the second-order condition can also be written as:

Slope of *MR* < Slope of *MC*

It implies that *MC* must have a steeper slope than *MR* or *MC* must intersect the *MR* from below. To conclude, profit is maximized where both the first and second order conditions are satisfied.

Example

We can now apply the profit maximization conditions to a hypothetical example and compute profit maximizing output. We know that $TR = P.Q$ and that $Q = f(P)$, i.e., the demand function.

Suppose demand function is given as:

$$Q = 50 - 0.5P \quad ...(4.8)$$

Given the demand function (4.8), price (P) function can be derived as

$$P = 100 - 2Q \quad ...(4.9)$$

Then $TR = (100 - 2Q)\ Q$

or $$TR = 100Q - 2Q^2 \quad ....(4.10)$$

Suppose also that the total cost function is given as:

$$TC = 10 + 0.5Q^2 \quad ....(4.11)$$

Let us now apply the first order condition of profit maximization and find profit maximizing output. We have noted that profit is maximum where

$$MR = MC$$

or $$\frac{\partial TR}{\partial Q} = \frac{\partial TC}{\partial Q}$$

Given the total *TR* function in Eq. (4.10) and *TC* function in Eq. (4.11),

$$MR = \frac{\partial TR}{\partial Q} = 100 - 4Q \quad ...(4.12)$$

and $$MC = \frac{\partial TC}{\partial Q} = Q \quad ...(4.13)$$

Thus, profit is maximum where

$$MR = MC$$

NOTES

NOTES

$$100 - 4Q = Q$$

or $$5Q = 100$$

$$Q = 20$$

The output 20 satisfies the second order condition also. The second order condition requires that

$$\frac{\partial^2 TR}{\partial Q^2} - \frac{\partial^2 TC}{\partial Q^2} < 0$$

In other words, the second-order condition requires that

$$\frac{\partial MR}{\partial Q} - \frac{\partial MC}{\partial Q} < 0 \quad \text{or} \quad \frac{\partial(100 - 4Q)}{\partial Q} - \frac{\partial(Q)}{\partial Q} < 0 - 4 - 1 < 0$$

Obviously, the second-order condition is also satisfied at output 20. Therefore, profit is maximum at $Q = 20$.

4.3.2 Controversy over Profit Maximization Objective: Theory vs. Practice

As discussed above, there is a good deal of controversy on whether business firms seek to maximize their profits. The traditional theory assumes profit maximization as the sole objective of a business firm. In practice, however, firms have been found to be pursuing objectives other than profit maximization. It is argued, in the first place, that the reason for the firms, especially the large corporations, pursuing goals other than profit maximization is the dichotomy between the ownership and management. The separation of management from the ownership gives managers an opportunity and also discretion to set goals for the firms they manage—goals other than profit maximization. Large firms pursue such goals as sales maximization, maximization of managerial utility function, maximization of firm's growth rate, making a target profit, retaining market share, building up the net, worth of the firm, etc.

Second, traditional theory assumes full and perfect knowledge about current market conditions and the future developments in the business environment of the firm. The firm is thus supposed to be fully aware of its demand and cost functions in both short and long runs. Briefly speaking, a complete certainly about the market conditions is assumed. On the contrary, it is widely recognized that market conditions keep changing and firms do not possess the perfect knowledge of their costs, revenue, and their environment. They operate in the world of uncertainty. Most price and output decisions are based on *probabilities.*

Finally, the marginality principle of equalizing *MC* and *MR* has been found to be absent in the decision-making process of the firms. Empirical studies of the pricing behaviour of the firms have shown that the marginal rule of pricing does not stand the test of empirical verification. Hall and Hitch[12] have found, in their study of pricing practices of 38 UK firms that the firms do not pursue the objective of profit maximization and that they do not use the marginal principle of equalizing *MR* and *MC* in their price and

output decisions. Most firms aim at long-run profit maximization. In the short-run, they set the price of their product on the basis of *average cost principle,* so as to cover $AC = AVC + AFC$ and a normal margin of profit (usually 10 per cent).

NOTES

In a similar study, Gordon has found that there is marked deviation in the real business conditions from the assumptions of the traditional theory and that pricing practices were notably different from the marginal theory of pricing. He has concluded that the real business world is much more complex than the one postulated by the theorists. Because of the extreme complexity of the real business world and ever-changing conditions, the past experience of the business firms is of little use in forecasting demand, price and costs. The firms are not aware of their *MR* and *MC*. The *average-cost-principle* of pricing is widely used by the firms. Findings of many other studies of the pricing practices lend support to the view that there is little link between pricing theory and pricing practices.

Defence of profit maximization hypothesis

The arguments against the profit-maximization assumption, however, should not mean that pricing theory has no relevance to the actual pricing policy of the business firms. A section of economists has strongly defended the profit maximization objective and 'marginal principle' of pricing and output decisions. The empirical and theoretical support put forward by them in defence of the marginal rule of pricing may be summed as follows.

In two empirical studies of '110 excellently managed companies', JS Earley has concluded that the firms do apply the marginal rules in their pricing and output decisions. Fritz Maclup has argued in abstract theoretical terms that empirical studies by Hall and Hitch, and Lester do not provide conclusive evidence against the marginal rule and these studies have their own weaknesses. He argues further that there has been a misunderstanding regarding the purpose of traditional theory of value. The traditional theory seeks to explain market mechanism, resource allocation through price mechanism and has a predictive value, rather than deal with specific pricing practices of certain firms. The relevance of marginal rules in actual pricing system of firms could not be established because of lack of communication between the businessmen and the researchers as they use different terminology like *MR, MC* and elasticities. Besides, businessmen even if they do understand economic concepts, would not admit that they are making *abnormal profits* on the basis of marginal rules of pricing. They would instead talk of a 'fair profit'. Also, Maclup is of the opinion that the practices of setting *price* equal to *average variable cost* plus a *profit margin* is not incompatible with the marginal rule of pricing and that the assumptions of traditional theory are plausible.

While the controversy on profit maximization objective remains unresolved, the conventional theorists, the marginalists, continue to defined the profit maximization objective and its marginal rules.

NOTES

Further arguments in defence of profit maximization hypothesis

The conventional economic theorists defend the profit maximization hypothesis also on the following grounds:

1. **Profit is indispensable for firm's survival:** The survival of all the profit-oriented firms in the long run depends on their ability to make a *reasonable* profit depending on the business conditions and the level of competition. The definition of a reasonable profit may be a matter of opinion. But, making a profit is a necessary condition for the survival of the firm. Once the firms are able to make profit, they try to make it as large as possible, i.e., they tend to maximize it.
2. **Achieving other objectives depends on firm's ability to make profit:** Many other objectives of business firms have been cited in economic literature, e.g., maximization of managerial utility function, maximization of long-run growth, maximization of sales revenue, satisfying all the concerned parties, increasing and retaining market share, etc. The achievement of such alternative objectives depends wholly or partly on the primary objective of making profit.
3. **Evidence against profit maximization objective not conclusive:** Profit maximization is a time-honoured objective of business firms. Although this objective has been questioned by many researchers, the evidence against it is not conclusive or unambiguous.
4. **Profit maximization objective has a greater predicting power:** Compared to other business objectives, profit maximization assumption has been found to be a much more powerful premise in predicting certain aspects of firms' behaviour. As Friedman has argued, the validity of the profit-maximization objective cannot be judged by *a priori* logic or by asking business executives, as some economists have done. The ultimate test of its validity is its ability to predict the business behaviour and the business trends.
5. **Profit is a more reliable measure of firm's efficiency:** Thought not perfect, profit is the most efficient and reliable measure of the efficiency of a firm. It is also the source of internal finance. Profit as a source of internal finance assumes a greater significance when financial market is highly volatile. The recent trend shows a growing dependence on the internal finance in the industrially advanced countries. In fact, *in developed countries, internal sources of finance contribute more than three-fourths of total finance.*

If one judges their motivations by their acts, profit maximization appears to be a more valid business objective.

