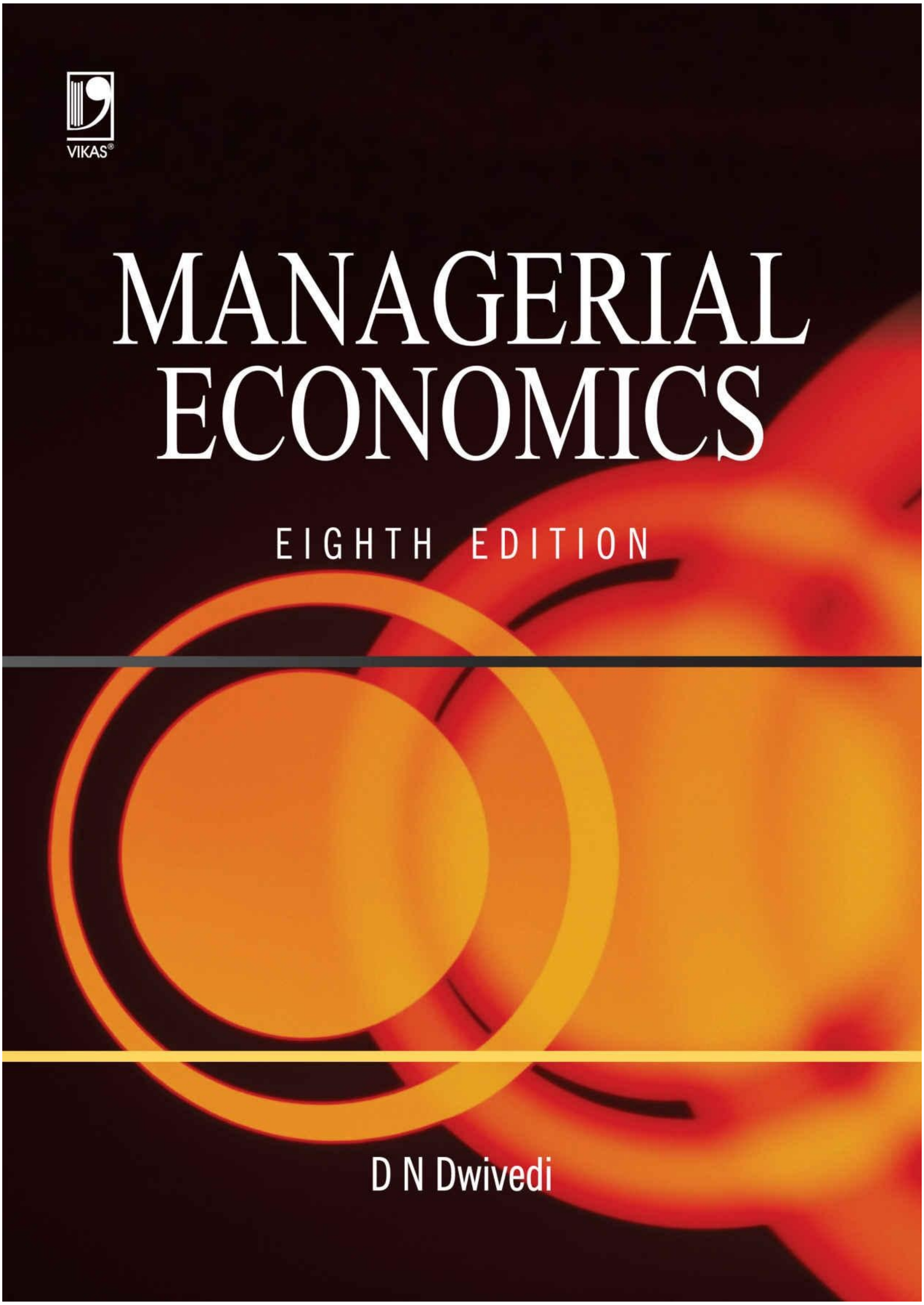
Managerial

Economics



Managerial

Economics

EIGHTH EDITION

D N DWIVEDI, Ph.D

Professor of Economics

Maharaja Agrasen Institute of Management Studies,

Delhi

Former Reader, Ramjas College

University of Delhi

VIKAS® PUBLISHING HOUSE PVT LTD

VIKAS® PUBLISHING HOUSE PVT LTD

E-28, Sector-8, **Noida**-201301 (UP) India

Phone: +91-120-4078900 • Fax: +91-120-4078999

Registered Office: 576, Masjid Road, Jangpura, **New Delhi-**110014. India

E-mail: ***helpline@vikaspublishing.com*** • Website: ***www.vikaspublishing.com***

• **Ahmedabad :** 305, Grand Monarch, 100 ft, Shyamal Road, Near Seema Hall,

Ahmedabad-380 051 • Ph. 079-65254204

• **Bengaluru :** First Floor, N.S. Bhawan, 4th Cross, 4th Main, Gandhi Nagar,

Bengaluru-560009 • Ph. +91-80-22204639, 22281254

• **Chennai**

: E-12, Nelson Chambers, 115, Nelson Manickam Road, Aminjikarai,

Chennai -600029 • Ph. +91-44-23744547, 23746090

• **Hyderabad :** Aashray Mansion, Flat-G (G.F.), 3-6-361/8, Street No. 20, Himayath Nagar,

Hyderabad-500029 • Ph. +91-40-23269992 • Fax. +91-40-23269993

• **Kolkata**

: 6B, Rameshwar Shaw Road, Kolkata-700014 • Ph. 033-22897888

• **Mumbai**

: 67/68, 3rd Floor, Aditya Industrial Estate, Chincholi Bunder,

Behind Balaji International School & Evershine Mall, Malad (West),

Mumbai-400064 • Ph. +91-22-28772545, 28768301

• **Patna**

: Flat No. 101, Sri Ram Tower, Beside Chiraiyatand Over Bridge,

Kankarbagh Main Road, Kankarbagh, Patna - 800 020, (Bihar)

Managerial Economics

ISBN: 978-93259-8668-8

Eighth Edition 2015

Dedicated to

The Memory of

My Mother

Preface to the Eighth Edition

With ever-changing business environment, more so because of globalization, business decision-making has become a very complex managerial function. The complexity of business decision-making is reflected by the growing need for integrating business decisions with a wide range of business issues related to market assessment, production planning, pricing, marketing, financing, human resource management and business planning taking in view the current business environment and economic prospects of the country. With growing complexity of business decision-making, the application of microeconomic theories and tools of analysis to business decision-making has been a well recognized phenomenon. With the speed of change of business environment, however, various universities, business schools and management institutes have revised their curricula of managerial economics to include the macroeconomic aspects of managerial decisions. The objective of this book has been to provide a comprehensive and rigorous analysis and application of both microeconomic and macroeconomic theories to business decision-making. Although the earlier editions of this book were well received and appreciated highly by the subject teachers and students, the book has been revised thoroughly with added interpretations of economic theories and concepts and their application to managerial decisions.

What Is New in This Edition

Although the organizational structure of the book as given in the earlier edition has been retained, many significant additions and chapter-wise changes have been made in this revised edition. Almost the entire text has been rewritten by adding some new features, new chapters, new topics and by enhancing clarity and precision to the explanations of economic concepts and theories. A greater emphasis has been laid on application of economic theories to business decision-making. Details of the new in this edition are as below.

New Chapters Added

The addition of new chapters is basically the result of reorganization of the subject matter of some chapters in the earlier editions of the book.

• The topics 'Market Demand' and 'Elasticity of Demand' were discussed earlier together in Chapter 8. In this edition, these topics have been discussed in a greater detail in two chapters—'Market Demand and Market Mechanism' in Chapter 7 and 'Elasticity of Demand' in Chapter 8. These topics now include 'Significance of Demand Elasticity' and 'Application of Elasticity in Pricing Decisions'.

• The 'Theories of Price and Output Determination' under different kinds of market conditions were discussed in the earlier editions in one chapter (Chapter 13) in a somewhat short version. On the advice of the reviewers of the book, subject teachers and the co-author, the discussion on the 'Theory of Price and Output Determination' has now been extended over ***four new chapters***: (i) Chapter 14: Price Determination under Perfect Competition; (ii) Chapter 15: Price and Output Determination under Monopoly; (iii) Chapter 16: Price and Output Determination under Monopolistic Competition, and (iv) Chapter 17: Price and Output Determination under Oligopoly. The purpose is to present the theories of price determination under different kinds of market conditions in a more exhaustive manner and with added explanation. Thus, four new chapters have been added to this edition of the book.

• In the Macroeconomics part of the book, a new chapter—Chapter 28: IS-LM Model—has been added with the purpose of presenting a complete analysis of macroeconomic system of the economy as a whole. Chapter 28 presents a brief analytical discussion on the integration of the working of the product and money markets. This aspect of macroeconomics has been discussed by taking in view the MBA syllabi of various universities and management institutes.

New Sections Added to Different Chapters

• Each chapter now gives *Chapter Objectives* to make students aware of what they are supposed to learn by going through the chapter.

• Two new sections have been added to Chapter 6: (i) Derivation of Demand Curve with Changing Marginal Utility of Money, and (ii) Why Demand Curve Slopes Downward to Right. The section on 'Comparison of Cardinal and Ordinal Utility Approaches' has been deleted as it does not serve any useful purpose in the current environment.

• A new section on 'Expansion Path' of the production based on modern concept has been added to Chapter 10. The derivation of the 'Expansion Path' shows the trend in production with increase in the scale of production.

• A new section on 'Price Determination in Long Run' under perfect competition has been added with an extended interpretation of the theory of price determination.

• In Chapter 15, a new section 'Equilibrium of Multi-plant Monopoly' has been added and the 'Deadweight Loss of Monopoly' has been re-explained with new examples of deadweight loss.

• In Chapter 18, a new section on 'Basis of Game Theory' has been added to show the importance of game theory in business decision-making under the condition of intensive competition.

• In Chapter 27, new sections on Kinds of Money, Functions of Money and Supply of Money have been added and sub-sections on Demand for Money have been revised thoroughly with added explanation.

• Two new sections: (i) 'Theory of Interest Rate Determination' and (ii) 'Monetary Sector Equilibrium', have been added in Chapter 27.

• A new section on 'Meaning and Extent of Globalization' has been added in Chapter 32 and sections remotely related to international business have been deleted.

• In Chapter 34, a new section on 'Current Foreign Trade Policy of India' has been added.

• Chapter 36 has been rewritten with greater details on various aspects of Balance of Payments and a new section on Equilibrium and Disequilibrium of BOP has been added.

• In Chapter 37 titled 'International Monetary System', a new section on 'The Current Role of the IMF' has been added.

• In Chapter 39 titled 'Monetary Policy', a new section on India's monetary policy has been added with a detailed discussion.

• The Chapter on 'Fiscal Policy' has been almost rewritten with a new section added on Meaning and Objective of Monetary Policy, Fiscal Instruments and India's Fiscal Policy.

• In Chapter 43, a new section on 'Current Scenario of CSR in India' has been added in this edition of the book.

Other Important Changes

In addition to the new chapters and sections, many significant revisions have been made in almost all chapters of the book. Some important ones are briefly noted down.

Introductory Chapter 1 has been almost rewritten with the objective of showing relevance and application of *managerial economics* with reference to real-life business decision-making problems. Chapter 2 dealing with business objectives has been revised by adding some business objectives. Schumpeterian theory of profit has been rewritten by adding a greater precision to the theory. The method of profit measurement has been discussed in a greater detail. Chapter 4 has been thoroughly revised by highlighting the need for and application of tools of economic analysis with added precision. The title of Chapter 6, 'Basics of the Market System and Market Equilibrium' has been changed to 'Market System and Market Equilibrium'. The section on 'Derivation of Individual Demand Curve' has been elaborated with added examples. While some unnecessary details have been deleted with the purpose of adding precision, some sections have been detailed with some added argument. Chapter 10 has been thoroughly revised and its content reorganised to put it in a more logical sequence. Major revisions made in this chapter include a detailed explanation of (i) long-run theory of production with added diagrams, and (ii) meaning and measurement of *marginal rate of transformation*.

In Chapter 14, the section on 'Price and Output Determination' has been revised with a greater theoretical orientation of price and output determination with a greater analytical clarity. In Chapter 16, Sweezy's model of oligopoly, collusive oligopoly model and joint profit maximization model have been revised thoroughly with added explanation and examples. The Williamson's model on 'Managerial Utility Maximization' discussed in Chapter 19 has been re-explained extensively with an additional diagram. Several important changes and additions have also been made in macroeconomics aspects of managerial economics. In Chapter 25, the section 'Definition of National Income' has been revised with the addition of some new concepts. Also, national income data has been updated. Chapter 31 on 'Theories of Inflation' has been reorganised by shifting the section on 'Desirability of Inflation' to a more appropriate place.

In Chapter 33 dealing with theories of international trade, the section on Introduction has been transformed to a new section 'The Basis of Foreign Trade' with current and relevant facts about modern trade. The section on Heckscher-Ohlin Theorem II has been revised with a detailed and theoretically sound analysis. Chapter 35 dealing with 'Balance of Payments' has been thoroughly revised for conceptual clarity and precision. In Chapter 38, the section on the 'Drawbacks of Free Market System' has been revised with additional empirical facts. The chapters on Monetary Policy (Chapter 39) and Fiscal Policy (Chapter 40) have also been revised with more recent and relevant data on India's monetary and fiscal policies. In Chapter 41, an introduction section has been added. Also, while dealing with India's industrial policy, the content has been revised with more appropriate and recent data. Besides, some extensive details which have lost their relevance have been deleted to make these chapters more relevant to managerial decision point of view.

Some more important revisionary changes have been made in this edition of the book. The introductory sections of almost all chapters have been revised to show the importance of the subject matter of the chapter and its relevance and applicability to business decision-making. At the end of each chapter, a section on *Summary* has been added in the end in the form of a short note for the benefit of students. 'Review Questions' of almost all chapters have been revised with addition of new questions and 'exercises'.

I am fully confident that this book will meet the requirement of the students of business management and also the manager in practice for their requirements of present times.

I am profoundly grateful to the editorial staff of Vikas Publishing House for getting the book reviewed by the subject experts. They provided me guidelines for reorganising and revising the text of the book. I express my gratefulness once again to all those who contributed to the earlier editions of the book. I would appreciate comments and suggestions from the readers of the book.