4.3.3 Alternative Objectives of Business Firms

While postulating the objectives of business firms, the conventional theory of firm does not distinguish between owners' and managers' interests. The

recent theories of firm called 'managerial' and 'behavioural' theories of firm, however, assume owners and managers to be separate entities in large corporations with different goals and motivation. Berle and Means were the first to point out the dichotomy between the ownership and the management and its role in managerial behaviour and in setting the goal(s) for the firm that they manage. Later, Galbraith wrote extensively on this issue which is known as Berle-Means-Galbraith (B-M-G) hypothesis. The B-M-G hypothesis states (i) that owner controlled firms have higher profit rates than manager controlled firms; and (ii) that managers have no incentive for profit maximization. The managers of large corporations, instead of maximizing profits, set goals for themselves that can keep the owners quiet so that managers can take care of their own interest in the corporation. Let us discuss briefly some important alternative objectives of business firms, especially of large business corporations.

(i) Sales revenue maximization: Baumol's hypothesis

Baumol has postulated maximization of sales revenue as an alternative to profit-maximization objective. The reason behind this objective he puts forward is the dichotomy between ownership and management in large business corporations. This dichotomy gives managers an opportunity to set their own goals other than profit maximization goal which most owner-businessmen pursue. Given the opportunity, managers choose to maximize their own utility function. According to Baumol, the most plausible factor in managers' utility *functions is maximization of the sales revenue.*

The factors which explain the pursuance of this goal by the managers are as follows:

First, salary and other earnings of managers are more closely related to sales revenue than to profits.

Second, banks and financial corporations look at sales revenue while financing the corporation.

Third, the trend in sales revenue is a readily available indicator of the performance of the firm. It also helps in handling the personnel problem.

Fourth, increasing sales revenue enhances the prestige of managers while profits go to the owners.

Fifth, managers find profit maximization a difficult objective to fulfill consistently over time and at the same level. Profits may fluctuate with changing conditions.

Finally, growing sales strengthen competitive spirit of the firm in the market and vice versa.

(ii) Maximization of firm's growth rate: Marris's hypothesis

According to Robin Marris, managers maximize firm's *balanced growth rate* subject to managerial and financial constraints. He defines a firm's balanced growth rate (G_B) in terms of the following equation:

$$G_B = G_D = G_C$$

NOTES

where G_D = growth rate of demand for firms product and G_C = growth rate of capital supply to the firm.

In simple words, a firm's growth rate is balanced when demand for its product and supply of capital to the firm increase at the *same rate.* The two growth rates are according to Marris, translated into two utility functions: *(i)* manager's utility function, and *(ii)* owner's utility function.

The manager's utility function (U_m) and owner's utility function (U_0) may be specified as follows.

$$U_m = f(\text{salary, power, job security, prestige, status})$$

and $U_0 = f(\text{output, capital, market-share, profit, public esteem}).$

Owners' utility function (U_0) implies growth of demand for firms product, supply of capital. Therefore, maximization of U_0 means maximization of 'demand for firm's product' or growth of capital supply. According to Marris, by maximizing these variables, managers maximise both their own utility function and that of the owners. The managers can do so because most of the variables (e.g., salaries, status, job security and power) appearing in their own utility function and those appearing in the utility function of the owners (e.g., profit, capital market and market share) are positively and strongly correlated with a single variables, i.e., *size of the firm.* Maximization of these variables depends on the maximization of the growth rate of the firms. The managers, therefore, seek to maximize a steady growth rate.

(iii) Maximization of managerial utility function: Williamson's hypothesis

Like Baumol and Marris, Williamson argues that managers have the discretion to pursue objectives other than profit maximization. The managers seek to maximize their own *utility function* subject to a minimum level of profit Managers' utility function (U) is expressed as

$$U = f(S, M, I_D)$$

where S = Additional expenditure on staff,

M = Managerial emoluments, and

I_D = Discretionary investments.

According to Williamson's, managers maximize their utility function subject to a satisfactory profit. A minimum profit is necessary to satisfy the shareholders or else manager's job security is endangered.

The utility functions which managers seek to maximize include both quantifiable variables like salary and slack earnings, and non-quantitive variable such as prestige, power, status, job security and professional excellence. The non-quantifiable variables are expressed, in order to make them operational, in terms of *expense preference* defined as 'satisfaction derived out of certain types of expenditures' (such as slack payments), and availability of funds for discretionary investment.

NOTES

(iv) Cyert-March hypothesis of satisficing behaviour

Cyert-March hypothesis is an extension of Simon's hypothesis of firms' 'satisficing behaviour' or satisfying behaviour. Simon had argued that the real business world is full of uncertainty; accurate and adequate data are not readily available; where data are available managers have little time and ability to process them; and managers work under a number of constraints. Under such conditions it is not possible for the firms to act in terms of rationality postulated under profit maximization hypothesis. Nor do the firm seeks to *maximize* sales, growth or anything else. Instead they seek to achieve a 'satisfactory profit', a 'satisfactory growth', and so on. This behaviour of firms is termed as 'Satisfaction Behaviour'.

Cyert and March added that, apart from dealing with an uncertain business world, managers have to satisfy a variety of groups of people—managerial staff, labour, shareholders, customers, financiers, input suppliers, accountants, lawyers, government authorities, etc. All these groups have their interest in the firms—often conflicting. The manager's responsibility is to 'satisfy' them all. Thus, according to Cyert-March, firm's behaviour is 'satisficing behaviour'. The 'satisficing behaviour' implies satisfying various interest groups by sacrificing firm's interest or objective. The underlying assumption of 'satisficing behaviour' is that a firm is a coalition of different groups connected with various activities of the firms, e.g. shareholders, managers, workers, input supplier, customers, bankers, tax authorities, and so on. All these groups have some kind of expections—high and low—from the firm, and the firm seeks to satisfy all of them in one way or another by sacrificing some of its interest.

In order to reconcile between the conflicting interests and goals, manager form an *aspiration level of the firm* combining the following goals: (a) production goal, (b) sales and market share goals, (c) inventory goal, (d) profit goal.

These goals and 'aspiration level' are set on the basis of the managers' past experience and their assessment of the future market conditions. The 'aspiration levels' are modified and revised on the basis of achievements and changing business environment.

(v) Rothschild's hypothesis of long-run survival and market share goals

Another alternative objective of firms—alternative to profit maximization—was suggested by Rothschild. According to him, the primary goal of the firms is long-run survival. Some other economists have suggested that attainment and retention of a constant market share is an additional objective of the firms. The managers, therefore, seek to secure their market share and long-run survival. The firms may seek to maximize their profit in the long-run though it is not certain.

(vi) Entry-prevention and risk-avoidance

Another related objective of the firms suggested by some economists is to prevent entry of new firms into the industry. The motive behind entry

NOTES

prevention may be (a) profit maximization in the long-run, (b) securing a constant market share, and (c) avoidance of risk caused by the unpredictable behaviour of new firms. The evidence of whether firms maximize profits in the long-run is not conclusive. Some economists argue, however, that where management is divorced from the ownership, the possibility of profit maximization is reduced.

The advocates of profit maximization argue, however, that only profit-maximizing firms can survive in the long-run. They can achieve all other subsidiary goals easily if they can maximize their profits.

It is further argued that prevention of entry may be the major objective in the pricing policy of the firm, particularly in case of limit pricing. But then, the motive behind entry-prevention is to secure a constant share in the market. Securing constant market share is compatible with profit maximization.

4.3.4 Reasonable Profit Target

As noted above, objectives of business firms can be various. There is no unanimity among the economists and researchers on the objective of business firms. One thing is, however, certain that the survival of a firm depends on the profit it can make. So whatever the goal of the firm—sales revenue maximization, maximization of firm's growth, maximization of managers' utility function, long-run survival, market share, or entry-prevention—it has to be a profitable organization. Maximization of profit in technical sense of the term may not be practicable, but profit has to be there in the objective function of the firms. The firms may differ on 'how much profit' but they set a profit target for themselves. Some firms set their objective of a 'standard profit', some of a 'target profit' and some of a 'reasonable profit'. A 'reasonable profit' is the most common objective.

Reasons for aiming at reasonable profits

Economists argue that for a variety of reasons modern large corporations aim at making a reasonable profit rather than maximizing the profit. Joel Dean has listed the following reasons:

1. **Preventing entry of competitors:** Profit maximization under imperfect market conditions generally leads to a high 'pure profit' which is bound to attract competitors, particularly in case of a weak monopoly. The firms, therefore, adopt a pricing and a profit policy that assures them a reasonable profit and, at the same time, keeps potential competitors away.
2. **Projecting a favourable public image:** It becomes often necessary for large corporations to project and maintain a good public image, for if public opinion turns against it, the firm may lose its market. Besides, high profits may attract government attention and if government officials start raising their eyebrows on profit figures, corporations may find it difficult to sail smoothly. So most firms set prices lower

than that conforms to the maximum profit but high enough to ensure a reasonable profit.