D N Dwivedi

dnd.dwivedi@gmail.com

Preface to the First Edition

The economic tools of analysis have of late gained a wide application to the process of business decision-making, possi bly, because modern business problems are so complex that a decision-maker's personal experience, intuition, insight, foresight and judgement alone are no longer adequate to provide an appropriate solution to complex business problems. The application of economic tools and logic to the business problems, not only reveals the behavioural pattern of economic entities and variables involved in business problems, but also helps in arriving at an optimum solution to the problem. It has therefore become essential for the decision-makers and for those who are associated with decision-making to possess at least a working knowledge of the relevant economic tools of analysis. Possibly in view of this, a course of study, *Managerial Economics* or applied economics, has been devised by assimilating the relevant pieces of economic theories.

Incidently, the scope of *Managerial Economics* is still in a state of flux and it may remain so because, in a dynamic society, the objective of business firms, outlook of management, and also the decision variables keep changing, necessitating a change in the subject matter of *Managerial Economics*. Many authors on the subject consider the relevant theories of microeconomics only as the subject matter of *Managerial Economics*. It is, however, difficult to justify the exclusion of those aspects of macroeconomics which very often figure in business decisions, from the purview of *Managerial Economics*. Besides, *Managerial Economics* syllabi of professional courses like M.B.A., C.A., I.C.W.A., I.C.S., and M.Com. examinations of certain Indian universities have been so framed that they include the relevant aspects of both micro and macroeconomics. Since this book has been written mainly to meet the requirements of students preparing for these examinations, it covers the relevant and important aspects of both micro and macroeconomic theories. Broadly speaking, this book deals with important aspects of Business Economics and Economic Environment and Policy.

The purpose of this book is to provide, in one volume, the various economic theories which are deemed to constitute the subject-matter of *Managerial Economics*. This book is intended to explain in non-technical language, the economic concepts, tools of analysis and their relevance to business decision-making, and also the influence of economic environment on business decisions. Attempt has also been made to illustrate the solutions to such business problems as optimization of output-mix, input combination, profit maximization and choice of investment projects. In writing this book, I have drawn heavily from scholarly writings of economists.

I acknowledge my indebtedness to my friend Shri M.C. Kuchhal of Shri Ram College of Commerce, University of Delhi, for his constant persuasion and encouragement. I am also grateful to my friend Dr. J.L. Raina of M.M.H. College, Ghaziabad, for his immense help in preparing the book for publication.

D N Dwivedi

Contents

Part I

INTRODUCTION

The objective of Part I of the book is to introduce managerial economics with its nature and scope and equip the readers with some basic economic concepts and analytical tools used in economic analysis and in business decision making. The subject matter of Part I is divided in four chapters.

CHAPTER

1 Nature and Scope of Managerial Economics

CHAPTER OBJECTIVES

The objective of this chapter is to introduce Managerial Economics as a subject of study. By going through this chapter, you will know:

- What economics is about
- What is managerial economics
- Economics concepts, theories and tools of analysis that constitute managerial economics
- How economics can be applied to managerial decision-making and
- Nature and scope of managerial economics

1.1 INTRODUCTION

Managerial Economics has emerged as a separate branch of economics. The emergence of managerial economics can be attributed to at least three factors: (*i*) growing complexity of the business environment and decision-making process; (*ii*) increasing application of economic logic, concepts, theories and tools of economic analysis in the process of business decision-making; and (*iii*) rapid increase in demand for professionally trained managerial manpower with good knowledge of economics. The growing complexity of the business world can be attributed to rapid growth of large scale industries, increasing number of business firms, quick innovation and introduction of new products, globalization and growth of multinational corporations, merger and acquisition of business firms, and large-scale diversification of business activities. These factors have contributed a great deal to the inter-firm, inter-industry and inter-country business rivalry and competition, enhancing uncertainty and risk in the business world.

Business decision-making in this kind of complex business environment has become a very complex affair. There was a time when family training, personal experience and business acumen were sufficient to make good business decisions and run an organization successfully. In today's business world, however, personal experience, knowledge and family training are no longer sufficient to meet the managerial challenges of the modern business world, though one can find a number of reputed businessmen[1] with no management training.

1. In his article "Dropout Champs *(Economic Times,* 20 March 2009), Nikhil Menon has cited a number of cases of reputed entrepreneurs who are college dropouts, let alone academic training in business management. Nevertheless, the importance of management training cannot be denied. Dhirubhai Ambani, a reputed entrepreneur, got his sons (Mukesh Ambani and Anil Ambani, world's top class businessmen) trained in business management in the US. In India today, demand for MBA degree by the persons with working experience has increased substantially. According to a survey report *(Business Lines,* August 5, 2013), 81 per cent MBA students for 2013–15 batch in MDI (Gurgaon) and HM-Bangalore and 71 per cent students in IIM-Calcutta have working experience of at least 2 years.

Making a reasonable business decision under the condition of growing complexity of business conditions has inevitably increased the application of economic concepts, theories and tools of analysis in business decision making. The reason is that making a reasonable business decision requires a good knowledge and comprehensive understanding of the business environment, production and cost conditions, the nature and degree of competition, and fundamentals of market mechanism. This requires an intensive and extensive analysis of the product, input and financial market conditions.

A systematic analysis of market conditions and business environment requires collecting large-scale data on demand and supply conditions, cost of production, pricing system, and morphology of market, the nature and degree of competition and so on. It is in this context of the managerial function that economics contributes a great deal. It is for this reason that managerial economics has emerged as an important aspect of management studies.

Before we proceed to discuss the nature and scope of managerial economics, let us first know "what economics is about".

1.2 WHAT IS ECONOMICS?[2]

Economists of different generations have defined economics in different ways according to their perception and subject matter of economics. For example, according to Adam Smith (1976), the "father of economics", economics is "an inquiry into the nature and causes of the wealth of nations"[3]. According to Alfred Marshall, (1922), an eminent economist of the neo-classical era, "Economics is the study of mankind in the ordinary business of life; it examines that part of individual and social actions which is most closely connected with the attainment and with the use of the material requisites of well-being."[4] Lionel Robbins (1932) defined economics more specifically: "Economics is the science which studies human behaviour as relationship between ends and scarce means, which have alternative uses."[5] One can find a number of other definitions of economics in economics literature. However, none of the definitions reveal the entire subject matter of the modern economics. Nevertheless, these definitions do reveal the essence of economics as a social science. The definitions given above, can be used to understand 'what economics is about'.

Going by Marshall's definition, economics is the "study of mankind in the ordinary business of life." What is the "ordinary business" of mankind? The 'ordinary business' of mankind is to work and earn their livelihood by using their resources and to improve their economic welfare. This kind of activity is *economic behaviour*. As Robbins has pointed out, human wants are endless but human resources are limited and have alternative uses. According to him, economic behaviour is making choice between the endless wants and between the alternative uses of their limited resources for meeting maximum number of their needs.

Economics can, thus be defined as a social science that studies economic behaviour of the people, the individuals, households, firms, and the government. *Economic behaviour* is essentially *economizing behaviour*. Economizing behaviour is the effort of the people to derive maximum gain from the use of their limited resources—land, labour, capital,

2. Students having done graduation with economics may skip this section.

3. Adam Smith, *The Wealth of Nations* (1776).

4. Alfred Marshall, *Principles of Economics,* 8th Edn. (Macmillan, New York, 1922), p. 1.

5. Lionel Robbins, *The Nature and Significance of Economic Science,* 2nd Edn. (Macmillan, London, 1935), p. 15.

time and knowledge, etc., which have alternative uses. Technically, the term 'economizing' means deriving maximum gains from a given cost and alternatively minimizing cost for a given gain. This is economizing behaviour—a natural behaviour.

Why do People Economize? People tend to adopt economizing behaviour because of the following facts of economic life of human beings.

1. ***Human wants, desires and aspirations are endless.*** Human wants are endless is the sense that they go on increasing with the availability of new kinds of goods and services and increase in ability to pay.

2. ***Resources available to the people are scarce.*** Resources (labour, land, capital, time and knowledge, etc.) available to the people at any point on time are scarce and limited; though they have alternative uses.

3. ***People are by nature economizers.*** Attempt to economize is a natural behaviour of the

people. For example, if metro service is easily available, one would not like to hire a taxi, and if a pizza is available at `50 in college canteen and for `100 in a nearby restaurant, students would prefer to eat pizza in the canteen. This is economizing behaviour.

In order to maximize their gains, people allocate their resources between their competing wants in such a way that their total gain is maximized. Thus, *economics as a social science studies how people allocate their limited resources to their alternative uses with the objective of deriving maximum possible gains from the use of their resources.*

For analysing economic behaviour of the people, economists use certain specific concepts, logic, tools of analysis and maximization techniques. The ultimate result of this kind of analysis is the formulation of economic theories.

Microeconomics and Macroeconomics

Economics as a social science has two major branches — *microeconomics* and *macroeconomics. Microeconomics* is the study of the economizing behaviour of the individual economic entities—individuals, households, firms, industries and factory owners. For example, *microeconomics* studies how *individuals* and *households* with limited income decide ‘what to consume’ and ‘how much to consume’ so that their total utility is maximized. In other words, microeconomics studies how individual consumers make choice of goods and services they want to consume and how they allocate their limited income between the goods and services of their choice to maximize their total economic welfare.

Similarly, in case of *business firms,* microeconomics studies how business firms take decision on what to produce, for whom to produce, how to produce, how much to produce, what price to charge, how to fight the competition, and how to promote sales to maximize their profit. In addition, microeconomics studies how demand for a product and its supply are affected by the change in price of the product and how price of a product is determined in the market. Obviously, decision making process in case business firms is a complex affair. Therefore, making a reasonable business decision requires a good knowledge and understanding of market conditions in respect of demand and supply conditions, cost of production, pricing system, nature and degree of competition, conditions in the financial market, and so on. This requires an extensive and intensive analysis of the market conditions. This is subject matter of *microeconomics.*

Macroeconomics, on the other hand, studies the economic phenomena at the national aggregate level. Specifically, macroeconomics is the study of working and performance of the economy as a whole. It studies what factors and forces determine the level of national output or national income, rate of economic growth, employment, price level, and economic welfare. Besides, macroeconomics studies how government of a country formulates its macroeconomic policies—taxation and public expenditure policies (the fiscal policy), monetary policy, price policy, employment policy, foreign trade policy, etc., to resolve the problems of the country. Macroeconomics studies how these policies affect the economy. From *managerial economics* point of view, the study of macroeconomics gives basis for judging the *economic environment* of the country. How the knowledge of business environment contributes to business decision making will be discussed later.

The scope of microeconomics and macroeconomics and their place and role in managerial economics will be discussed later.

1.3 WHAT IS 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’’6.

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.”7

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."8

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."9

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

6. Mansfield, E., *Managerial Economics and Operations Research* (W.W, Norton & Co. Inc., New York, 1966), p. 11.
7. Spencer, M. H. and Seigelman, L., *Managerial Economics* (Irwin, Illinois, 1969), p. 1.
8. Davis, R. and Chang, S., *Principles of Managerial Economics* (Prentice-Hall, N.J., 1986), p. 3.
9. Douglas, Evan J., ***Managerial Economics: Analysis and Strategy*** (Prentice-Hall, N.J., 1987), p. 1.

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.

Managerial Decision Areas

• Choice of business area given the resources
• Choice of product
• Determining optimum output
• Choice of technology—factor combination
• Acquisition of inputs—labour and capital
• Determining price of the product
• Assessing economic environment and business scope

Economic Science

Quantitative Methods

Microeconomics and
• Mathematical Tools
Macroeconomics:
• Statistical Tools
Economic Principles, Theories,
• Game Theory
Concepts and Tools of Analysis
• Econometrics

Managerial Economics

Application of economic theories and quantitative methods to find solution to business 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.

1.4 HOW ECONOMICS CONTRIBUTES TO MANAGERIAL DECISIONS

The primary function of managers is to take appropriate decisions and implement them effectively to achieve the objective of the organization to maximum possible extent, given the resources. Application of economics contributes a great deal to managerial decision-making as it provides guidance in finding an appropriate solution to the business problem. Just as biology contributes to medical profession and physics to engineering, economics contributes to managerial functions. As such, a working knowledge of economics is essential for managers. Managers are, in fact, practicing economists.

Let us now see how economics contributes to managerial decisions. All the areas of managerial decisions, as noted in Block 1 of Figure 1.1, have economic perspective. Therefore, economic theories, concepts and tools of analysis are applied as roadmap to find solution to business problems. It has been found empirically that application of economic theories and tools of analysis makes significant contribution to the process of business decision making in many ways.

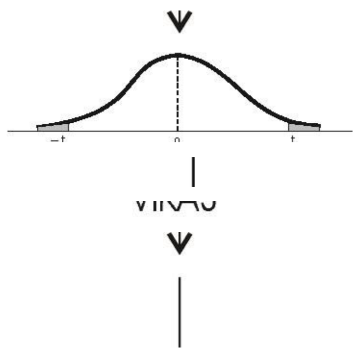
According to Baumol,10 a Nobel laureate in economics, economic theory contributes to business decision making in *three important ways.*

First, 'one of the most important things which the economic theory can contribute to management science' is providing framework for building analytical models which can help recognize the structure of managerial problem, determine the important factors to be managed, and eliminate the minor factors that might obstruct decision making.

Secondly, economics provides 'a set of analytical methods' which may not be directly applicable to analyse specific business problems but they do widen the scope of business analysis and enhance the analytical capability of the business analyst in understanding the nature of the business problems.

Thirdly, various economic terms are used in common parlance, which are not applicable to business analysis and decision making. Economic theory offers clarity to various economic concepts used in business analysis, which enables the managers to avoid conceptual pitfalls. For example, in general sense, 'demand' means quantity demanded at a point of time. But, in economic sense, 'demand' means the quantity people are willing to buy at a given price and they have ability and willingness to pay.

10. Baumol, William J., "What Can Economic Theory Contribute to Managerial Economics", *American Economic Review,* Vol. 51, No. 2, May 1961.



Apart from providing analytical models and methods and conceptual clarity, economics contributes to business decision in many other ways also. Most business conditions are taken under the condition of *risk* and *uncertainty*. Risk and uncertainty arise in business because of continuous change in business conditions and environment, and unpredictable market behaviour. Economics provides models, tools and technique to predict the future course of market conditions, ways and means to assess the risk and, thereby, helps in business decision making.

It is because of these important contributions of economics to business decision making that economics has been integrated with managerial decisions. Managerial decision making without applying economic logic, theory and analytical tools may not offer a reasonable solution.

1.5 APPLICATION OF ECONOMICS TO BUSINESS DECISION: AN EXAMPLE

We have discussed above, in general terms, of course, how economics contributes to business decision making. In this section, we illustrate the application of economic theories and analytical tools to business decision making with a hypothetical example.

Suppose a mobile phone company plans to introduce a new brand with extra facilities, say, making calls with spoken numbers and talking face-to-face to the contacted person. In order to take the final decision, the company will have to go through a process of decision making. In general, the decision making process[11] is comprised of first four main stages as shown in Fig. 1.2.

1: Determining and Defining Objective

2: Identifying Business Related Issues

3: Collections and Analysis of Relevant Data

4: Inventing, Developing Possible Course of Action

5: Selection of the Best Solution

6: Implementation of the Decision

Fig. 1.2 *Decision-Making Process*

11. See also Simon, Herbert A., "The Decision-Making Process", in Mansfield, E. (ed), *op. cit.*

The firm has already determined its objective—to introduce a new product. Thus, the stage one is already accomplished. The decision-making problem arises at stage 2. Identifying business related issues and finding alternative solutions to the problems is a very crucial area of business decisions. The business issues of the mobile phone company can be classified under the following two categories.

1. Production related issues and
2. Sales-related issues

Production and sales related issues have a very wide perspective. For example, *production related issues* can be identified as listed below.

(*a*) Available techniques of production of the planned product,

(*b*) Cost of production associated with each production technique,

(*c*) Supply position of required inputs,

(*d*) Price structure and affordability of input prices,

(*e*) Supply position and price of competitive substitute,

(*f*) Availability of foreign exchange if imported inputs are required.

Similarly, *sales related managerial issues* include making assessment of at least the following aspects of the market conditions.

(*a*) Market size, demand prospects and general market trends,

(*b*) General trend in the introduction of new products and their sales,

(*c*) Major existing and potential competitors and their market share,

(*d*) Prices of the competitive products,

(*e*) Pricing strategy of competitors of the substitute products, and

(*f*) Market structure and degree of competition.

Decision making by the mobile phone companys on all these managerial decision issues requires collection and analysis of relevant data. It is in this kind of managerial function that economic theories and analytical tools contribute a great deal in managerial decision making. Application of economic theories and analytical tools facilitates decision making in the following ways.

(*i*) It helps in making choice of the variables in regard to each business issue for collection and analysis of data. Collection of relevant and adequate data provides a sound basis for analysing the issue at hand.

(*ii*) Economic theories provide a logical basis for determining the nature and extent of relationship between the interrelated and interdependent variables. For example, theory of demand states that there is inverse relationship between demand and price and change in demand with change price depends on the elasticity of demand. The elasticity of demand guides in pricing decision—whether or not to change the price. Similarly, theory of production states that production increases with increase in inputs and production may increase at increasing rate, decreasing rate or at constant rate. The estimated input-output relationship guides in taking a decision on whether or not increase the production.

(*iii*) All business firms have certain objective(s) to achieve, given their limited resources. Economic theories combined with economic tools of analysis provide a sound basis for determining the optimum levels of inputs and output, and a sustainable price

given the market conditions. In case profit maximization is the firm's objective, economic theories and mathematical tools help in determining the profit maximizing production and price in view of the nature of competitive market.

It may thus be concluded that the application of economic theories and tools of analysis ensures consistency of business analysis and streamlines the decision-making process. Not only that, it ensures the validity of the decisions taken.

1.6 THE SCOPE OF MANAGERIAL ECONOMICS

The scope of managerial economics is comprised of economic concepts, theories and tools of analysis that can be applied in the process of business decision making to analyse business problems, to evaluate business options, to assess the business prospects, with the purpose of finding appropriate solution to business problems and formulating business policies for future. As noted above, *economic science* has two major branches, viz., *microeconomics* and *macroeconomics.* Both Microeconomics and Macroeconomics are applied to business analysis and business decision making depending on the nature of the issue to be examined. Managerial decision issues can be divided broadly under two broad categories: (*a*) *Internal managerial issues* and (*b*) *External environmental issues.* In brief, microeconomic theories and analytical tools are applied to *internal managerial* or *operational issues* and macroeconomic theories and analytical techniques are applied to assess the *external and environmental issues.* Let us now look at the nature and areas of internal managerial issues and environmental issues and how micro and macro-economic theories are applied to resolve the business issues.

1.6.1 Microeconomic Theories Applied to Internal Issues

Internal managerial issues refer to decision-making issues arising in the management of the firm. Internal managerial issues include problems that arise in operating the business organization. All such managerial issues fall within the purview and the control of the managers. Some of the basic internal management issues can be listed as follows.

- What to produce—choice of the business
- How much to produce—determining the size of the firm
- How to produce choice of efficient and affordable technology
- How to price the product—determining the price of the product
- How to promote sale of the product
- How to face price competition from the competing firms
- How to enlarge the scale of production—planning new investment
- How to manage profit and capital.

The microeconomic theories and tools of analysis that provide a logical basis and ways and means to find a reasonable solution to business problems constitute the microeconomic scope of managerial economics. The main microeconomic theories that fall within the scope of managerial economics are following.

Theory of consumer demand: Theory of consumer demand analyses the decision -making behaviour of the consumers. The decision-making behaviour of the consumer relates to such questions as: how consumers decide what to consume; how much to consume; how

much to buy, and how consumers react to change in price of the products they consume and price of their substitutes. Demand theory combined with quantitative tools helps in assessing the total demand for a product at different prices. Thus, the consumer demand theory helps in deciding 'what to produce'.

Theory of production: Theory of production analyses the nature of input-output relationship. It explains how output changes with change in inputs—labour or capital—given the technology. It provides guidance in the choice of technology and in maximizing the output from the resources of the firm. Thus, the knowledge and application of the theory of production helps in determining the optimum level of production, the size of the firm, and the employment of labour and capital.

Theory of cost: Theory of cost analysis the nature and pattern of change in cost of production with change in output. Specifically, theory of cost reveals the change in total, marginal and average cost of production. Application of cost theory helps in knowing the cost behaviour with increase in production and in determining the output that minimizes the average cost production. In view of profit-maximization objective, this theory helps in determining the profit maximizing output, give the price of the product.

Theory of price determination: The theory of price determination offers an analysis of how price is determined under different kinds of market conditions. Market conditions are determined on the basis of degree of competition between the firms of the industry—perfect competition, monopolistic competition, oligopoly and monopoly. The theory of price determination combined with the cost theory helps firms in determining the profit maximizing price of their product.

Theory of capital and investment decisions: Capital is the foundation of business firms. An efficient management of capital is one of the most important functions of the managers as it is the determinant of the success of the firm. The major issues in capital management are (*i*) the choice of investment avenues, (*ii*) assessing the efficiency and productivity of capital investment avenues, and (*iii*) making the choice of most efficient investment project. The theory of capital contributes a great deal in making appropriate investment decisions.

1.6.2 Macroeconomics Applied to Business Decision

As noted above, macroeconomics is the study of economic conditions of the economy as a whole whereas a firm is a small unit of the economy. As such, macroeconomic theories are not directly applicable to managerial decisions. However, business managers, while making business decisions, cannot assume the economic conditions of the country to remain the same for ever. As a matter of fact, economic conditions of the country keep changing. Changing economic conditions change the economic environment of the country, and thereby business environment and business prospect. And, as management experts Weihrich and Koontz point out, "... managers cannot perform their task well unless they have an understanding of, and are responsive to the many elements of the external environment—economic, technological, social, political, and ethical factors that affect their areas of operations."12 Therefore, while making business decisions, managers

12. Heinz and Harold Koontz, Management—A Global Perspective, 10th edn, (Tata McGraw-Hill, New Delhi, 1993), p. 5.

have to take into account the economic environment of the country. The factors which, in general, determine the economic environment of a country are (*i*) the general trend in national income (GDP), saving and investment, prices, employment, etc., (*ii*) the structure and role of the financial institutions, (*iii*) the level and trend in foreign trade, (*iv*) economic policies of the government, (*v*) socio-economic organizations like trade unions, consumer associations, and (*vi*) political environment.

It is far beyond the powers of a single firm, howsoever large it may be, to determine the course of economic, political and social conditions of the country. But the *environmental factors* have a far reaching bearing on the functioning and performance of the business firms. Therefore, it is essential for business decision-makers to take in view the present and future economic environment of the country. It is essential because business decisions taken ignoring the environmental factors may not only fail to produce the result but may also cause heavy losses. This is what happened in case of establishing a SEZ in Nandigram and Tata's small car project in Singrur district of West Bengal in 2009–10. Therefore, while taking decision regarding forward planning and programmes on (*a*) expanding the scale of production, (*b*) setting up new plants, and (*c*) introduction of new products. The major macroeconomic issues and factors that business managers are supposed to take in view while making decisions with long-term implications are described here briefly.

Macroeconomic Factors

The major macroeconomic environmental factors that figure in business decisions, especially those related to forward planning and formulation of strategy, may be described under the following three categories.

1. ***Trend in the Economy*** The overall trend in the economy is determined by the trend in macroeconomic factors, viz., Gross national product (GNP), economic growth rate, saving and investment, saving and investment, general price level, employment, and investment climate. The trend in these factors determines the future prospect of business corporations. For example, economic slowdown in the Indian economy and a high rate of inflation during 2012–12 has led to decline in demand and hence decline in industrial production from 4.3 per cent in 2012 to 2.2 per cent in Q2 of 2013. Therefore, firms planning to expand their size or to set up a new unit take the view of the general trend in the economy and demand level and its effect on the demand for their product. A comprehensive view of these factors can be obtained by using macroeconomic concepts and theories.

2. ***International Economic Conditions*** Business corporations involved in foreign trade and financial transactions are affected directly and more heavily than the rest of the domestic economy by the fluctuations in the international economy. Fluctuations in the international trade, exchange rate and financial flows affect the economy of inter related countries and, thereby, the business corporations. For example, the 'sub-prim financial crisis' of the US in 2008–09 led to collapse of the European countries and caused decline in growth rate of China and India. The decline in growth rate has adverse impact on the domestic industries. Therefore, it is in the interest of the business corporations having international transactions to take a comprehensive view of trends in the global economy. International economics which constitutes a branch of macroeconomics provides guidelines for assessing the international economic conditions.

3. ***Government Policies*** Government's economic policies, especially monetary, fiscal, industrial and foreign trade policies, play a crucial role in determining the internal economic environment of the country. Economic policies affect functioning of the business firms adversely or favourably depending on policy objectives. Therefore, while taking business decisions, business managers are required to have the knowledge of how government's economic policies can affect their business. Macroeconomic theories provide the theoretical framework and analytical models to measure the effects of government policies.

In brief, the scope of managerial economics consists of economic theories, tools of analysis, concepts and methodology that can be applied to managerial decision making. Economics has two major branches—microeconomics and macroeconomics. The main economic theories and tools of analysis of both microeconomics and macroeconomics constitute the subject matter of managerial economics. *Microeconomic theories and analytical tools* are applied to business decision pertaining to internal management problems like the choice of product, technology, quantity of product, price of the product, scale of production in future, method of facing competition and promotion of sales, etc. Macroeconomic theories and analytical methods are applied to external environmental issues like assessing the present and future economic conditions of the country, growth prospects, price level and business prospects.

1.7 SOME OTHER DISCIPLINES OF MANAGERIAL ECONOMICS

As mentioned above, the scope of managerial economics consists mainly of micro and macro economic theories, concepts, and tools of analysis that can be applied to managerial decision making. There are, however, certain other disciplines which contribute to quantitative economic analysis of business problems and hence to business decision making. Therefore, managerial economics includes the study of the disciplines that contribute to managerial decisions. The disciplines on which economics draws heavily are mathematical tools, statistics, operations analysis, management theory and accountancy. How these disciples contribute to managerial economics are described here briefly.

1.7.1 Mathematical Tools

Business managers deal primarily with variables that are essentially quantitative in nature, e.g., demand, supply, price, cost, product, capital, wages, interest rate, inventories, etc. These variables are interrelated, directly or indirectly. What is needed in economic analysis is to have clarity of these concepts in order to have, as far as possible, accurate estimates of these economic variables and a measure of their relationships. The use of mathematical tools in the analysis of economic variables provides not only clarity of concepts, but also a logical and systematic framework for measuring the quantitative relationships between the relevant variables. More importantly, mathematical tools are widely used in 'model' building, for exploring the relationship between related economic variables. Mathematical logic and tools are, therefore, a great aid to economic analysis.

Furthermore, a major problem that managers face is how to minimize cost, maximize profit or optimize sales under certain constraints. Mathematical concepts and techniques are extensively used in economic analysis with a view to finding answers to these questions. Besides, certain mathematical tools and optimization techniques, relatively more

sophisticated and advanced, designed during and after World War II have found wide ranging application to business management, viz., linear programming, inventory models and game theory. A working knowledge of these techniques and other mathematical tools is essential for managers. These topics, therefore, fall very much within the scope of managerial economics. The tools of analysis and optimization techniques widely used in the process of decision-making will be discussed in Chapters 3 and 4.

1.7.2 Statistics

Like mathematical tools, statistical tools provide a great aid in business decision-making. Statistical techniques are used in collecting, processing and analyzing business data, testing the validity of economic laws with the real life economic data before they are applied to business analysis. A good deal of business decisions are based on probable economic events. The statistical tools, e.g., theory of probability, and forecasting techniques help the decision-makers in predicting the future course of economic events and probable outcome of their business decisions. Regression technique is a widely used statistical method to measure the relationship between the related variables. Thus, the scope of business economics includes also the study of statistical tools and techniques that are applied to analyze the business data and to forecast economic variables. The mathematical and statistical techniques are the tools in the armoury of decision-makers that solve the complex problems of business.

1.7.3 Operations Research (OR)

Operations Research (OR) is an inter-disciplinary technique of finding solutions to managerial problems. It combines economics, mathematics and statistics to build models for solving specific business problems and to find a quantitative solution thereto. Linear programming and goal programming are two widely used OR techniques in business decision-making.

The scope of managerial economics is, thus, very wide. It is difficult to do justice to the entire subject matter of managerial economics in one volume. In this book, we have covered only that part of microeconomic theory which has direct application to business decisions, as mentioned above. In addition, some broad aspects of macroeconomic theory, international trade and government policies, which often figure in business decision-making have also been covered.