3. **Restraining trade union demands:** High profits make labour unions feel that they have a share in the high profit and therefore they raise demands for wage-hike. Wage-hike may lead to wage-price spiral and frustrate the firm's objective of maximizing profit. Therefore, profit restrain is sometimes used as a weapon against trade union activities.
4. **Maintaining customer goodwill:** Customer's goodwill plays a significant role in maintaining and promoting demand for the product of a firm. Customer's goodwill depends largely on the quality of the product and its 'fair price'. What consumers view as fair price may not be commensurate with profit maximization. Firms aiming at better profit prospects in the long-run, sacrifice their short-run profit maximization objective in favour of a 'reasonable profit'.
5. **Other factors:** Some other factors that put restraint on profit maximization include (a) managerial utility function being preferable to profits maximization for executives, (b) congenial relation between executive levels within the firm, (c) maintaining internal control over management by restricting firm's size and profit, and forestalling the anti-trust suits.

Standards of reasonable profits

When firms voluntarily exercise restraint on profit maximization and choose to make only a 'reasonable profit', the questions that arise are (i) what form of profit standards should be used, and (ii) how should reasonable profits be determined?

Forms of profit standards

Profit standards may be determined in terms of (a) aggregate money collection, (b) percentage of sales, or (c) percentage return on investment. These standards may be determined with respect to the whole product line or for each product separately. Of all the forms of profit standards, the total net profit of the enterprise is more common than other standards. But when the purpose is to discourage the potential competitors, then a target rate of return on investment is the appropriate profit standard, provided competitors' cost curves are similar. The profit standard in terms of 'ratio to sales is an eccentric standard' because this ratio varies widely from firm to firm, even if they all have the 'same return on capital invested'. This is particularly so when there are differences in (a) vertical integration of production process, (b) intensity of mechanization, (c) capital structure, and (d) turnover.

Setting the profit standard

The following are the important criteria that are taken into account by the firms while setting the standards for a reasonable profit:

(a) **Capital-attracting standard:** An important criterion of profit standard is that it must be high enough to attract external (debt and equity) capital. For

NOTES

example, if the firm's stocks are being sold in the market at 5 times their current earnings, it is necessary that the firm earns a profit of 20 per cent of the book investment.

There are however certain problems associated with this criterion: (i) capital structure of the firms (i.e., the proportions of bonds, equity and preference shares) affects the cost of capital and thereby the rate of profit, and (ii) whether profit standard has to be based on current or long-run average cost of capital is a problem as it varies widely over time and may at times prove treacherous.

(b) **'Plough-back' standard:** In case a company intends to rely on its own sources for financing its growth, then the most relevant standard is the aggregate profit that provides for an adequate 'plough-back' for financing a desired growth of the company without resorting to the capital market. This standard of profit is used when maintaining liquidity and avoiding debt are main considerations.

(c) **Normal earnings standard:** Another important criterion for setting standard of reasonable profit is the 'normal' earnings of firms of an industry over a normal period. Company's own normal earnings over a period of time often serve as a valid criterion for reasonable profit, provided it succeeds in (i) attracting external capital, (ii) discouraging growth of competition, and (iii) keeping stockholders satisfied. When average of 'normal' earnings of a group of firms is used, then only comparable firms and normal periods are chosen.

However, none of these standards of profits is perfect. A standard is, therefore, chosen after giving a due consideration to the prevailing market conditions and public attitudes. In fact, different standards are used for different purposes because no single criterion satisfies all the conditions and all the people concerned—owners, managers, labour, banks, etc.

Conclusion

Although profit maximization continues to remain the most popular hypothesis in economic analysis, there is no reason to believe that this is the only objective that firms pursue. Modern corporations, in fact, pursue multiple objectives. The economists have suggested a number of alternative objectives that firms seek to achieve. The main factor behind the multiplicity of the objective, particularly in case of large corporations, is the dichotomy of management and the ownership and divergence in their own respective goals.

Moreover, profit maximization hypothesis is a time-honoured one. It is more easy to handle. The empirical evidence against this hypothesis is not unambiguous and conclusive. The alternative hypotheses are not strong enough to replace the profit maximization hypothesis. What is more important is that profit maximization hypothesis has a greater explanatory and predictive power than any of the alternative hypotheses. Therefore, profit maximization hypothesis

still forms the basis of most economic analysis. In the subsequent units, we have adopted this hypothesis and explained the price and output decisions of the business firms.

NOTES

Check Your Progress

4. What is the condition to earn the maximum profit under the first order condition?
5. How does Robin Marris define a firm's balanced growth rate?
6. What is the most plausible factor in managers' utility functions according to Baumol?

4.4 ANSWERS TO 'CHECK YOUR PROGRESS'

1. There are two approaches to estimate probabilities of outcomes of a business decision, viz., (i) a priori approach, i.e., the approach based on deductive logic or intuition, and (ii) posteriori approach, i.e., estimating the probability statistically on the basis of the past data.
2. A decision tree is a graphical device to map all possible managerial decisions in a sequence and their expected outcomes under different states of the economy.
3. The first step in the simulation technique is to build a mathematical model, called simulation model.
4. The necessary or the first order condition requires that for profit to be maximum marginal revenue (MR) must be equal to marginal cost (MC), i.e., MR = MC.
5. Robin Marris defines a firm's balanced growth rate (GB) in terms of the following equation: GB = GD = GC.
6. According to Baumol, the most plausible factor in managers' utility functions is maximization of the sales revenue.

4.5 SUMMARY

- Certainty is the state of perfect knowledge about the market conditions.
- In the state of certainty, there is only one rate of return on the investment and that rate is known to the investors.
- Risk refers to a situation in which a business decision is expected to yield more than one outcome and the probability of each outcome is known to the decision-makers or can be reliably estimated.
- There are two approaches to estimate probabilities of outcomes of a business decision, viz., (i) a priori approach, and (ii)) posteriori approach.

NOTES

- Uncertainty refers to a situation in which there is more than one outcome of a business decision and the probability of no outcome is known or can be meaningfully estimated.
- For the purpose of decision making, uncertainty is classified as (a) complete ignorance (b) partial ignorance.
- Pay-off Matrix was originally developed as a tool of Game Theory applied to decision-making under oligopoly. It is widely used in general business decision-making and also in investment decisions.
- The states of nature refer to the future market conditions in the long run on which the firm has no control.
- In the risk-adjusted discount rate, it is the rate of discount (r) which is adjusted for accounting for the risk.
- In the certainty-equivalent approach, it is the expected return (R) which is adjusted for the risk factor.
- The probability theory method can be applied only if a large number of similar investments are to be undertaken and if one project fails to yield the expected return, others faring so well that the loss is more than compensated.
- A decision tree is a graphical device to map all possible managerial decisions in a sequence and their expected outcomes under different states of the economy.
- The simulation technique uses the whole range of probability distributions of each decision variable, and sensitivity analysis uses probabilities in the high range only.
- The Laplace criterion uses the Bayesian rule to calculate the expected value of each strategy.
- Conventional theory of firm assumes profit maximization as the sole objective of business firms.
- In practice, however, firms have been found to be pursuing objectives other than profit maximization.
- Some important alternative objectives of business firms, especially of large business corporatins includes sales revenue maximization, maximization of firms growth rate, maximization of managerial utility function etc.
- Economists argue that for a variety of reasons modern large corporations aim at making a reasonable profit rather than maximizing the profit.
- Profit standards may be determined in terms of (a) aggregate money collection, (b) percentage of sales, or (c) percentage return on investment.

4.6 KEY TERMS

- **Pay-off Matrix:** It is a tabular array of strategic actions and their corresponding pay-offs under different states of nature.

- **Bayesian Rule:** This rule states that where meaningful estimate of probabilities is not available, the outcome of each strategy under each state of nature must be assigned the same probability and that the sum of probabilities of outcome of each strategy must be equal to 1.
- **Simulation:** It refers to a mathematical technique used to produce alternative target variables under certain stipulated conditions.
- **Net Present Value:** This value applies to a series of cash flows that occur at various intervals.
- **Sensitivity Analysis:** It is a tool used in financial modelling to analyze how the different values of a group of independent variables affect a specific dependent variable under certain specific conditions.
- **Marginal Cost:** It is the additional cost incurred for the production of an additional unit of output.
- **Marginal Revenue:** It refers to the incremental change in earnings obtained one additional unit is sold.
- **Profit Target:** It refers to a predetermined price point at which an investor will exit a trade for a positive gain.

4.7 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Differentiate between risk-adjusted discount rate and certainty-equivalent approach.
2. What are the limitations of the probability theory approach?
3. How do conventional economic theorists defend the profit maximization hypothesis?
4. What does the B-M-G hypothesis state?
5. Which technique is used to produce alternative target variables under certain stipulated conditions? Also, list the steps involved in it.

Long-Answer Questions

1. Explain the state of certainty with the help of an example.
2. Describe Baumol's hypothesis and its postulates.
3. 'Modern large corporations aim at making a reasonable profit rather than maximizing the profit.' Discuss the reasons.
4. What kind of data is required to construct a pay-off matrix? How does it help in investment decision-making?

4.8 FURTHER READING

NOTES

Dwivedi, D.N. 2015. *Managerial Economics,* Eighth Edition. New Delhi: Vikas Publishing.

Keat, Paul G. and Philip, K.Y. Young. 2004. *Managerial Economics: Economics Tools for Today's Decision Makers,* Fourth Edition. Singapore: Pearson Education.

Keating, B. and J.H. Wilson. 2004. *Managerial Economics: An Economic Foundation for Business Decisions,* Second Edition. New Delhi: Biztantra.

Peterson, H.C. and W.C. Lewis. 1999. *Managerial Economics*, Fourth Edition. Singapore: Pearson Education.

Salvantore, Dominick. 2001. *Managerial Economics in a Global Economy*, Fourth Edition. Australia: Thomson-South Western.

Thomas, Christopher R. and Maurice S. Charles. 2005. *Managerial Economics: Concepts and Applications*, Eight Edition. New Delhi: Tata McGraw-Hill.

UNIT 5 ADVERTISING AND WELFARE ECONOMICS

NOTES

Structure

5.0 INTRODUCTION

Advertising is a form of commercial communication in which a product, service, or concept is promoted or sold through the use of an explicitly sponsored, non-personal message. Various mass media, including traditional media such as newspapers, magazines, television, radio, outdoor advertising or direct mail; and new media such as search results, blogs, social media, websites or text messages are used to communicate advertising. The actual presentation of the message in a medium is referred to as an advertisement: advert or ad for short. Commercial ads often seek to generate increased consumption of their products or services through 'branding', which associates a product name or image with certain qualities in the minds of consumers.

Regarding the origin of welfare economics, it is very difficult to point out the period in the history of economic thoughts which marks the beginning of welfare economics. Nor it is reasonable to associate the emergence of welfare economics with any particular economist, because it does not appear at any time to have wholly engaged the labours of any one economist. Many economic problems involving the enhancement of social welfare are left unsolved by positive microeconomics. Nor it suggests appropriate policy measures that can maximize the economic well-being of the society as a whole. The branch of economic analysis which is concerned with these problems is called welfare economics. It must be checked that the allocation of resources is efficient or not in a company. This means whether any state or situation regarding resource allocation maximizes social welfare. In welfare economics, an attempt is made to define standards or norms by which alternative economic situations and policies can be judged or evaluated from the standpoint of efficiency or social well-being. This unit will discuss advertising, its objectives and scope. In addition, it will explain the concept of welfare economics in detail.

NOTES

5.1 OBJECTIVES

After going through this unit, you will be able to:

- Explain advertising and its objectives
- Discuss the role of advertising in economy
- Describe the concept of welfare economics

5.2 ADVERTISING

While watching television, all of us at some point have been driven mad by an advertisement interrupting our favourite show. You must have wondered at that point, why is the point of all these ads? Advertising communicates to people how the advertised product/service can satisfy some of their relevant needs in a meaningful way. It is a substitute of human salesmanship and includes all the visual or oral messages transmitted to people that aim to inform them and influence their buying behaviour, or to incline the target audience favourably towards institutions, ideas or persons that have been featured.

There is a saying in the world of marketing - doing business without advertising is like talking alone in an empty room; you know what you are saying, but nobody else does. This speaks volumes about the importance of advertising in the modern world. Its importance can also be judged by the fact that in today's world it is tough to imagine any organization which does not advertise. Every business strives to make profit. In this context, advertising is extensively used by business organizations to promote goodwill, increase sales, create awareness; in short reach people with the right sense at the right time. It is a tool for making an organization and its products and/or services known to the people at large, in such a manner that it brings about a desire to buy the product and/or avail the services being offered by the company.

Before moving on to the definitions of advertising, it is important to differentiate between advertising and advertisements. Advertising is a process which uses advertisements to promote a product/service/idea. It includes the process of planning, developing, preparing and producing, placing the advertisements in the selected media and finally assessing the effects of the advertisements. On the other hand, advertisements can be defined as materials that are used to convey persuasive messages to target prospects. Advertisements could be printed or broadcasted.

Advertising also involves significant financial implications. At the same time, the objectives achieved by way of advertising may or may not be commensurate with the incurred expenses. Successful advertising depends on the right combination of numerous factors such as market trends, brand preference and various media used to reach the target audience.

NOTES

Definitions of advertising

The word 'advertising' has been derived from the Latin word — *advertere* — which means to attract people's attention to a specific product or service. The most commonly used definition of advertising was given by the American Marketing Association (1960). It defined advertising as any paid form of non-personal presentation and promotion of ideas, goods or service by an identified sponsor. In the simplest of terms, advertising can be defined as a paid form of non personal communication about an organization or its products that is transmitted to a target audience, through an appropriate medium.

- According to Mason & Roth, advertising is a salesmanship without a personal salesman.
- Krick Patrick defines advertising as mass communication of information intended to persuade buyer so as to maximize dollar profit.
- Advertising is basically a paid form of (or commercial) mass communication. Professor James E Littlefield and Professor CA Kirkpatrick, in their book *Advertising: Mass Communication in Marketing*, have defined advertising while differentiating commercial communication, mass communication and commercial mass communication:
  (i) The basic function of commercial communication is to change a person's attitudes about or actions toward a subject or object.
  (ii) The purpose of mass communication is to change or reinforce the attitudes of many individuals and if possible, to cause them to take actions that are favourable to the communication.
  (iii) The basic objective of advertising, as commercial mass communication is mostly to induce purchase. Advertising tries to succeed in this objective by trying to change the buying behaviour of potential customers, by sustaining or reinforcing the buying behaviour of existing customers and of course, by trying to create or maintain a favourable action towards the advertised product or company.
- Albert Laskar gave a definition of advertising in the 1920s, when the only advertising media available were newspapers and magazines. According to him, 'Advertising is salesmanship in print'.
- The Definition Committee of the American Marketing Association (AMA) gave one of the most commonly used definitions of advertising in 1948. It defined advertising as any paid form of non-personal presentation and promotion of idea, goods, or services by an identified sponsor. Although this definition is widely accepted, yet there are some shortcomings in the definition. In 1988, Dorothy Cohen, in her book *Advertising* pointed out that this definition does not mention about the persuasive nature of advertising, its creative aspects and the large number of advertising media.

NOTES

So Dorothy Cohen suggested the following definition:

> 'Advertising is a business activity employing creative techniques to design persuasive communication in mass media that promote ideas, goods and services in a manner consistent with the achievement of the advertiser's objectives, the delivery of consumer satisfaction and the development of social and economic welfare.'

- A British advertising industry's professional body, The Institute of Practitioners in Advertising (IPA), describes advertising as a fascinating fusion of business and art. According to IPA, the planner, the negotiator, the writer, the artist, the production expert, the management coordinator combine to form a team which delivers pertinent and hopefully first-rate advertising solutions to help clients achieve their business goals.
- Subroto Sengupta, in his highly acclaimed book, *Brand Positioning: Strategies for Competitive Advantage,* defines advertising as the discovery and communication of a persuasive difference for a brand to the target prospect.
- Prof. Jaishri N. Jethwaney, who teaches advertising at the Indian Institute of Mass Communication, New Delhi, has provided a working definition of advertising in her book. According to Jethwaney, advertising is the art and science of building brands through persuasive communication and positioning them in consumers' perception with constant vigil on the market situation and consumer expectations.
- According to another definition, advertising is an organized method of communicating information about a product or service, which a company or an individual intends to sell. It is a paid announcement that is communicated with the help of words, pictures, illustrations and music in a medium/media found appropriate for reaching the target audience/ prospective buyers.
- Burt Manning of J Walter Thompson, defines advertising as one of the important forces which serves the public interest. It is a form of open communication between those who sell and those who buy. It is a form of advocacy open to any company or cause that wants to argue in case. The jury is the public and every purchase is a vote.