1.8 MANAGERIAL ECONOMICS BRIDGES THE GAP BETWEEN THEORY AND PRACTICE

We have noted above that application of theories to the process of business decision-making contributes a great deal in arriving at appropriate business decisions. In this section, we highlight the gap between the theoretical world and the real world and see how managerial economics bridges this gap.

1.8.1 Theory *vs.* Practice

It is widely known that there exists a gap between theory and practice in all walks of life, more so in the world of economic theory and behaviour. A theory which appears logically sound may not be directly applicable in practice. For example, if there is a

constant relationship between inputs and outputs, it may be theoretically concluded that if inputs are doubled, output will be doubled and if inputs are trebled, output will be trebled. This theoretical conclusion may not hold in practice. This gap between theory and practice has been very well illustrated in the form of a story by a classical economist, J.M. Clark.[13] He writes:

'There is a story of a man who thought of getting the economy of large scale production in plowing, and built a plow three times as long, three times as wide, and three times as deep as an ordinary plow and harnessed six horses [three times the usual number] to pull it, instead of two. To his surprise, the plow refused to budge, and to his greater surprise it finally took fifty horses to move the refractory machine... [and] the fifty could not pull together as well as two'.

The gist of the story is that managers—assuming they have abundant resources—may increase the size of their capital and labour, but may not obtain the expected results. Most probably the man in Clark's story did not get the expected result because he was either not aware of or he ignored or could not measure the resistance of the soil to a huge plow. This incident clearly shows the gap between theory and practice.

The gap between theory and practice can be explained as follows. Let us first look at the reality of economic life. The *real economic world is extremely complex*. The reason is that economic decisions and economic activities of economic entities—individuals, households, firms, and the government are interconnected and interdependent. Change in any one important economic variable generates a wave of changes, beginning with a change in the directly related areas, which create counter-changes. In economic system, a change in one economic variable causes change in a large number of related variables. As a result, the entire economic environment changes. An alterning economic environment changes people's economic goals, motivations and aspirations which, in turn, change managers' decisions. In fact, decision-making becomes a continuous process. The entire system looks 'hopelessly chaotic'. Under the condition of changing environment and economic decisions, it is extremely difficult to predict human behaviour.

On the contrary, *economic theories are rather simplistic* because they are propounded on the basis of economic models built on simplifying assumptions—most economic models assume 'other things remain constant'. In fact, through economic models, economists create a simplified world with its restrictive boundaries. It is from such models that they derive their own conclusions and formulate their theories. It is another thing that some economic, rather econometric, models are more complex than the real world itself. Although economic models are said to be an extraction from the real world, how close it is to reality depends on how realistic are the assumptions of the model. The assumptions of economic models are often claimed to be unrealistic. The most common assumption of the economic models is the *ceteris paribus* assumption, i.e., all other things remain constant. For example, consider the law of demand. It states that demand for a commodity decreases when its prices increases and *vice versa*, *other things remaining constant*. The 'other things' include consumers' income, prices of substitute and complementary goods, consumer's tastes and preferences, advertisement, consumer's expectations about the commodity's future price, 'demonstration effect', and 'snob effect', etc. In reality, however,

13. *Studies in the Economics of Overhead Costs*, University of Chicago, 1923, p. 116, quoted in K.K. Seo, *Managerial Economics, Text, Problems, and Short Cases*, (Richard, D. Irwin, Inc., Homewood, Ill, 1984), p. 325.

these factors do not remain constant. Since 'other things' do not remain constant, the *ceteris paribus* assumption is alleged to be the most *unrealistic assumption*. Nevertheless, the law of demand does state the nature of relationship between the demand for a product and its price, in isolation of the factors determining the demand for that product.

Economic theories are, no doubt, *hypothetical in nature but not away from reality*. Economic theories are, in fact, a caricature of reality under certain specified conditions. In their abstract form, however, they do look divorced from reality. Besides, abstract economic theories cannot be straightaway applied to real life problems. This should, however, not mean that economic models and theories do not serve any useful purpose. 'Microeconomic theory facilitates the understanding of what would be a hopelessly complicated confusion of billions of facts by constructing simplified models of behaviour, which are sufficiently similar to the actual phenomenon and thereby help in understanding them'.[14] Nevertheless, it cannot be denied that there is apparently *a gap between economic* *theory and practice*. This gap arises mainly due to the inevitable gap between the abstract world of economic models and the real world.

1.8.2 How Managerial Economics Fills the Gap

There is undeniably a gap between economic theory and the real economic world. Therefore, economic theories do not offer a custom-made or readymade solution to business problems. However, economic theories do provide a framework for logical economic model-building and systematic analysis of economic issues. The need for such a framework arises because the real economic world is too complex to permit considering every bit of relevant facts that influence economic decisions. In the words of Keynes, 'The objective of [economic] analysis is not to provide a machine, or method of blind manipulation, which will furnish an infallible answer, but to provide ourselves with an organized and orderly method of thinking out particular problem…'.15 In the opinion of Boulding, the objective of economic analysis is to present the 'map' of reality rather than a perfect picture of it.16 In fact, economic analysis equips us with a road map; it guides us to the destination, but does not carry us there. This is how managerial economics bridges the gap between economics and business decision-making. As an example, **managerial economics** can also be compared with medical science. Just as the knowledge of medical science helps in diagnosing the disease and prescribing an appropriate medicine, managerial economics helps in analyzing the business problems and in arriving at an appropriate decision.

Let us now see how managerial economics bridges this gap. On one side, there is the complex business world and, on the other, are abstract economic theories. 'The big gap between the problems of logic that intrigue economic theorists and the problems of policy that plague practical management needs to be bridged in order to give executive access to the practical contributions that economic thinking can make to top management policies'.17 Managerial decision-makers deal with the complex, rather chaotic, business

14. Lerner, A.P. "Microeconomic Theory" in *Perspective in Economics— Economists Looks at Their Field of Study by* A.A. Brown, E. Neuberger, and M. Pakmastier (eds.), (McGraw-Hill, NY), p. 36.

15. Keynes, J.M. *The General Theory of Employment Interest and Money*, (Harcourt Brace, New York, 1936), p. 297.

16. Boulding, K.E. *Economic Analysis*, (Harper and Bros., New York, 1948), p. 14.

17. Dean, Joel *Managerial Economics*, (Prentice Hall of India, New Delhi, 1977), p. vii.

conditions of the real world and have to find the way to their destination, i.e., achieving the goal that they set for themselves. Managerial economics applies economic logic and analytical tools to sift wheat from the chaff. Economic logic and tools of analysis guide business managers in the following ways.

(*i*) Identifying their problems in achieving their goal,

(*ii*) Collecting the relevant data and related facts,

(*iii*) Processing and analyzing the facts,

(*iv*) Drawing relevant conclusions,

(*v*) Determining and evaluating the alternative means to achieve the goal, and

(*vi*) Taking a decision.

Without application of economic logic and tools of analysis, business decisions are most likely to be irrational and arbitrary, which may often prove counter-productive.

SUMMARY

• Economics is a social science which studies economic behaviour of the people, i.e., how people maximize their gains from the use of their limited resources.

• Managerial economics is the study of economic laws, theories, concepts and tools of analysis that can be applied to business decision-making. Managerial economics is basically applied economics.

• In reality, there is gap between the business theory and business practice. Managerial economics fills the gap between the theory and practice.

• Economics contributes to business decision-making in three important ways: (*i*) it provides framework for building an appropriate model to analyze business issues, (*ii*) it provides a set of analytical methods to analyze the business problems and to derive a reasonable conclusion, (*iii*) it provides a sound basis for determining the data required for analyzing business problems.

• The scope of managerial economics includes both microeconomics and macroeconomics.

Microeconomics deals with internal issues of the business firms, such as what to produce, how to produce, how much to produce, how to determine price, etc. Macroeconomics deals with economic environmental issues like economic conditions of the country, expected trend in the economy, the trend in the general price level, government economic policies and their probable impact on business, the likely impact of international economic conditions on domestic business, etc.

• Some other disciplines linked to managerial economics are mathematical tools, statistics, operation research, and management theory.

REVIEW QUESTIONS

1. Define managerial economics and disucss its scope. What is importance of managerial economics.
2. What is the nature and scope of managerial economics? How does managerial economics contribute to business decision making?
3. Managerial economics is the discipline which analyses the application of 'economic theory to business management'. Comment.
4. What are the major areas of business decision-making? How does economic theory contribute to managerial decisions?

5. Discuss the nature and scope of managerial economics. What are the other related disciplines?
6. "Managerial economics bridges the gap between economic theory and business practice". Explain with examples.
7. Managerial economics is essentially the application of microeconomic theory of business decision-making. Discuss the statement.
8. What are the related topics other than microeconomic theories in managerial economics? How do they contribute to managerial economics?
9. "Managerial economics is applied microeconomics". Elucidate.
10. What are the basic functions of a manager? How does managerial economics contribute to business decision-making?
11. "Managerial economics is the integration of economic theory with business practice for the purpose of facilitating decision-making and forward planning by the management." Explain.
12. How does the study of managerial economics help a business manager in decision-making? Illustrate your answer with examples from production and pricing issues.
13. What are the operational issues in business management? How does microeconomics contribute to decision-making in the operational issues?
14. What is macroeconomics? In what way is macroeconomics applicable to business decision-making.
15. What macroeconomic issues figure in business decision-making? How does macro-economics help in understanding the implications of macroeconomic issues?
16. What is meant by business environment? What branch of economics is related to the environmental issues of private business?
17. What are the basic functions of business managers? How does economics help business managers in performing their functions?
18. What are the major macroeconomic issues related directly to business decision-making? What is their significance in business decisions?

CHAPTER

2 Objectives of Business Firms

CHAPTER OBJECTIVES

Profit maximization is generally taken as the objective of business firms; however, this is not the case in reality. By going through this chapter, you will know:

• The accounting and economic concepts of profit
• Traditional theories of profit dealing – the economists' concepts and sources of profit
• Problems in profit measurement from the viewpoint of managers, accountants, tax authorities and government policy makers
• Alternative objectives of business firms in modern times

2.1 INTRODUCTION

The first and foremost responsibility of business managers is to achieve the objective of the firm they manages. Therefore, all their managerial decisions, policies, strategy and functions are directed towards the attainment of the objective of the firm. A question that arises here is: What is the objective of business firms? The quick answer to this question is ‘making maximum profit, of course!’ However, economists do not agree with this answer. According to Baumol, a Nobel laureate in Economics and an authority on business economics, “It is most frequently assumed in economic analysis that the firm is trying to maximize its total profit. However, there is no reason to believe that all business firms pursue the same objective”’1. “In fact, it is common experience when interviewing executives to find that they will agree to every plausible goal about which they are asked. The say they want to maximize sales and also to maximize profits; that they wish ... to minimize cost; and so on”2.

In fact, researches conducted by the economists over time reveal that business firms pursue different objectives under different kinds of market conditions. The conventional theory of firm is based on the assumptions that the objective of business firms is to maximize profits. However, economists have found that business firms pursue many other objectives, viz., (*i*) maximization of sales revenue (Baumol), (*ii*) maximization of firm’s growth rate (Marris), (*iii*) maximization of managerial utility function (Williamson), (*iv*) satisficing behaviour (Cyert and March), (*v*) long-run survival and market share goal, and (*vi*) entry-prevention and risk-avoidance. The economists have pointed out some other

1. Baumol, William J., *Economic Theory and Operations Analysis*, 4lh Ed. (Prentice-Hall of lndia, New Delhi, 1985), p. 378.

2. Baumol, *op. cit.,* p. 377.

objectives such as achieving a target growth rate, making a target profit, and making a satisfactory or reasonable profit.

Since it is the responsibility of business managers to achieve the objective of the firm, they need to have a clear perception and understanding of the objective they have to achieve. Therefore, we begin the study of managerial economics with a brief discussion on the nature and purpose of different objectives of business firms and why and what objective of the firm is determined. The various objectives of business firms are generally classified under two categories:

(*i*) Profit maximization, and

(*ii*) Alternative objectives.

The objectives of business firms and their nature and purpose are discussed here under these two categories. We begin our discussion with a description of meaning of profit, theories and sources of profit, and problems in measuring profit. This will be followed by a detailed discussion on profit maximization as the business objective. Finally, we will discuss briefly the *alternative objectives* of business firms.

2.2 PROFIT AS BUSINESS OBJECTIVE

2.2.1 Meaning of 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…”3 In a general sense, ‘profit’ is regarded as income accruing to entrepreneurs, in the same sense as wages accrue to the labour; rent accrues to the owners of rentable assets; and interest accrues to the moneylenders. To a layman, profit means all income that flows to the investors. To an accountant, ‘profit’ means the excess of revenue over all paid-out costs including both manufacturing and overhead expenses, also known as ‘accounting profit’. It is more or less the same as ‘net profit’. For all practical purposes, profit means accounting profit plus non-allowable expenses.4 Profit figures published by the business firms are profits conforming to accounting concept of profit. Economist’s concept of profit is of ‘pure profit’, also 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. The accounting and economic concepts of profit are discussed below in detail.

2.2.2 Accounting Profit *vs.* Economic Profit

The two important concepts of profit that figure in business decisions are ‘economic profit’ and ‘accounting profit’. From conceptual clarity point of view, it is useful to understand the difference between the two concepts of profit. As already mentioned, in accounting sense, profit is surplus of total revenue over and above all paid-out costs, including both manufacturing and overhead expenses. Accounting profit may be calculated as follows.

Accounting profit = $TR - (W + R + I + M)$

3. Dean, Joel, *Managerial Economics*, (Asia Publishing House, Bombay), 1960, p. 3.

4. For example, Indian Income Tax Act makes only partial allowance for expenses on 'entertainment and advertisement'.

where TR = total revenue, 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 costs recorded in the books of accounts, are considered.

The concept of 'economic profit' differs from 'accounting profit'. Economic profit also takes into account the implicit or imputed costs. The *implicit cost* is *opportunity cost*. *Opportunity cost is defined as the income foregone due to withdrawal of 'the factors of production from their most remunerative alternative employment'.* Alternatively, opportunity cost can be defined as the income foregone which a businessman could make from the second best use of his resources. For example, if an entrepreneur uses his capital in his own business, he foregoes dividend which he could earn by purchasing shares of other companies or the interest which he could earn by depositing his money with banks for a period. Furthermore, if an entrepreneur uses his time and labour in his own business, he foregoes his income (salary) which he could earn by working as a manager in another firm or working as a consultant.

It should also be noted that *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,

Pure profit = Total revenue – (explicit costs + implicit costs)

Alternatively, *pure profit* may be defined as follows.

Pure profit = Accounting profit – (opportunity cost + unauthorized payments, e.g., bribes)

Pure profit so defined may not be necessarily positive for a single firm in a single year—it may be 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 market conditions.

2.3 TRADITIONAL THEORIES OF PROFIT: THE ECONOMISTS' PERCEPTION AND SOURCES OF PROFIT

What are the sources of profit? Economists are not unanimous on this issue. It is in fact this question that has been a source of an unsettled controversy and has led to the emergence of various theories of profit. In this section, we discuss briefly the main theories of profit.

2.3.1 Walker's Theory of Profit: Profit as Rent of Ability

One of the most widely known theories of profit was propounded by F.A. Walker. According to him, profit is the rent of "exceptional abilities that an entrepreneur may possess" over others. Just as land *rent* is the difference between the yields of the least and the most fertile lands, profit is the difference between the earnings of the least and the most efficient entrepreneurs. In formulating his profit theory, Walker assumed a state of perfect competition in which all firms are presumed to possess equal managerial ability. Each firm would receive only the wages of management. Thus, according to Walker, under perfectly competitive conditions, there would be no pure profit and all firms would earn only managerial wages, popularly known as ' *normal profit*'.

2.3.2 Clark's Dynamic Theory of Profit

According to J.B. Clark,5 profits arise in a dynamic economy, not in a static one. A static economy is one in which things do not change significantly; population and capital are stationary; production process remains unchanged over time; goods continue to remain homogeneous; factors enjoy freedom of mobility but do not move because their marginal product in every industry is the same; there is no uncertainty and hence no risk; and if there is any risk, it is insurable. In a static economy therefore, all firms make only the 'normal profit', i.e., the wages of management.

On the other hand, a dynamic economy is characterized by the following generic changes: (*i*) increase in population, (*ii*) increase in capital, (*iii*) improvement in production technique, (*iv*) changes in the forms of business organization, and (*v*) increase in and multiplication of consumer wants. The major functions of entrepreneurs or managers in a dynamic world are to take advantage of the generic changes and promote their business, expand their sales and reduce their costs. The entrepreneurs who successfully take advantage of changing conditions in a dynamic economy make *pure profit*, i.e., profit in addition to 'normal profit'. Pure profits, however, exist only in the short run. In the long run, competition forces other firms to imitate the changes made by the leading firms. This leads to a rise in demand for factors of production and therefore rise in factor prices and rise in cost of production. On the other hand, rise in output causes a decline in product prices, given the demand. The ultimate result is that pure profit disappears. In Clark's own words, "Profit is an elusive sum which entrepreneurs grasp but cannot hold. It slips through their fingers and bestows itself on all members of the society".

This, however, should not mean that profits arise in a dynamic economy only for a short period and disappear for ever. In fact, in a dynamic economy, generic changes take place continuously and managers with foresight continue to take advantage of the changing market conditions and make profit in excess of normal profit. According to Clark, emergence, disappearance and re-emergence of profit is a continuous process.

2.3.3 Hawley's Risk Theory of Profit

The risk theory of profit was propounded by F.B. Hawley in 1893. Risk in business may arise for such reasons as obsolescence of a product, sudden fall in prices, non-availability of certain crucial materials, introduction of a better substitute by a competitor, and risks due to fire, war, etc. Hawley regarded risk-taking as an inevitable accompaniment of dynamic production and those who take risks have a sound claim to an additional reward, known as "profit". According to Hawley, profit is simply the price paid by society for assuming business risks. In his opinion, businessmen would not assume risk without expecting adequate compensation in excess of **actuarial value**, i.e., the premium on calculable risk. They would always look for a return in excess of the wages of management for bearing risk. Assuming risk gives the entrepreneur a claim to a reward in excess of actuarial value of risk. Profit, according to Hawley, consists of two parts: (*i*) compensation for actuarial or average loss incidental to the various classes of risks necessarily assumed by the entrepreneur; and (*ii*) an inducement to suffer the consequences of being exposed to risk in their entrepreneurial adventures.

5. Clark, J.B. "Distribution as Determined by the Law of Rent". *Q.J.E*. April 1891.

2.3.4 Knight's Theory of Profit

Frank H. Knight[6] treated profit as a residual return to uncertainty bearing, not to risk bearing. Obviously, Knight made a distinction between risk and uncertainty. He divided risk into calculable and non-calculable risks. Calculable risks are those whose probability of occurrence can be statistically estimated on the basis of available data. For example, risk due to fire, theft, accidents, etc. are calculable and such risks are insurable. There remains, however, an area of risk in which probability of risk occurrences cannot be calculated. For instance, cost of eliminating competitors may not be accurately calculable and the strategies of the competitors may not be precisely assessable. The risk element of such incalculable events are not insurable. The area of incalculable risk is the area of uncertainty.

It is in the area of uncertainty that decision-making becomes a crucial function of an entrepreneur and a manager. If his decisions are proved to be correct by the subsequent events, the entrepreneur makes profit and *vice versa*. Thus, according to Knight, profit arises from the decisions taken and implemented under the conditions of uncertainty.

2.3.5 Schumpeter's Innovation Theory of Profit

The innovation theory of profit was developed by Joseph A. Schumpeter[7] His theory of profit is, in fact, the constituent of his *theory of economic development*. According to him, economic development takes place only when there are innovations in goods and service, manufacturing techniques, and methods of supply. Innovations are made and introduced by the business firms to make *pure profit*— profit in excess of normal profit. Introduction of innovations creates conditions for additional investment and labour employment and this leads to economic development and generate new business opportunities and also for profit. Schumpeter developed his theory of economic development by assuming a stationary economic condition in the country. Under the condition of stationary equilibrium, demand is equal to supply, prices are equal to cost, and total revenue of firms is exactly equal to total cost. Under these conditions, firms make only normal profit, i.e., *managerial wages*— there is no *pure profit.* Pure profit, i.e., profit in excess of *management wages,* can be made only by making innovations in goods and service, manufacturing techniques and in the methods of supply goods. According to him, innovations may include:

(*i*) introduction of new products and/or better quality goods and services,

(*ii*) introduction of a new production technology,

(*iii*) creating or finding new sources of raw materials,

(*iv*) opening new markets for the innovated products, and

(*v*) introduction of a more efficient and innovative management.

These kinds of innovations provide opportunities to innovative firms to fix a price of their product higher than the static equilibrium price. In simple words, innovative firms charge a price higher than production cost and hence make a *net profit.* Thus, according to Schumpeter, innovation is the source of profit.

6. Knight, Frank H., *Uncertainty and Profit* (Houghton Mifflin Co., Boston, New York, 1957).

7. Joseph A. Schumpeter, *Theory of Economic Development,* (Harvard University Press, Cambridge, Mass., 1934).

2.3.6 Monopoly Power as a Source of Profit

Monopoly is said to be another source of pure profit. Monopoly characterizes a market situation in which there is a single seller of a commodity without a close substitute. Monopoly may arise due to such factors as: (*i*) economies of scale, (*ii*) sole ownership of certain crucial raw materials, (*iii*) legal sanction and protection, and (*iv*) mergers and takeovers.

A monopolist may earn ‘pure profit’ or what is generally called in this case, ‘monopoly profit’, and maintain it in the long run by using its monopoly powers. Monopoly powers include: (*i*) powers to control supply and price; (*ii*) powers to prevent the entry of competitors by price cutting, and (*iii*) in some cases, monopoly power to exercise control over certain input markets. These powers help a monopoly firm to make pure profit (or monopoly profit). In such cases, monopoly is the source of pure profit.

It may be added at the end that pure monopoly too is a rare phenomenon. Monopolies, wherever they exist, are in the government sector (e.g., production and supply of electricity, water, transport services, etc.) or come into existence by governmental sanction and are under government control and regulation.

2.4 PROBLEMS IN PROFIT MEASUREMENT

In business analysis, generally two kinds of profit concepts are used: (*i*) accounting profit and (*ii*) economic or pure profit. Accounting profit equals revenue *minus* all explicit costs, and economic profit equals revenue *minus* both explicit and implicit costs. Once profit is defined, it should not be difficult to measure the profit of a firm for a given period. But two questions complicate the task of measuring profit: (*i*) which of the two concepts of profit should be chosen for measuring profit and with what purpose? and (*ii*) what costs should be or should not be included in the implicit and explicit costs?

The answer to the first question is that the profit concept should be chosen on the basis of the purpose of using the profit figure. Accounting concept of profit should be chosen when the purpose is to produce a profit figure for (*i*) the shareholders to inform them of progress of the firm, (*ii*) financiers and creditors who would be interested in the firm’s creditworthiness, (*iii*) the managers to assess their own performance, and (*iv*) for computation of tax-liability. For measuring accounting profit for these purposes, necessary revenue and cost data are, in general, obtained from the firm’s books of account.

On the other hand, if the objective is to measure ‘true profit’, the concept of economic profit should be used. However, ‘true profitability of any investment or business cannot be determined until the ownership of that investment or business has been fully terminated.’ But then life of a corporation is eternal. Therefore, true profit cannot be measured reasonably. This concept of business income is, as noted above, an ‘unattainable concept’ and hence is of little practical use.

If follows from the above discussion that, for all practical purposes, profits have to be measured on the basis of ***accounting concept***. But, measuring even the accounting profit is not an easy task. The main problem is to decide as to what costs should be and should not be included in the total cost. In fact, profit figures published by the joint stock companies are taken to be their actual accounting profit and are used for analyzing firms’ performance. There are, however, three specific items of cost and revenue which pose conceptual problems. These items are: (*i*) depreciation, (*ii*) capital gains and losses, and (*iii*) current vs. historical costs. Measurement problems arise for two reasons:

(*a*) economists' view on these items differs from that of accountants, and (*b*) there is more than one accepted method of treating these items. We discuss below the problems related to these items in detail.

2.4.1 Problem in Measuring Depreciation

Problems in measuring depreciation arise becuase of the different views and methods suggested by the economists and accountants. Economists view depreciation as capital consumption. From their point of view, there are two distinct ways of charging for depreciation: (*i*) the depreciation of an equipment must be equal to its *opportunity cost*, or alternatively, (*ii*) the depreciation must be equal to replacement cost that will produce comparable earning. *Opportunity cost* of an equipment is 'the most profitable alternative use of it that is foregone by putting it to its present use'. But, if equipment has no alternative use, the problem is then how to measure the opportunity cost. One method of estimating opportunity cost, suggested by Joel Dean, is to measure *the fall* in the value of the equipment during a year. This method, however, cannot be applied when a capital equipment has no alternative use, like a harvester, a printing machine and a hydro-power project, etc. In such cases, replacement cost is the appropriate measure of depreciation.

From accountants points of view, there are different methods of charging depreciation over the lifetime of an equipment. The use of different methods of charging depreciation results in different levels of profit reported by the accountants. The firm can apply any of the following four methods of charging depreciation: (1) straightline method (2) reducing balance method, (3) annuity method, and (4) sum-of-the-year's digit approach. These four methods yield four different measures of annual depreciation and, hence, the different levels of profit. So the problem arises 'what method to use'?

2.4.2 Treatment of Capital Gains and Losses

Capital gains and losses are regarded as 'windfalls'. Fluctuation in the stock prices is one of the most common sources of 'windfalls'. In a progressive society, according to Dean, capital losses are, on balance, greater than capital gains. Many of the capital losses are of insurable nature, and when a businessman over-insures, the excess insurance premium becomes eventually a capital gain. Treatment of capital gains and losses is another problem in profit measurement. Profit is affected by the way capital gains and losses are treated in accounting. As Dean suggests, "A sound accounting policy to follow concerning windfalls is never to record them until they are turned into cash by a purchase or sale of assets, since it is never clear until then exactly how large they are …"8 In practice, however, compoanies follow a diverse method. Most companies do not record capital gains until it is realized in money terms, but they do write off capital losses from the current profit. It means, they treat capital gains and losses differently. If 'sound accounting policy' is followed, there will be one profit, and if the other method is followed, there will be another figure of profit. That is the problem.

2.4.3 Current *vs.* Historical Costs

Accountants prepare income statements typically in terms of historical costs, i.e., the actual purchase price, not in terms of current price. The reasons given for this practice

8. Dean, Joel, *op. cit.*, p. 19.

are: (*i*) historical costs produce more accurate measurement of income, (*ii*) historical costs are less debatable and more objective than the present replacement value, and (*iii*) accountants' job is to record historical costs whether or not they have relevance for future decision-making. The accountant's approach ignores certain important changes in earnings and losses of the firms, e.g., (*a*) the value of assets presented in the books of accounts is understated in times of inflation and overstated at the time of deflation and (*b*) depreciation is understated during deflation. Historical cost recording does not reflect such changes in values of assets and profits. This problem assumes a critical importance in case of inventories and material stocks. The problem is how to evaluate the inventory and the goods in the pipeline.

There are three popular methods of inventory valuation: (*i*) first-in-first-out (FIFO), (*ii*) last-in-first-out (LIFO), and (*iii*) weighted average cost (WAC).

Under FIFO method, material is taken out of stock for further processing in the order in which they are acquired. The material stocks, therefore, appear in the firm's balance sheet at their actual cost price. This method is suitable when price has a secular trend. However, this system exaggerates profits at the time of rising prices.

The LIFO method assumes that stocks purchased most recently become the costs of the raw material in the current production. If inventory levels are stable, the cost of raw materials used at any point in the calculation of profits is always close to market or replacement value. But, when inventory levels fluctuate, this method loses its advantages.

The WAC method takes the weighted average of the costs of materials purchased at different prices and different points of time to evaluate the inventory.

All these methods have their own weaknesses and, therefore, they do not reflect the 'true profit' of business. So the problem remains that how to evaluate inventories so that it yields a true profit figure.

2.5 PROFIT MAXIMIZATION AS BUSINESS OBJECTIVE

The conventional economic theory assumes profit maximization as the only objective of business firms —profit measured as TR–TC. Profit maximization as the objective of business firms has a long history in economic literature. It forms the basis of conventional price theory. Profit maximization is regarded as the most reasonable and analytically the most 'productive' business objective. The strength of this assumption lies in the fact that this assumption 'has never been unambiguously disproved'.