Nonetheless, advertising is different from salesmanship. Unlike salesmanship, which involves direct face-to-face communication, it is clear from the various definitions that advertising is non-personal and indirect means of communication with the prospects by means of various media. Salesmanship involves communication through personal interview between the salesman and the prospect while by way of advertising; the advertiser reaches a large number of prospects simultaneously.

Advertising also differs from publicity in the sense that publicity may or may not have an identified sponsor, whereas advertising always has an identified sponsor. Unlike publicity, advertising is a paid form of communication. In fact, it would not be incorrect to say that paid publicity is advertising.

5.2.1 Purpose and Functions of Advertising

NOTES

Advertising is always goal oriented and it attempts to accomplish a wide variety of goals. The objectives of advertising may be to:

- Catch the attention of the prospective consumers/customers.
- Promote or persuade prospective consumers/customers to find out more about the product or service, for example, visiting the company's website.
- Create awareness about the product.
- Attract the prospective consumers/customers to try the product or avail the services by offering introductory prices, trial offers, free gifts and coupons.
- Introduce a new product/service/idea effectively.
- Reassure the customers that the brand is trustworthy and reliable.
- Disseminate any additional information regarding the advertised product/service, such as any new feature of the product.
- Reaffirm an existing brand image or to create a new brand.
- Increase immediate profits.
- Influence a desired change in the buying behaviour and consumption pattern of the people. For example, it may try to persuade us to join a particular computer course or donate for a charitable cause.

Advertising should concentrate on clear goals that should be measurable.

These objectives are known as DAGMAR (Defining Advertising Goals for Measured Advertising Results). This concept was given by Russell Colley in the year 1960 in his book, *Defining Advertising Goals for Measured Advertising Results*. He defined advertising as a paid form of mass communication, the ultimate purpose of which is to impart information, develop attitude and induce action that is beneficial to the advertiser, generally the sale of a product.

In the same book, Colley concluded that 'All commercial communication that weigh on (or aim at) the ultimate objective of a sale, must carry a prospect (recipient of advertising message) through four levels of understanding. The prospect must first be aware of the existence of a brand or a company. He must have a comprehension of what the product (or company) is and what will it do for him. He must arrive at a mental conviction to buy the product. Finally, he must stir himself to action.'

A good (read great) advertisement should satisfy the objectives of both the advertiser and the consumer. Advertising should meet the objectives of consumers by engaging them and delivering the desired message. Nonetheless, it must achieve the advertiser's objectives. Advertiser and consumer have different expectations from advertising. The advertiser must ensure that the advertisement satisfies all the objectives/expectations of both the advertiser and the consumer.

NOTES

Functions of advertising

At the elementary level, advertising is all about influencing the target audience by providing them with information. Although each advertisement has specific objectives to accomplish, advertising performs the following three primary functions:

(a) Giving information regarding product and brand: Providing relevant information to the target audience is the chief function of advertising. This information, in turn, helps the prospects of decision-making. The kind of information in advertisement will depend on the product/service/idea being advertised and the needs of the target audience. If the product is technical, the information carried in the advertisement will be detailed.

(b) Giving incentives to take action: Various researches indicate that consumers are averse to switching brands even when they are not completely satisfied with the product they are using. It is not very easy to change their established behaviour. Advertising helps in bringing about the required change in the buying behavior of the consumers by way of providing them reasons and offering incentives to switch product brands. Incentives being offered may include high quality, competitive price, trial packs, free gifts and warranties.

(c) Giving reminders and reinforcements: Advertising is not only directed at the prospective consumers/customers, it is directed at current customers as well. Over a period of time, the consumers may forget why they bought a particular brand of TV or car. Advertising repeatedly reminds the consumers about their purchased brand, it unique features, benefits value, and so on. Constant exposure to such messages (in the form of advertisements) helps reinforce the consumer's decision and helps in creating brand loyalty.

Other functions of advertising can be summed up as follows:

- Registering the brand name
- Reminding the brand name
- Reinforcing the brand name
- Creating brand differentiation
- Creating brand preference
- Creating brand loyalty
- Reinforcing brand image
- Changing brand image
- Introducing a new product/idea/service
- Creating corporate image
- Reinforcing corporate image
- Changing corporate image

In the early 20th century, Daniel Starch, pioneer in advertising research, gave a formula to explain the functions of advertising. This formula is popularly known as the Starch formula. According to the Starch formula, in order to be effective, an advertisement should be:

- Seen, read or heard
- Believed

- Remembered
- Acted upon

5.2.2 Nature and Scope of Advertising

NOTES

Advertising is both an art and a science. It is the art of bringing about the desired effects amongst the target group by way of exposing them to mass-produced messages. It involves the science of human psychology to ensure effectiveness of the message.

- Advertising provides a link between the producer and the consumer.
- Advertising is non-personal and thus excludes interpersonal communication.
- Advertising promotes tangible products like car, iPod; ideas like prevention from cervical cancer, family welfare and services like hospitality, insurance policy and others.
- Advertising is a paid form of communication where the sponsor (advertiser) is always identified. The sponsor could be a commercial or non-commercial organization.
- It tries to bring about a desired change in the behaviour of the target audience.
- Advertising appears in a recognized media such as newspapers, magazines, hoardings, radio, TV and direct mail.

One, must also venture into the world of advertising for better understanding in terms of those who play an important role in bringing advertisements to the consumers. There are five key players, namely:

(i) The advertiser

(ii) The advertising agency

(iii) The media

(iv) The vendor

(v) The target audience

(i) **Advertiser:** Advertising begins with the advertiser. An advertiser is usually the initiator of the advertising process. An advertiser could be an individual or an organization or an institution. An advertiser may wish to communicate information about product/service/idea to a target segment. It is the advertiser who takes the final call on all important decisions like the advertising budget, various media to be used to reach the target audience and the campaign duration.

(ii) **Advertising agency:** It is the second player in the advertising world. Advertising agencies are hired by the advertiser to plan and execute an individual commercial or a complete advertising campaign. This understanding between the advertiser and the advertising agency is called the agency-client partnership. Advertising agencies provide strategic and creative expertise, media knowledge and workforce talent to the advertiser. These agencies also negotiate good deals for their clients.

NOTES

Large companies may even have their own in-house advertising agencies. This ensures closer monitoring and better control over advertising. These in-house agencies perform most advertising functions and many a times they also undertake all the functions of an external advertising agency in a more cost-effective manner. It was the Liverpool-based White Star Line's in-house advertising agency that planned and created most of Titanic's advertising.

(iii) **Media:** Media can be termed as the third player. Media is the term used to describe the channels of communication that are used to carry messages from the advertiser to the target audience. The selected media must deliver advertising messages in a way that is consistent with the creative effort.

Media sells space in newspapers, magazines as well as in support media like World Wide Web to the advertiser. It also sells time in electronic media like TV and radio. Media organizations also help advertiser/advertising agency in selection of appropriate media for transmitting message to the audience and offer assistance in ad designing.

There are different media that can be used for advertising. It is not rare to find advertisements on billboards, walls, web banners, shopping carts, web pop-ups, bus stop benches, logo-jets (advertisements on the sides of aeroplanes), taxi doors and even in airbuses on the seat-back tray tables and overhead storage bins. These days we even find advertisements on fruits. It is not rare to find advertising stickers on apples when we go to a supermart to purchase them. There are even advertisements on the back of movie tickets and supermarket receipt bills. Advertisements may seem interesting and entertaining at times and annoying at other time.

(iv) **Vendor:** Vendors are the fourth player in the advertising world. Vendors are a group of service organizations that are responsible for providing assistance to advertisers, advertising agencies and the media. Vendors include freelancers, consultants and other professionals who assist in meeting the advertising objectives and planning as well as designing an advertisement. It is not feasible to hire all kinds of professionals as it means a lot of money. Also, many a times, the advertiser or the agency handling his advertising campaign may not have an expertise in a particular field or they may also have their hands full when it comes to work. Thus, an advertiser or an advertising agency may avail services of copywriters, graphic designers, market researchers and public relations consultants.