Besides, profit maximization assumption has a greater predictive power. It helps in predicting the behaviour of business firms in the real world and also the behaviour of price and output under different market conditions. No other hypothesis explains and predicts the behaviour of firms better than the profit maximization assumption. Nevertheless, the profit maximization has been questioned strongly by some modern economists. This created a controversy on objectives of business firms. The controversy has been discussed below in detail.

2.6 CONTROVERSY ON PROFIT MAXIMIZATION: THEORY *VS*. PRACTICE

As discussed above, the conventional theory of firm assumes profit maximization as the sole objective of the business firms. Some modern economists, however, refute the profit

maximization assumption because, in their opinion, it is practically non-achievable. Their own findings reveal that business firms, especially big corporations, pursue several other objectives, rather than profit maximization. However, some modern economists have strongly defended the profit maximization objective. This has created a controversy on the profit maximization objective of the business firms. In this section, we discuss briefly the arguments against and for profit maximization objective.

2.6.1 Arguments against Profit-Maximization Objective

(*i*) The ***first argument*** against the profit maximization objective is based on the dichotomy between the *ownership* and *management* of business firms . It is argued that, in modern times, due to rapid growth of large business corporations, management of business firms has got separated from the ownership. The separation of management from ownership gives managers an opportunity and also the discretion to set firm's goals other than profit maximization. The researches conducted by the economists reveal that, in practice, business managers pursue such objectives as (*a*) *maximization of sales revenue,* (*b*) *maximization of the value of the firm, i.e., the net worth of the firm,* (*c*) *maximization of managerial utility function,* (*d*) *maximization of firm's growth rate*, (*e*) *making a target profit,* (*f*) *retaining and increasing market share,* and so on.

(*ii*) Another argument against profit maximization objective is that traditional theory of firm assumes managers to have full and perfect knowledge of market conditions and of the possible future development in business environment of firm. The firm is thus supposed to be fully aware of its demand and cost conditions in both short and long runs. Briefly speaking under profit maximization objective, a complete certainty about the market conditions is assumed. Some modern economists question the validity of this assumption. They argue that the firms do not possess the perfect knowledge of their costs, revenue and future business environment. They operate in the world of uncertainty. Most price and output decisions are based on *probabilities*. Besides, it is further argued that the equi-marginal principle of profit maximization, i.e., equalizing *MC* and *MR*, has been claimed to be ignored 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 Hitch9 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. According to them, 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 (where *AC* = Average cost, *AVC* = Average variable cost, *AFC* = Average fixed cost) and a normal margin of profit (usually 10 per cent).

In a similar study, Gordon10 has found (*i*) that there is a marked deviation in the real business conditions from the assumptions of the traditional theory, and (*ii*) that pricing

9. Hall, R.L., and C.J. Hitch, "Price Theory and Business Behaviour", *Oxford Economics Papers* (1939), reprinted in '*Studies in the Price Mechanism*' (ed.) by Wilson, T. and Andrews, P.W.S. (Oxford University Press, 1952).

10. Gordon, R.A., "Short Period Price Determination in Theory and Practice", *Am. Eco. Rev.,* 1948.

practices were notably different from the marginal theory of pricing. Gordon 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.

2.6.2 The Defence of Profit Maximization

The arguments against profit-maximization objectives have been strongly rejected by other economists. They argued strongly that pricing theory does have 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 profit maximization objective and marginal rule of pricing may be summed as follows.

In two empirical studies of 110 'excellently managed companies', J.S. Earley11 has concluded that the firms do apply the marginal rules in their pricing and output decisions. Fritz Maclup12 has argued in abstract theoretical terms that empirical studies by Hall and Hitch and by Lester do not provide conclusive evidence against the marginal rule and that 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 dealing with pricing practices of individual firms. The relevance of marginal rules in actual pricing system of firms could not be established for lack of communication between the businessmen and the researchers as they use different terminology. Researchers use technical terms like *MR*, *MC* and elasticities which are often abstract for businessmen. 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. This point has been discussed further in Chapter 14.

While the controversy on profit maximization objective remains unresolved, the conventional theorists, the marginalists, continue to defend the profit maximization objective.

Other Arguments in Defence of Profit Maximization Hypothesis

The conventional economic theorists defend the profit maximization hypothesis on the following grounds also.

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. What profit is reasonable may be

11. Early, J.S., 'Recent Developments in Cost Accounting and the Marginal Analysis', *Journal of Political Economy,* 1955, and 'Marginal Policies of Excellently Managed Companies', *Am. Eco. Rev.*, 1956.

12. Maclup, Fritz 'Marginal Analysis and Empirical Research', *Am. Eco. Rev.,* 1946, and 'Theories of the Firm: Marginalist, Managerialist, Behavioural', *Am. Eco. Rev.* (1967).

a matter of opinion. But, making profit is a necessary condition for the survival of the firm. Once the firms begin to make profit, they try 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 at least partly on the primary objective of making profit.

3. *Evidence against profit maximization objective is not conclusive.* Profit maximization is a time-honoured objective of business firms. Although this objective has been questioned by many researchers, some economists have argued that 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 objective has been found to provide a much more powerful basis for 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. In his opinion, ultimate test of its validity lies in its ability to predict the business behaviour and the business trends.

5. *Profit is a more reliable measure of a firm's efficiency.* Though not perfect, profit is the most quick 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 much 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 the total finance.*

6. *Finally,* according to Milton Friedman, whatever one may say about firms' motivations, if one judges their motivations by their managerial acts, profit maximization appears to be **a more valid business objective.** [13]

2.7 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 motivations. Berle and Means[14] were the first to point out the dichotomy between the ownership and the management of firm and its role in setting the goal(s) for the firm that they manage. Later on, Galbraith[15] wrote extensively on this issue. Their conjoint contribution to this issue 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

13. Friedman, Milton, *Essays in Positive Economics* (Chicago University Press, Chicago, 1953), pp. 3–43.

14. Berle, A.A., and G.C. Means, *The Modern Corporation and Private Property* (Commerce Clearing House, New York, 1932).

15. Galbraith, J.K., *Amercian Capitalism: The Concept of Countervailing Power,* 1952; *The Affluent Society,* 1958; *The New Industrial State*, 1967, all by Houghton Mifflin, Boston.

corporations, instead of maximizing profits, set goals that can keep the owners satisfied and managers can take care of their own interest in the corporation. The controversy arising out of dichotomy of management and ownership led to extensive research on the objective of business firms, which revealed several alternative objectives. In this section, we will discuss very briefly some important alternative objectives of business firms, especially of large business corporations. A detailed discussion follows in a subsequent chapter.

2.7.1 Baumol's Hypothesis of Sales Revenue Maximization

Baumol[16] has postulated maximization of sales revenue as an alternative to profit-maximization objective. He attributes this objective to the dichotomy between ownership and management in large business corporations. This dichotomy gives managers an opportunity to set their goals other than profit maximization goal which most owners and 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* for the following reasons.

First, managers' salary and other benefits are more closely related to sales revenue than to profits.

Secondly, banks and financial corporations look at and lay a great emphasis on sales revenue while financing a corporation.

Thirdly, trend in sales revenue is a readily available indicator of the performance of the firm. It helps also in handling the employee's problem of awarding efficiency and penalizing inefficiency.

Fourthly, increasing sales revenue enchances the prestige, reputation and perks of managers while profits go to the owners.

Fifthly, 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 whereas decreasing sales put the survival of the firm at risk.

Criticism: Baumol's sales maximization hypothesis has been questioned on the following grounds.

First, so far as **empirical validity** of sales revenue maximization objective is concerned, factual evidences are inconclusive.17 Most empirical works are, in fact, based on inadequate data simply because requisite data is mostly not available.

Secondly, even theoretically, if total cost function (*TC*) intersects the total revenue function (*TR*) before it reaches its climax, Baumol's theory collapses.

Finally, it is also argued that, in the long run, sales maximization and profit maximization objective converge into one. For, in the long run, sales maximization tends to yield only normal levels of profit which turns out to be the maximum under competitive conditions. Thus, profit maximization is not incompatible with sales maximization.

16. Baumol, W.J., *Business Behaviour, Value and Growth* (Macmillan, New York, 1959). Revised edition published by Harcourt, (Brace & World Inc., 1967).

17. Koutsoyiannis, A., *Modern Microeconomics* (Macmillan, London, 1979), pp. 346–51.

2.7.2 Marris's Hypothesis of Maximization of Firm's Growth Rate

According to Marris,18 managers maximize firm's *balanced growth rate* subject to managerial and financial constraints. He defines firm's balanced growth rate (*G*) as

$$G = G_D = G_C$$

where G_D = growth rate of demand for firm's 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*. Marris translates the two growth rates into two utility functions: (*i*) manager's utility function and (*ii*) owner's utility function. Manager's utility function (U_m) and owner's utility function (U_o) are expressed as follows.

$U_m = f$ (salary, power, job security, prestige, status),

and

$U_o = f$ (output, capital, market-share, profit, public esteem).

Owners' utility function (U_o) implies growth of demand for firm's product and supply of capital to the firm. Therefore, maximization of U_o 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, power, etc.) appearing in their own utility function and those appearing in the utility function of the owners (e.g., profit, capital market, share, etc.) are positively and strongly correlated with a single variable, i.e., *size of the firm*. Therefore, managers seek to maximize the size of the firm. Maximization of size of the firm depends on the maximization of its growth rate. *The managers, therefore, seek to maximize a steady growth rate.*

Marris's theory, though more rigorous and sophisticated than Baumol's sales revenue maximization, has its own **weaknesses**. It fails to deal satisfactorily with oligopolistic interdependence. Another serious shortcoming of his model is that it ignores price determination which is the main concern of profit maximization hypothesis. In the opinion of many economists, Marris's model too, does not seriously challenge the profit maximization hypothesis.

2.7.3 Williamson's Hypothesis of Maximization of Managerial Utility Function

As mentioned above, in modern corporations, owners (or stockholders) and managers (paid salary for their managerial services) are two separate entities with different objectives. Under these conditions, the problem of determining firm's objective is generally known also as ***Principal-Agent Problem.*** In *principal-agent problem*, 'principal refers to owner(s) of the business corporation' and 'agent' refers to manger(s) of the organisation. The *principal-agent problem* refers to the problem that arises because of different objecitves of the owners and the managers. In general, while owners' objective is to gain maximum profit, managers' objective is to maximise their own gains (salary and fringe benefits).

18. Marris, Robin, "A Model of the Managerial Enterprise", *Q.J.E.* , 1963 and *Theory of Managerial Capitalism* (N.Y., Macmillan, 1963).

Like Baumol and Marris, Williamson[19] argues that managers have discretion to pursue objectives other than profit maximization. Instead of maximizing profit, the managers of modern corporations 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 hypothesis, managers maximize their utility function subject to a satisfactory profit. A reasonable 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-quantitative variables such as prestige, power, status, job security, professional excellence, etc. 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 ready availability of funds for discretionary investment.

Like other alternative hypotheses regarding business objecitves, Williamson's theory too suffers from certain **weaknesses**. His model fails to deal with the problem of oligopolistic interdependence. Williamson's theory is said to hold only where rivalry between firms is not strong. In case of strong rivalry, profit maximization is claimed to be a more appropriate hypothesis. Thus, Williamson's managerial utility function too does not offer a more satisfactory hypothesis than profit maximization.

2.7.4 Cyert-March Hypothesis of Satisficing Behaviour

Cyert-March[20] 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 firms seeks to *maximize* sales or growth rate 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 own interest in the firms. Their interests are often conflicting. The manager's responsibility is to 'satisfy' them all. Thus, according to the Cyert-March hypothesis, firm's behaviour is 'satisficing behaviour'. The 'satisficing

19. Williamson, O.E., "Managerial Discretion and Business Behaviour", *Am. Eco. Rev*., 1963 and *The Economics of Discretionary Behaviour: Managerial Objectives the Theory of Firm* (Markhamm, Chicago, 1967).

20. Cyert, Richard M., and March James, G, (ed), *A Behavioural Theory of the Firm* (Prentice-Hall, Engleward Cliffs, N.J.: 1963). This hypothesis was earlier developed by H.A. Simon, in his "A Behavioural Model of Rational Choice", *Q.J.E.* , 1955, pp. 99–118.

behaviour' implies satisfying various interest groups by sacrificing firm's interest to some extent. The underlying assumption of 'Satisficing Behaviour' is that a firm is a coalition of different interest groups connected with various activities of the firm, e.g., shareholders, managers, workers, input suppliers, customers, bankers, tax authorities, and so on. All these groups have some kind of expectations–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 own interest.

In order to reconcile between the conflicting interests and goals, managers form an *aspiration level of the firm* combining the following goals: (*a*) Production goal, (*b*) Sales and market share goals, (*c*) Inventory goal, and (*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.

Criticism. The behavioural theory has, however, been criticised on the following grounds ***First,*** though the behavioural theory deals realistically with the firm's activity, it does not explain the firm's behaviour under dynamic conditions in the long run. ***Secondly***, it cannot be used to predict exactly the future course of firm's activities, ***Thirdly***, this theory does not deal with the equilibrium of the firm or the industry. ***Fourthly***, like other alternative hypotheses, this theory too fails to deal with interdependence of the firms and its impact on firms' behaviour.

2.7.5 Rothschild's Hypothesis of Long-Run Survival and Market Share Goals

Another alternative objective of a firm–as an alternative to profit maximization—was suggested by Rothschild.21 According to him, the primary goal of the firm 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.

2.7.6 Entry-prevention and Risk-avoidance

Yet another alternative objective of the firms suggested by some economists is to prevent entry of new firms into the industry. The motive behind entry-prevention may be (*a*) profit maximization in the long run, (*b*) securing a constant market share, and (*c*) avoidance of risk caused by 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 ownership, the possibility of profit maximization is reduced.

The advocates of profit maximization objective 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, no doubt, 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.

21. Rothschild, K.W., "Price Theory and Oligopoly", *E.J.* , 1947, pp., 297–320.

2.8 MAXIMIZATION OF VALUE OF THE FIRM: THE MODERN THEORY

The modern theory of firm makes a distinction between the 'principal' (owner) and the 'agent' (manager) of the firm and also between the objectives of owners and managers.

According to the modern theory of firm, owner's objective is the maximization of the ***value of the firm*** whereas objective of business managers is to maximize ***profit*** of the firm. The profit maximization objective is implicitly assumed to be current or short-run profit, whereas maximization of firm's value is a long-term objective. It implies that owners pursue the long-term objective and managers pursue short-term objective. However, as we will see below, both objectives turn out to be the same. Let us first look at the value maximization approach of the modern theory of firm.

According to the modern theory of firm, owners of the firms have been found to pursue the objective of maximizing the value or the worth of the firm. To this end, they prefer to sacrifice the short-run objective of profit maximization. Therefore, the recent development in the **theory of the firm** postulates that the objective of business firms is to maximize the **value or the wealth of the firms** in the long run, i.e., the value at which can be sold out.

The future value of the firm is worked out by estimating the ***present value*** of the expected future profits of the firm by discounting the profit by the degree of risk. The formula for estimating the ***present vlue (PV)*** is given below.

$$PV = \frac{\pi_1}{(1+r)^1} + \frac{\pi_2}{(1+r)^2} + \frac{\pi_3}{(1+r)^3} + \ldots + \frac{\pi_n}{(1+r)^n} = \sum_{i=1}^{n} \frac{\pi_i}{(1+r)^n}$$

(where *PV* = sum of the present value of expected future profits; $\pi_1, \pi_2, \pi_3 \ldots, \pi_n$ represent the expected future profits from 1 to *n* years; *r* = rate of discount22; and subscript *t* in Eq. 2.1b indicates the total period of profit earning.

Note that the present value of the firm (*PV*), estimated by the formula given above depends on the anticipated rate of profit (π). Since, the rate of profit is estimated as *TR –TC*, the maximization of the ***value of the firm*** would depend on the rate of profit (π). It implies that maximization of profit would maximize the *present value of the firm*

– the higher the value of π , the higher the value of the firm. It may thus be concluded

s

that the objective of maximization of the value of the firm is not different from the profit maximization objective of the firm.

2.9 MAKING A REASONABLE PROFIT: A PRACTICAL APPROACH

As noted above, objectives of business firms can be various. There is no unanimity among the economists and researchers on the objectives of business firms. One thing is, however,

22. Present value of future profit is lower in money terms because of compounding rate of interest. Therefore, to work out the present value of future profit a discount rate is used. For example, suppose a firm expects to make a profit of `100 crore after one year and the market rate of interest or return is 10% (used as discount rate). The present value of `100 crore expected after one year can be estimated by using the formula: $p/(1 + r)$. Thus, the present value of `100 crore can be worked as `100/(1 + 0.10) = 91 crore. It implies that by investing `91 crore today, the firm can earn `100 crore after one year.

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 the value of the firm, maximization of managers' utility function, long-run survival, market share, or entry-prevention—it has to make a profit. The firms, therefore, adopt a more practical approach. Maximization of profit in technical sense of the term may not be practicable, but making a profit has to be the objective of the firms. The firms may differ on 'how much profit' but they do set a profit target for themselves. Some firms set their objective of a 'standard profit', some a 'target profit' and some a 'reasonable profit'. A 'reasonable profit' is the most common objective.

Let us now look into the policy questions related to setting standard or criteria for a reasonable profit. The important policy questions are:

(*i*) Why do modern corporations aim at a "reasonable profit" rather than maximizing profit?

(*ii*) What are the criteria for a reasonable profit?

(*iii*) How is the "reasonable profit" determined?

Let us now briefly examine the policy implications of these questions.

2.9.1 Reasons for Aiming at "Reasonable Profits"

For a variety of reasons, modern corporations aim at making *a reasonable profit* rather than maximizing the profit. Joel Dean23 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.24 The firms, therefore, adopt a pricing and a profit policy that assure them a reasonable profit and, at the same time, keep potential competitors away.

2. ***Projecting a favourable public image.*** It becomes often necessary for large corporations to project and maintain a good public image. The reason is, if public opinion turns against the firm, its sales begin to fall. Also, if profits are high, government officials start raising their eyebrows on profit figures. So most firms set prices lower than those conforming to the maximum profit but high enough to ensure a "reasonable profit".

3. ***Restraining trade union demands.*** High profits make trade unions feel that they have a share in the high profit and therefore they raise demands for wage-hike. Hiking wage under pressure may lead to wage-price spiral and frustrate the firm's objective of maximizing profit. Therefore, profit restrain is sometimes used as a measure to prevent 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 (*d*) forestalling the anti-trust suits.

23. Dean, Joel, *op. cit.*, pp. 29–33.

24. A weak monopoly is one that has no strong barriers to protect its strategic material markets, patent rights, etc., and where production of a close substitute is technically a near possibility.

2.9.2 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?

(i) Forms of Profit Standard Profit standards may be determined in terms of (*a*) aggregate money terms—total net profit, (*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 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'.

(ii) Setting the Profit Standard The following are the important criteria that are taken into account while setting the standards for a 'reasonable profit'.

(a) Capital-attracting standard. An important criterion used in setting standard profit is that it must be high enough to attract external (debt and equity) capital. For example, if a firm's stocks are being sold in the market at five times their current earnings, it is necessary that the firm earns a profit at the rate of 20 per cent of the booked investment. There are however certain problems associated with this criterion: (*i*) capital structure of the firm (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 as it varies widely from company to company 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 especially by those firms for whom maintaining liquidity and avoiding debt are main considerations in profit policy.

(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 of 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 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 conditions and all the people concerned.

2.10 PROFIT AS A MEASURE OF MANAGERIAL CONTROL

An important managerial aspect of profit is that it is used in measuring and controlling performance of the executives of the large business undertakings. Researches have revealed that business executives of middle and high ranks often deviate from profit objective and try to maximize their own utility functions.[25] They think in terms of job security, personal ambitions for promotions, larger perks, etc., which often conflict with firms' profit-making objective. Keith Powlson[26] has pointed out three common deviationist tendencies:

(*i*) more energy is spent in expanding sales volume and product lines than in raising profitability;

(*ii*) subordinates spend too much time and money doing jobs to perfection regardless of its cost and usefulness; and

(*iii*) executives cater more to the needs of job security in the absence of any reward for imaginative ventures.

In order to control these deviationist tendencies and orienting managerial functions towards the profit objective, the top management uses 'managerial decentralization and control-by-profit techniques'. These techniques have distinct advantage for a big business corporation. Managerial decentralization is achieved by changing over from functional division of business activities (e.g., production branch, sales division, purchase department, etc.) to a system of product-wise division. Managerial responsibilities are then fixed in terms of profit. Managers enjoy autonomy in their operations under the general policy framework. They are allotted a certain amount of money to spend and a profit target to be achieved by the particular division. Profit is then the measure of executive performance, not the sales or quality. This kind of reorganization of management helps in assessing profit-performance of various product lines in a multi-product organization. It serves as a useful guide in reorganization of the product lines.

The use of this technique, however, raises many interesting technical issues that complicate its application. These issues centre around the method of measuring divisional profits and profit standards to be set. The two important problems that arise are: (*i*) should profit goals be set in terms of total net profit for the divisions or should they be confined to their share in the total net profit? and (*ii*) how should divisional profits be determined when there is a long ladder of vertical integration?

In respect to question (*i*), the most appropriate profit standard of divisional performance is revenue minus current expenses. In allocating different costs, however, some arbitrariness is bound to be there. However, where a long vertical integration is involved, relative profitability of a division can be fixed in terms of a lower 'transfer price' compared to the market price. But, control measures are not all that simple to apply. It is difficult to set a general formula. It has to be settled differently under varying conditions.

CONCLUSION

Although profit maximization continues to remain the standard business objective in economic analysis, there is no reason to believe that profit maximization is the only

25. Managerial utility functions have already been discussed above under 'Alternative Objectives of Business Firms'.

26. Quoted in Joel Dean, *Managerial Economics,* pp. 39–40.

objective that firms pursue. Modern corporations, in fact, pursue multiple objectives. Through their study of business firms, the economists have postulated a number of alternative objectives pursued by them. The main factor behind the multiplicity of the objectives, particularly in case of large corporations, is the dichotomy between the management and the ownership.

Moreover, profit maximization hypothesis is a time-tested one. It is more easy to handle. The empirical evidence against this hypothesis is not conclusive and unambiguous. Nor are the alternative hypotheses strong enough to replace this hypothesis. More importantly, profit maximization hypothesis has a greater explanatory and predictive power than any of the alternative hypotheses. Therefore, it still forms the basis of firms' behaviour. In the subsequent chapters, we will use this hypothesis to explain the price and output decisions of the business firms.

SUMMARY

- The objective of this chapter is to have a glance at different kinds of business objectives and how they guide the behaviour of business firms—an important aspect of the subject matter of managerial economics.
- Making a reasonable profit is the main objective of business firms. But the term 'profit' has different meaning under different perspectives, e.g., normal profit, abnormal profit, economic profit, etc.
- For all practical purposes, however, the term 'profit' means the difference between total revenue (*TR*) and total paid-out costs (*TC*), i.e., profit = *TR* – *TC*, known as accounting profit.
- The economists of different generations have conceptualised different sources of profit and different concepts of profit: (*i*) profit as wage for managerial function; (*ii*) profit as the result of taking advantage of dynamic economic conditions; (*iii*) profit as reward for risk bearing; (*iv*) profit as return on innovations, and (*v*) profit as outcome of monopoly.
- As regards the objective of business firms, in general, profit maximization is assumed to be the basic objective of the business firms.
- In practice, however, economists have found that firms seek to achieve various alternative objectives under different kinds of market conditions.

• Alternative objectives of business firms suggested by the economists include: (*i*) sales revenue maximization objective (W. J. Baumol), (*ii*) maximization of firm's growth rate (Robin Marris), (*iii*) maximization of managerial utility function (O. E. Williamson), (*iv*) satisficing behaviour – satisfying all related directly or indirectly to firms (R. M Cyert and J. G. March), (*v*) long-term survival (K. W. Rothschild), and (*vi*) entry prevention of rival firms.

• Other economists have, however, argued that, in spite of various alternative objectives pointed out by the economist, making a reasonable profit, if not maximization of profit, is the main objective of business firms.

REVIEW QUESTIONS

1. Distinguish between the following concepts of profit:

(*a*) Accounting profit and economic profit;

(*b*) Normal profit and monopoly profit;

(*c*) Pure profit and opportunity cost.

2. Explain the following statements:

(*i*) Profit is the rent for exceptional ability of an entrepreneur (Walker).

(*ii*) Profits arise only in a dynamic world (J.B. Clark).

(*iii*) Profit is a reward for risk bearing (F.B. Hawley).

(*iv*) Profit is a return to uncertainty bearing (F.H. Knight).

(*v*) Profit is reward for innovations (J.A. Schumpeter).

3. What is the most plausible objective of business firms? Why is there the controversy on profit maximization hypothesis? How will you react to the controversy?

4. What problems do the depreciation and capital gains cause in measuring profit? What are the methods of resolving the problems?

5. Examine critically profit maximization as the objective of business firms. What are the alternative objectives of business firms?

6. What are the objectives of business firms other than profit maximization? Are these alternative objectives of business firms theoretically sustainable?

7. Why do firms in general aim at a reasonable profit rather than pursuing other goals? What are the standards of reasonable profits?

8. State whether the following statements are true or false:

(*a*) Pure profit is the return in excess of the opportunity cost.

(*b*) Profit maximization is the sole objective of all the firms.

(*c*) Profit is maximum when MR – MC is maximum.

(*d*) Pure profit is nil when opportunity cost equals actual earning.

9. Profit maximization remains the most important objective of business firms in spite of multiplicity of alternative business objectives suggested by the modern economists. Comment.

10. Profit maximization is theoretically the most sound but practically unattainable objective of business firms. Do you agree with this statement? Give reasons for your answer.

11. Maruti Udyog Ltd. (MUL) is losing its market to other car manufacturing companies. Suppose MUL management is faced with a dilemma. If it retains its present level of profit margin, it will continue to lose its market share. And, if cuts down the price to retain the demand for its car and, thereby, its market share, it loses the profit volume which leads to fall in its share prices, and market worth of MUL goes down. What in your opinion should be the objective of MUL?

12. State precisely the major propositions of the Williamson's model of managerial discretion in determining the firm's goal. How are his propositions different from profit maximization assumption?

13. Explain how profit is used as a control measure. What problems are associated with the use of profit figure as a control measure?

14. What are the considerations in aiming at a reasonable profit target? What standards are used in determining a reasonable profit?

[**Ans.** Q. 8. (*a*) True, (*b*) False, (*c*) False, (*d*) True.]

FURTHER READING

Baumol, W.J., *Economic Theory and Operations Analysis* (Prentice-Hall of India Pvt. Ltd., 1985).

Cyert, r.m. and J.G. marCh, *A Behavioural Theory of the Firm* (Prentice-Hall, Englewood, NJ, 1963).

Dean, Joel, *Managerial Economics* (Asia Publishing House, Bombay, 1960), Indian edition, Ch. 1.
haWley, F.B., 'The Risk Theory of Profit' *Qly. Jl. of Eco.* , July 1893.
KniGht, F.B., *Risk Uncertainty and Profit* (Houghton Miffin Co., Boston, New York, 1957).
Koustoyiannis, a., *Modern Microeconomics*, 2nd ed. (Macmillan, 1979), Chs. 15–18.
salvatore, D., *Managerial Economics* (McGraw-Hill, 1989), Ch. 1.
sChumpeter, J.a., *Theory of Economic Development* (Harvard University Press, Cambridge, Mass., 1938).
WalKer, F.a., 'The Source of Business Profits', *Qly. Jl. of Eco.*, April 1887.
Williamson, o.e., 'Managerial Discretion and Business Behaviour', *Am. Eco. Rev., 1963*.

CHAPTER

3 Some Fundamental Concepts and Business Decision Rules

CHAPTER OBJECTIVES

Decision making is the basic function of business managers. Certain fundamental concepts and decision rules are adopted in business decision making. The objective of this chapter is to acquaint you with the following concepts and decision rules:

- The concept of opportunity cost and decision rule
- Marginality principle and decision rule
- Incremental principle of decision rule
- The equi-marginal principle of business decision

3.1 INTRODUCTION

In Chapter 2, we have discussed alternative objectives of business firms. Whatever may be the objectives of a firm, all its managerial decisions are directed towards achieving the predetermined goal. Managerial decisions are taken at different levels of sophistication. While some business decisions require only 'rule-of-thumb' technique, some others involve the use of sophisticated techniques. The 'rule-of-thumb' technique, evolved out of the traditional business manag ement practices, is used where routine type of business decisions are involved. Sophisticated techniques of business decision-making are used where business decisions involve the problem of handling complex business issues.