(v) **Target audience:** The fifth player in the world of advertising is the target audience. The advertising strategy, for any product or service or idea, always keeps in mind the target audience. In terms of marketing, target audience denotes the person who purchases a product (also called a customer). The advertiser/agency spends a reasonably good amount of money on

ascertaining the target audience, they need to communicate with. However, many a times in spite of a lot of research, a campaign may fail to reach particular target audience. One must keep in mind that in advertising there may be more than one segment that the advertiser needs to target. The latest advertising of Knorr soupy noodles targets both children and their mothers.

5.2.3 Advertising as a Marketing Tool

The American Marketing Association defines advertising as 'The process of planning and executing the conception, pricing, promotion, and distribution of ideas, goods and services to create exchange that satisfies individual (customer) and organizational objectives.'

Accordingly, organizations aim at creating products that may satisfy the needs and requirements of the customer. This, in turn, requires a set of well coordinated activities. Advertising plays a vital role here by communicating this information about the product to the target audience. Although there may be different objectives behind marketing a product, the primary marketing objectives include introducing product; motivating new customers to buy the product; inducing the present customers to make more purchases of the product; obtaining more shelf space; maintaining sales in off-seasons and competing with other brands.

Marketing mix

The process of marketing involves three major steps. First, satisfying the wants and needs of customers. Secondly, coordinating all marketing efforts across the organization and thirdly, achieving long-term goals.

There are four major components of marketing—product, price, place and promotion. All these four components are termed as controllable variables as the marketers can control these components. Based on market research and keeping in mind the customers' needs and expectations, marketers often vary these variables. They may create a mediocre quality product and put a reasonable price tag or they may decide to manufacture high quality product and price it high, depending on the target audience's profile. Marketers may choose personal selling to sales promotion or vice versa depending upon the means which is most suitable for advertising their product. Thus, advertising and marketing are related.

Promotion is an important component of marketing and it helps in achieving marketing objectives. A product can be promoted only with the help of communication. Hence, it is also termed as marketing communication. The process of promotion involves apprising the target audience about the suitability of the product for them in terms of quality, availability and affordability of the product. In order to ensure that product promotion gives the marketer desired results, it must be carried out in an attractive, appealing, adequate and convincing manner.

NOTES

For the purposes of promoting a product, there are four tool options available with the marketer. These are – sales promotion, personal selling, public relations and advertising. They are also known as promotion mix. Just as in case of marketing mix, marketers exercise control over the promotional tool options. They decide which promotion tool has to be used and to what extent. Earlier, these promotional tools were used separately. However, these days there is a change in the trend. Now, marketers adopt an integrated approach, which means that marketers integrate more than one promotional tool for the purposes of promoting their product. When marketers adopt an integrated approach for promotion, then it is known as integrated marketing communication. Let us study these tools.

The first promotional tool is 'sales promotion'. Sales promotion is a term used for all the activities that are beneficial in promoting sales of a product. These activities could be trade-related or consumer-related. Trade-related sales promotion is a part of push strategy. Under this strategy, the trade chain (which includes wholesalers, dealers and retailers) is encouraged to push the product. Just in contrast is the consumer-related sales promotion which is a part of the pull strategy. Here, the focus is on consumers who are encouraged to begin and continue buying a particular brand. It is through advertising that the consumers are apprised of attractive offers, to increase the sales.

The second promotional tool is 'personal selling'. It can be a very effective tool of promotion, but mostly in case of industrial goods. Many in service sector, such as insurance, also make use of personal selling. However, one may not find it feasible for many products, particularly for consumer goods. Personal selling requires a large sales force and is very time-consuming. It is practically not possible to reach out to a very large number of prospective consumers.

The third promotional tool is 'public relations' (PR). It is a long-term process which helps in creating goodwill and enhancing the reputation of the organization. It is with the help of public relations that mutually beneficial relations are created and then sustained between the organization and different stakeholders like shareholders, customers, investors, media and employees. The positive image of the organization created with the help of public relations in turn helps in achieving the marketing objectives. PR is beneficial in creating and sustaining loyalty.

5.2.4 Economic Effects of Advertising

Advertising has the following economic effects:

(i) **On gross domestic product (GDP):** Advertisements stimulate demand, spending by consumer rises, thus increasing gross revenues of firms and contributing to an increase in GDP, helping in introducing new products.

(ii) **On business cycles:** Advertising has a stabilizing effect on downturns in business activities. Many firms increase advertising expenditure in times of recession. Firms that maintain advertising during recession perform better afterwards as compared to firms that slash their advertising expenditure. This is because many competitors cut expenses, thus increasing the

prominence of the company's advertising. Advertising during an economic downturn also sends positive signals to consumers about the company's brand image, as they perceive the company is doing well even during the period of recession.

(iii) **On competition:** Advertising stimulates competition and hence companies come up with better products and production methods which benefit the economy as a whole. Advertising provide a gateway to gain entry into new markets and consequently, competition across economies is encouraged. However, critics have highlighted that amount of money needed to compete efficiently in many industries is often unaffordable. As a result, advertising acts as a barrier to entry into an industry, that is, a firm may have all the potential to compete in an industry except the advertising expenditure, which is very high. Advertising can actually decrease the overall amount of competition.

(iv) **On prices:** Cost for advertising is built into price but this must be judged against how much effort a consumer would have to put into searching for a product without benefit of advertising. Advertising by stimulating demand causes economies of scale resulting in lower costs and hence lower price.

(v) **On value:** Value is the perception by the consumer that a product provides beyond the cost incurred to acquire the product. Value is added to consumption process by advertising. The experience of eating at McDonald's or drinking a Coke is enhanced by the expectation the advertising has created and reinforced within the consumer. Advertising affects a consumer's perception of value by contributing to symbolic value and social meaning of a brand. Symbolic value refers to what a product means to consumers in a non-literal way. For instance, automobiles symbolized self-concept for some consumers. Social meaning refers to what a product means in a social context. Social class is marked by any number of products used and exhibited to suggest class membership such as car, clothes and others. Time and again, the product's connection to a social class tackles a need within consumers to move up in a class. Advertising develops brand image. It adds directly to the consumers' perception of the value of the brand. The more the value consumers see in a brand, the more they can pay to acquire the brand.

Check Your Progress

1. Who gave the concept of DAGMAR?
2. Define the term 'advertiser.'
6. Mention any three promotional tools.

5.3 WELFARE ECONOMICS

Welfare economics is the study of economic welfare of the members of a society as a group. In the words of Oscar Lange, 'Welfare economics is concerned with the conditions which determine the total economic welfare of a community.' Reder

NOTES

defines 'welfare economics' as that 'branch of economics science that attempts to establish and apply the criteria of propriety to economic policies.' In his survey of welfare economics, Mishan defines 'theoretical welfare economics' as 'that branch of study which endeavours to formulate propositions by which we may rank on the scale of better and worse, alternative economic situations open to society'. The function of welfare economics is to evaluate the alternative economic situations and determine whether one economic situation yields greater economic welfare than others. Welfare economics may also be defined as that branch of economic science which evaluates alternative economic situations (i.e., alternative patterns of resource allocations) from the viewpoint of economic well-being of the society as a whole.

Nature of welfare economics

Economists hold different views on the question *whether welfare economics is a positive (pure) or normative (applied) science.* Although welfare economics has been closely associated with positive economics from the inception of economic thinking, 'at one point in economic thought, it was felt that welfare economics was unscientific; that it was normative and was hence a branch of Ethics.' It was also argued that welfare economics is concerned with 'what ought to be' and, hence, it is 'normative' in character. This view, however, has not been universally held. Pigou, for example, was concerned, in his *Economics of Welfare*, with the causes of economic welfare and did not make any policy recommendation. Pigou's *Economics of Welfare* is, therefore, a positive study.

A widely held view on this issue is that *welfare economics is both a positive and a normative science.* Positive economics is primarily concerned with understanding, explaining and predicting the working of the economic system. Welfare economics is a positive science insofar as it attempts to explain and predict the outcome of the functioning of the economic system. Welfare propositions 'may be subjected to test in the same way as those of positive economics,' though testing welfare propositions is much more difficult than the propositions of general positive economics.[7] The information gained through positive analysis is useful in devising appropriate policy measures to maximise the welfare of the society. Welfare economics is a normative science as it provides guidelines for policy formulations to maximise social welfare. Maximisation of economic welfare presumes a welfare function which consists essentially of value judgements. Given the welfare function, welfare economics, as a normative science, provides guidelines for appropriate policy measures.

5.3.1 The Concept and Measurement of Social Welfare

The term 'welfare' has been defined in many diverse ways, perhaps, because it is extremely difficult to give it a precise meaning. The difficulty arises from the fact that welfare of an individual or of a group of individuals depends on diverse and

NOTES

immeasurable factors including social, political, economic and philosophical attitude of the people towards life and society. In economics, however, the concept of welfare is used in a narrow sense: it is limited to only economic welfare. Even the term 'economic welfare' eludes a precise definition. Nevertheless, economists have tried to give it a concrete meaning.

(a) Cardinal utility approach: Social welfare is sum of individual utility

The concept of economic welfare used in economic literature is synonymous to utility or happiness. The economic welfare of an individual, thus, equals the total utility he derives from the goods and services that he consumes, the leisure that he enjoys, and deeds that he performs, but not all of these. Economic welfare of an individual equals the total utility derived from only those goods and services and only that part of leisure and deeds which can be exchanged for money or can be brought under the measuring rod of money.

On the basis of this concept of individual welfare, Bentham defined social welfare as 'the sum total of the happiness (or welfare) of all the individuals in society.' Following Benthamite doctrine, Pigou defined social welfare as the arithmetic sum of the individual welfare. In a nutshell, social welfare, i.e., the aggregate of welfare of society, was regarded (by the economists of cardinal utility tradition) as the arithmetic sum of the cardinally measurable utilities of the individual members of society.

The concept of social welfare based on cardinal utility has, however, met with certain serious objections.

First, it is argued that utility cannot be cardinally measured and, hence, cannot be added to obtain the aggregate welfare. That is, the concepts that cannot be quantified cannot be added together. It is therefore meaningless to define social welfare as the sum of the individual welfare. This objection is universally accepted.

Second, since ordinal measurement of utilities is not possible, interpersonal comparison of utilities is not possible in an objective or scientific way. It would, therefore, not be possible to determine how a change in existing pattern of resource allocation affects the aggregate welfare unless it is unrealistically assumed that all individuals have identical income utility and commodity utility functions.

Owing to these problems, Benthamite and Pigovian concept of social welfare had become inoperational as it could not be used objectively in any policy formulation. Therefore, the cardinal utilitarian thesis that the welfares of different individuals could be added up to arrive at the welfare of society had to be abandoned.

(b) Pareto's approach: Social welfare as social optimum

It was Vilfred Pareto, an Italian economist, who broke away from the cardinal utility tradition and gave a new orientation to welfare economics. He rejected cardinal utility concept and additive utility function on the ground of their limitations.

NOTES

With the rejection of cardinal utility thesis, the attempts to quantify the social welfare ended, at least temporarily, perhaps because welfare is not an observable quantity like a market price or an item of personal consumption.

Pareto introduced a new concept, i.e., the concept of *social optimum*. This concept is central to Paretian welfare economics. The basic idea behind this concept is that while it is not possible to add up utilities of individuals to arrive at the total social welfare, it is possible to determine whether social welfare is optimum. Conceptually, social welfare is said to be *optimum* when nobody can be made better-off without making somebody worse-off. In the words of Boulding, 'a social optimum is defined as a situation in which nobody can move to a position which he prefer without moving somebody else to a position which is less preferred.'

The basic point in regard to the concept of social optimum which needs to be noted is that social optimum does not define (or determine) a quantity or magnitude of welfare. It is rather associated with the question whether the magnitude of social welfare from a given economic situation can be or cannot be increased by changing the economic situation. The test of increase in social welfare is that at least one person should be made better-off without making any body else worse-off.

However, it is difficult to conceive economic policies which can improve the welfare of an individual without injuring another. To overcome this problem, the economists, viz., Kaldor, Hicks and Scitovsky, have evolved the compensation principle. This principle states that the person who benefits from an economic policy or reoganisation must be able to compensate the person who becomes worse-off due to this policy and yet remain better-off.

To conclude, modern welfare economics does not attempt to quantify the total social welfare. It concerns itself with only the indicators of change in welfare. It analyses whether total welfare increases or decreases when there is a change in economic situation. This approach is based on the premise that while cardinal measurement of utility is not possible, expression of utility in ordinal sense is possible and it is an adequate guide to change in the welfare of an individual. It is this principle on which the modern welfare criteria are based.

Theories of welfare and welfare ceiteria

Having introduced the welfare economics and the concept of economic welfare, let us now discuss various theories of welfare and welfare criteria divised by welfare economists. In this section, we will discuss some important contributions to welfare economics and welfare criteria beginning with Adam Smith. The whole discussion is broadly divided in four stages: (i) classical welfare economics; (ii) neo-classical welfare economics; (iii) paretian welfare economics; and (iv) post-Paretian contributions to welfare economics.

Let us discuss the first two stages in detail.

NOTES

(i) Classical welfare economics: Smith's welfare criteria

Classical economists did not develop any specific theory of welfare. One may, however, find welfare propositions in Adam Smith's *Wealth of Nations*. Smith's criteria of welfare may be summarised as follows. The final aim of all production is consumption and increase in consumption results in increase in the level of satisfaction. Therefore, increase in the level of national output (or wealth) leads to increase in the level of welfare. Adam Smith thus implicitly assumed that the level of welfare of a society is proportional to the level of its total physical output. Adam Smith regarded growth in national product (i.e., the national wealth or dividend) as the main determinant of welfare.

Distribution of national wealth (i.e., allocative efficiency) was only a subsidiary theme to Adam Smith. In respect of resource allocation he showed that the equilibrium process of the competitive market will lead to an optimum allocation of resources among different industries. Free competition will thus ensure that right commodities would be produced in right quantities and price of each commodity would be equal to its cost. Adam Smith's welfare criterion that growth in gross national output adds to the social welfare assumes existing income distribution as just. Growth in national income however might lead to inequalities in incomes which may reduce the social welfare.

(ii) Neo-classical welfare economics: Pigou's welfare economics

The neo-classical welfare economics refers to the welfare propositions 'which emerged from the mixture between classical economic ideas and those of the marginal utility school.' The major contributors to the neo-classical welfare economics were, chronologically, Sidgewick, Marshall and Pigou. Pigou was the main neo-classical economist who made significant contribution to welfare economics.

As regards the views of other neo-classical economists, 'with Sidgewick's *Principles of Political Economy'*, economic thought started moving away from the classical materialistic level to the modern subjective level of analysis'. Sidgewick was the first to stress the distinction between production and distribution welfare economics, whereas classical welfare economists had concentrated only on Production Welfare. While he accepted the classical viewpoint that increase in production of wealth increases total happiness, he favoured the view that 'a more equal distribution of wealth tends *prima facie* to increase happiness or welfare. He was, however, of the view that redistribution might cause reduction in aggregate production of wealth. His clear-cut distinction between production and distribution welfare economics led him directly to consider the possible divergence between *private* and *social net* products.

Marshall's major contribution to the welfare economics is the introduction of the concept of consumer's surplus. Marshall's doctrine of consumer's surplus did generate high hopes for being used as an instrument of social betterment. Consumer's surplus was treated as a measure in welfare propositions. Although Marshall himself eventually lost his faith in his doctrine of consumers' surplus, 'the

prospect of its extension as a practical tool of welfare analysis has tempted the ingenuity of several eminent economists, among whom the best known is Prof JR Hicks.'

NOTES

It is however Pigou's *Economics of Welfare* which marks the culmination of neo-classical welfare economics. Some argue that Pigou's *Economics of Welfare* represents the emergence of 'Welfare Economics' as a separate branch of economic theory. Let us, therefore, discuss Pigovian Economics in detail.

Pigovian welfare economics

Let us study understand the meaning of welfare according to Pigou.

Meaning

Pigou's approach to welfare economics is pragmatic. In his own words, 'The goal (of his enquiry into the welfare economics) is to make more easy practical measures to promote welfare—practical measures which statesmen may build upon the work of the economists.' Therefore, he limited his enquiry to only that part of social welfare which could be measured with the measuring rod of money. Accordingly, his enquiry 'becomes limited to that part of social welfare that can be brought directly or indirectly into relation with the measuring rod of money.' This part of social welfare he called economic welfare. Economic welfare according to him, is limited to the extent to which individual's wants are satisfied. Thus, economic welfare of an individual implies the total utility he derives from the goods and services that can be exchanged for money. Following Bentham, as mentioned earlier, Pigou defined the welfare of a society as the arithmetic sum of utilities of individuals of the society.

Pigovian criteria of welfare

In order to devise his welfare criteria, Pigou made the following postulates:

1. The objective of each individual is to maximise his satisfaction from his expenditure on goods and services. To this end, consumers spend their incomes on different goods and services in accordance with the Marshallian principle of equimarginal utility.
2. Individuals have equal capacity for satisfying their wants. That is, people with similar characteristics with respect to their race, country, environment, habits, cultural heritage, income, social status, etc., are on the average much like in their consumption behaviour. This postulate enabled him to assume that utilities, though not measurable strictly in the cardinal sense, are comparable both *intrapersonally* and *interpersonally.* In his own words, "on the basis of analogy, observation and intercourse, interpersonal comparisons can properly be made; and moreover, unless we have a special reason to believe the contrary, a given amount of stuff may be presumed to yield a similar amount of satisfaction, not indeed as between any one man and any other, but as between representative members of groups of individuals. Thus, Pigou postulates that individuals derive, by and large, the same satisfaction out of the same real income.

3. Real income of the people consists of a single commodity, i.e., money, and is subject to the law of diminishing marginal utility and that the marginal unit of money must yield the same amount of satisfaction to each individual.

NOTES

Pigou's dual criterion

On the basis of the above assumptions, Pigou adopted a dual criterion for judging the improvement in social welfare. His dual criterion may be stated as (i) an increase in the value of national dividend (or income) without corresponding increase in the supply of factors, and (ii) a transfer of income from rich to poor, indicate increase in social welfare. Pigou provided additional qualifications to his dual criterion as follows.

In regard to the first criterion, he felt that national dividend could be increased, factor supply remaining constant, either by increasing some goods without diminishing the production of others or by transferring factors of production to activities in which their social value is higher. Any such increase in national dividend without decrease in the share of the poor is to be regarded as an improvement in the social welfare.

Regarding the second criterion, he suggested that redistribution of national income must not lead to decrease in the national dividend. Thus, he proposed his second criterion as any reorganisation of the economy or redistribution of incomes which increases the share of the poor without causing reduction in the national dividend, to be accepted as an improvement in the social welfare.

Criticism of Pigovian welfare economics

First, Pigou's definition of social welfare is questionable. Since utility (or satisfaction) is not cardinally measurable, it cannot be added, and therefore, social welfare cannot be defined, as Pigou did, as the arithmetic sum of the utilities of all individuals of a society.

Second, Pigou's postulate that both interpersonal and intrapersonal comparisons of utility are possible has no scientific basis. As IMD Little has observed, 'any such comparison is a value or an ethical judgement and not an ordinary empirical judgement about a matter of fact.'

Third, Pigou's assumption that all individuals have identical utility functions for the same money incomes is also empirically unsustainable because attitude towards money differs from individual to individual.

Fourth, while Pigou intended to make his inquiry a positive study, the involvement of value judgement in his postulates (as indicated in the second criticism) makes his study normative in character.

Finally, even if the postulates of Pigou are granted, mere increase in national output cannot guarantee increase in the social welfare. Composition of national product, apart from equal distribution of incomes, is equally important. Consider the famous example of gun and butter. If proportion of gun in the national product increases, and that of butter decreases, social welfare will not increase—it will rather decrease.

NOTES

Check Your Progress

4. Why is it difficult to give a precise meaning to the term, welfare?
5. Who rejected the cardinal utility concept and additive utility function?
6. What is the major contribution of Marshall to welfare economics?

5.4 ANSWERS TO 'CHECK YOUR PROGRESS'

1. The concept of DAGMAR was given by Russell Colley in the year 1960 in his book, *Defining Advertising Goals for Measured Advertising Results*.
2. An advertiser is usually the initiator of the advertising process.
3. The three promotional tools are sales promotion, personal selling and public relations.
4. The difficulty to give a precise meaning to the term, welfare arises from the fact that the welfare of an individual or of a group of individuals depends on diverse and immeasurable factors including social, political, economic and philosophical attitudes of the people towards life and society.
5. Vilfred Pareto rejected the cardinal utility concept and additive utility function.
6. Marshall's major contribution to welfare economics is the introduction of the concept of consumer's surplus.

5.5 SUMMARY

- The word 'advertising' has been derived from the Latin word — *advertere* — which means to attract people's attention to a specific product or service.
- Advertising also differs from publicity in the sense that publicity may or may not have an identified sponsor, whereas advertising always has an identified sponsor.
- Advertising is a tool for making an organization and its products and/or services known to the people at large, in such a manner that it brings about a desire to buy the product and/or avail the services being offered by the company.
- A good (read great) advertisement should satisfy the objectives of both the advertiser and the consumer.
- The five key players in the advertising process are the advertiser, advertising agency, media, vendor and target audience.
- The process of marketing involves three major steps. First, satisfying the wants and needs of customers. Secondly, coordinating all marketing efforts across the organization and thirdly, achieving long-term goals.

NOTES

- There are four major components of marketing—product, price, place and promotion.
- Promotion mix includes four tools available with marketers to promote a product—sales promotion, personal selling, public relations and advertising.
- Welfare economics is the study of economic welfare of the members of a society as a group.
- The function of welfare economics is to evaluate the alternative economic situations and determine whether one economic situation yields greater economic welfare than others.
- There are two broad approaches to defining and measuring social welfare: Cardinal utility approach, social welfare is a sum of individual utility.
- Pareto introduced a new concept, i.e., the concept of social optimum.
- The theories of welfare and welfare criteria are drivided into four stages: classical welfare economics, neo-classical welfare economics, paretian welfare economics and post paretian contributions to welfare economics.

5.6 KEY TERMS

- **Advertising:** It is a type of marketing communication in which a product, service or idea is sold through an openly sponsored or non-personal message.
- **Advertiser:** This is a person or company that promotes a product, service, or event.
- **Marketing Mix:** It is a set of marketing tools or methods that are used to promote a product or service in the market and sell it.
- **Welfare Economics:** It is a branch of economics that assesses well-being at the aggregate level using microeconomic techniques.
- **Compensation Principle:** This is the amount that must be paid to another economic agent in order to compensate them for their loss of economic welfare.

5.7 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. What are the criteria for an advertisement on the basis of the Starch formula?
2. What is the role of media in the advertising process?
3. Define marketing mix. What are the stages involved in the process?
4. How does advertising impact the economy of a country?

NOTES

5. Write the dual criterion for judging the improvement in social welfare given by Pigou.
6. Write a short note on 'Classical Welfare Economics'.
7. Differentiate between classical and neo-classical welfare economics.

Long-Answer Questions

1. Explain the primary functions of advertisement.
2. Analyse the postulates made by Pigou to devise his welfare criteria.
3. Discuss the criticism of Pigovian welfare economics.

5.8 FURTHER READING

Dwivedi, D.N. 2015. *Managerial Economics,* Eighth Edition. New Delhi: Vikas Publishing.

Keat, Paul G. and Philip, K.Y. Young. 2005. *Managerial Economics: Economics Tools for Today's Decision Makers,* Fourth Edition. Singapore: Pearson Education.

Keating, B. and J.H. Wilson. 2005. *Managerial Economics: An Economic Foundation for Business Decisions,* Second Edition. New Delhi: Biztantra.

Peterson, H.C. and W.C. Lewis. 1999. *Managerial Economics*, Fourth Edition. Singapore: Pearson Education.

Salvantore, Dominick. 2001. *Managerial Economics in a Global Economy*, Fourth Edition. Australia: Thomson-South Western.

Thomas, Christopher R. and Maurice S. Charles. 2005. *Managerial Economics: Concepts and Applications*, Eight Edition. New Delhi: Tata McGraw-Hill.

MBA, First Year
Paper - II

MANAGERIAL ECONOMICS

Madhya Pradesh Bhoj (Open) University, Bhopal
(Established under an Act of State Assembly in 1991)

मध्यप्रदेश भोज (मुक्त) विश्वविद्यालय, भोपाल

Kolar Road, Near Swarn Jayanti Park, Yashoda Vihar Colony,
Chuna Bhatti Bhopal, Madhya Pradesh-462016

ISBN 978-93-5453-813-1