There are certain fundamental economic concepts which are used—explicity or implicity, consciously or unconsciously, deliberately or otherwise—in business decisions of complex nature. This chapter presents a brief explanation of some major economic concepts and their use in business decision-making.

3.2 THE OPPORTUNITY COST AND DECISION RULE

The opportunity cost is essentially opportunity lost because of scarcity of resources. The concept of opportunity cost is related to the alternative uses of scarce resources. As noted earlier, resources, both natural and man-made, are *scarce* in relation to demand for them to satisfy the ever growing human needs. Resources, though scarce, have *alternative uses*. The *scarcity* and the *alternative uses* of the resources give rise to the concept of *opportunity cost*.

In the context of a business firm, resources available to a business unit—be it an individual firm, a joint stock corporation or a multinational—are limited. But the limited resources available to a firm can be put to alternative uses. For example, suppose a firm has `100 million at its disposal and the firm finds three risk-free alternative uses of the fund

available to it: (*i*) to expand the size of the firm, (*ii*) to set up a new production unit in another city, and (*iii*) to buy shares in another firm. Suppose also that the expected annual return from each of the three alternative uses of a finance of `100 million is given as follows.

Alternative 1 : Expansion of the size of the firm	`20 million
Alternative 2 : Setting up a new production unit	`18 million
Alternative 3 : Buying shares in other firms	`16 million

All other things being the same, a rational decision for the firm would be to invest the money in alternative 1. This implies that the manager would have to sacrifice the annual return from the second best alternative, i.e., a return of `18 million expected from alternative 2. In economic jargon, `18 million is treated as an annual *opportunity cost* of an annual income of `20 million. Thus, *the opportunity cost of availing the best opportunity is the foregone income expected from the second best opportunity of using the resources*.

In our example, the opportunity cost of `20 million per annum is `18 million per annum. The difference between *actual earning* and its *opportunity cost* is called ***economic gain*** *or* ***economic profit***. The concept of opportunity cost assumes a great significance where economic gain is neither insignificant nor very large because then it requires a careful evaluation of the two alternative options.

The applicability of the opportunity cost concept is not limited to decisions on the use of financial resources. The concept can be applied to all other kinds of issues involved in business decisions, especially where there are at least two alternative options involving costs and benefits. For example, suppose a firm has to take a decision on whether to fire an efficient labour officer (for treating labour unkindly) in settlement of a dispute with the labour union or to allow the matter to be taken to the labour court. If the firm decides to fire the labour officer, then the loss of an efficient and reliable labour officer is the *opportunity cost* of buying peace with the labour union. If the firm decides to retain the labour officer, come whatever may, then the cost of prolonged litigation, the cost arising out of a possible labour strike and the consequent reduction in output are the *opportunity costs* of retaining the labour officer. Given the two options, the firm will have to evaluate the cost and benefit of each option and take a decision accordingly.

3.3 MARGINALITY PRINCIPLE AND DECISION RULE

The concept of 'marginal' value is widely used in economic analysis, for example *marginal utility in consumer analysis, marginal productivity in production analysis* and *marginal revenue and marginal cost in pricing theory*. Marginality concept assumes special significance where maximization or minimization problem is involved, for example, maximization of a consumer's utility, maximization of a firm's profit, minimization of cost, etc. The term 'marginal' refers to the change (increase or decrease) in total quantity or value due to a one-unit change in its determinant. For example, given the factor prices, the total cost of production of a commodity depends on the number of units produced.

In this case, 'marginal cost' (*MC*) can be defined as the rise in total cost as a result of producing one additional unit of a commodity. The marginal cost (*MC*) can be worked out as follows.

Marginal cost (MC) = $TC_n - TC_{n-1}$

where TC_n = total cost of producing n units and TC_{n-1} = total cost of producing $n - 1$ units.

For example, suppose total cost (*TC*) of producing 100 units of a commodity is `2500. When production is increased to 101 units *TC* increases tò2550. In this case, $TC_n$ = `2550, TC_{n-1} = `2500 (where $n = 101$ and $n - 1 = 100$).

Then

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

= `2550 – `2500 = `50

Similarly, *MR* can be defined as the change in *TR* due to the sale of one additional unit of a product. It can also be defined as:

Marginal revenue (MR) = $TR_n - TR_{n-1}$

where TR_n = total revenue from the sale of n units and TR_{n-1} = total revenue from the sale of $n - 1$ units.

Alternatively, if *TC* and *TR* are given in the form of functions, then *MC* and *MR* are defined as the *first derivatives* of the *TC* and *TR* functions, respectively. Suppose *TC* and TR functions are given as

$TC = f(Q)$

and

$TR = f(Q)$

Then the first derivative is calculated through calculus as follows.

$$MC = \frac{\partial TC}{\partial Q} \text{ and } MR = \frac{\partial TR}{\partial Q}$$

If *TC* and *TR* are given in the form of curves, then *MC* and *MR* are defined in terms of the *slopes* of the *TR* and *TC* curves. Marginal value derived from a function in the form of its first derivative is not the same as defined in terms of a unit. More on this aspect of the 'marginal' concept will follow in a subsequent section.

The decision rule Suppose a profit-maximizing firm is faced with a problem—how much to produce so that profit is maximized. One simple decision rule under the marginal principle is that a business activity (production and sale) must be carried out so long as its $MR > MC$. As regards profit maximizing output, economists use the marginality principle to set a necessary condition for profit maximizing output. The necessary condition for profit maximizing output is that *MC* must be equal to *MR*. That is, profit is maximum where

$MR = MC$

In simple words, the profit of a firm is maximized at that level of output and sale where the cost of producing one additional unit equals the revenue from the sale of that unit of output. The use of the marginality concept in profit maximization and in other optimization problems will be discussed in detail in chapter 4.

The application of the marginal principle for profit maximization has certain serious **limitations** which must be borne in mind.

One, it can be applied only where the management has the *TC* and *TR* data for each and every unit of output or where the management is fully aware of the cost of producing one additional unit and the price expected to be received from the sale of that unit.

Two, the concept of 'marginal' value, when used in cost analysis, reduces the value of *MC* to the change in *variable cost* only. Therefore, marginal analysis can be applied to a situation in which only the variable cost changes.

3.4 INCREMENTAL PRINCIPLE AND DECISION RULE

The concept of 'incremental' value is similar to the concept of 'marginal' value but with a difference. While 'marginal' concept is basically a theoretical concept, 'incremental' is a practical concept. Marginal principle can be applied only when change in quantity or value, e.g., *MC* and *MR*, can be calculated precisely in unit terms. In general, however, firms find it difficult to estimate *MC* and *MR* as defined conceptually. The reason is that most business firms produce and sell their products in bulk, not in terms of units unless, of course, it is the case of production and sale of such large-unit goods as aeroplanes, ships, large buildings, turbines, etc. Where production and sale activities are carried out in bulk and where both *fixed* and *variable* costs are subject to change, business managers use the *incremental principle* in their business decisions.

The ***incremental principle*** is applied to business decisions which involve bulk production and a large increase in total cost and total revenue. Such an increase in total cost and total revenue is called 'incremental cost' and 'incremental revenue' respectively, related to 'incremental output'.

Let us first explain the concept of ***incremental cost***. Conceptually, *incremental costs* can be defined as the costs that arise due to a business decision. For example, suppose a firm decides to increase production by using more inputs or by adding a new plant to the existing capacity. This decision increases the firm's total cost of production from `100 million tò115 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*.

The ***incremental revenue***, on the other hand, is the increase in revenue due to a business decision, e.g., a decision to increase production and sale of the firm's product. 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 the 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 tò150 million. Thus, the post-decision total revenue of `150 million *less* the pre-decision total revenue of `130 million = `20 million is the *incremental revenue*.

Incremental Reasoning in Business Decision

The use of the incremental concept in business decisions is called *incremental reasoning*. The incremental reasoning is used for accepting or rejecting a business proposition or option. For instance, suppose that in our example, the firm is considering 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 `20 million. The incremental revenue exceeds the incremental cost by `5 million which means a 33.33 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.

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, especially

of cost and revenue are involved. Besides, incremental concept and reasoning are used in business decisions more frequently than the marginal concept. There are at least **two reasons** for this. ***First,*** marginal 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 decisions. However, marginal concept is of great significance in theoretical analysis.

3.5 CONTRIBUTION ANALYSIS

Associated with the concept of incremental cost and incremental revenue is the concept of *contribution*. The *contribution* of a business decision can be defined as the difference between the *incremental revenue* and the *incremental cost* associated with that particular decision. Contribution analysis is generally applied to analyze the contribution made by a business decision to *overhead costs* and *revenue* to work out the net result of that particular business decision. It is a useful technique for taking a decision on such issues as:

(i) whether or not to set up a project,

(ii) whether or not to introduce a new product,

(iii) whether or not to accept a fresh order,

(iv) whether or not to add an additional plant,

(v) whether to make or to buy inputs, and so on.

For the use of contribution analysis for a business decision, it is important to know what is the relevant *incremental cost* and what is the *relevant incremental revenue*. That is, it is important to know what is included and what is not in *incremental cost* and *incremental revenue*.

The relevant ***incremental costs*** that are taken into account in contribution analysis include the following costs:

(*i*) Present explicit costs including:

(*a*) Explicit variable costs consisting of Direct labour costs, Direct material costs, and Direct variable overheads.

(*b*) Fixed costs including New additional equipment and New additional personnel.

(*ii*) Opportunity cost: Foregone contribution expected from the second best alternative use of resources.

(*iii*) Future incremental costs: Expected present value of probable future costs arising out of a present business decision.

For the purpose of contribution analysis of a business decision, the following costs are considered to be the *irrelevant costs* and are not taken into account.

(*i*) *Committed costs*, i.e., the costs which have already been committed by the firm and are bound to be incurred irrespective of the decision under consideration, e.g., payment of old debts, interest on old debts, committed raise in salaries and wages of managers and workers, etc.

(*ii*) *Sunk costs*, i.e., the costs which have already been made on purchase of assets (building, plant and equipment) and non-recoverable advance payments.

The relevant ***incremental revenue*** includes (*i*) explicit present revenue, (*ii*) a possible opportunity revenue and (*iii*) a possible future revenue. The *explicit present period*

revenue is the revenue which arises directly from a decision in the present period. An *opportunity revenue* is the cost which could be avoided by taking a particular decision. For example, if a firm bids a low price for winning a contract, it avoids the cost arising out of laying off its labour. A *possible future revenue* is the revenue that might arise due to the present decision. For example, if a firm wins a contract because of its low bids, it builds a reputation and is treated as a favoured bidder for future projects. This yields a revenue which would not exist otherwise.

3.6 THE EQUI-MARGINAL PRINCIPLE

The equi-marginal principle was originally associated with 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 distribution of consumption expenditure maximizes a consumer’s total utility.

The law of equi-marginal principle was over time applied by business managers to 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 worth `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 3.1.

Table 3.1 *Marginal Productivity (MP) Schedule of Projects A, B and C*

(` *in million*)

Units of Expenditure	*Marginal Productivity (MP) (` 10 million)*		
	Project A	*Project B*	*Project C*
1st	501	403	354
2nd	452	305	306
3rd	357	208	209
4th	2010	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., *MPA* = *MPB* = *MPC*. It can be seen from Table 3.1, 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. It suggests that a profit (gain) maximizing firm 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 propotios 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.

3.7 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 have short-run outcome or pay-off and, therefore, involve short-run time perspective. For example, a decision to buy explosive materials for manufacturing crackers involves short-run demand prospects. Similarly, a decision regarding building inventories of finished product involves a short-run time perspective. There are, however, a large number of business decisions which have long-run repercussions, e.g., investment in 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, for example, the decision regarding the introduction of a new product may not be profitable in the short-run but may prove very profitable in the long-run. For example, the introduction of a newly designed laptop computer—a book size laptop priced at `10,000—may not succeed in the market quickly and smoothly. It may be difficult to even cover the variable costs because potential buyers have already one laptop or they may be uncertain about its usefulness, quality, serviceability and cost of operation. But in the long-run, it may enjoy 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 than the increase in cost. Therefore, while taking a business decision with long-run implications, it is immensely important to keep a well worked out time perspective in view.

The 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 analysis point of view and decisions based thereon may result in poor pay-offs. For example, in a business decision 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 time perspective is unwise.

SUMMARY

• Chapter 3 presents a brief discussion on the fundamental economic concepts and principles figuring often in the process of business decisions.

• The concepts and decision principles discussed here include (*i*) opportunity cost, (*ii*) marginality principle, (*iii*) incremental principle, (*v*) contribution analysis, and (*iv*) equi-marginal principle.

• *Opportunity cost* is opportunity lost due to scarcity of resources with alternative uses. In real terms, the opportunity cost is the income or gain expected from the second best use of resources lost due to the best use of the scarce resources.

• *Marginality principle* is a theoretical principle. Marginal principle can be stated as the principle of maximizing the gain from the use of resources. The condition of marginal principle is that marginal cost is equal to marginal benefit. For example, profit is maximized when cost of production of marginal unit is equal to the revenue from the sale of the marginal unit.

• *Incremental principle* is the principle or rule of taking decisions on increasing production, labour employment, investment, etc. in terms of a large quantity or amount, not on the basis of a single marginal unit.

• *Contribution analysis* is the analysis of estimating possible contribution from a business decision related to incurring incremental cost, increasing investment, increasing advertisement expenditure, etc.

• *Equi-marginal principle* is the rule of equalizing marginal gain from all possible actions. For example, consumers equalize the utility from the consumption of different goods and firms equalize marginal profits from different business actions.

REVIEW QUESTIONS

1. Define the concept of opportunity cost. What are its implications in business decisions?
2. What is the basis of opportunity cost? Under what conditions is the opportunity cost equal to zero?
3. What is marginal principle? What is the application of this principle in business decision-making? Illustrate the application of marginal principle in case of a firm maximizing its profit.
4. How are marginal cost and marginal revenue estimated. What is the marginal condition for profit maximization?
5. Define the concept of 'incremental' with reference to incremental revenue and incremental cost. How are incremental revenue and incremental cost different from the concept of marginal revenue and marginal cost?
6. What is the equi-marginal principle? Under what conditions is the application of this law relevant? Illustrate how this law helps in optimizing the allocation of resources between the competing ends.
7. What is meant by time perspective in business decision? Under what kind of business decisions time perspective become an important consideration?

FURTHER READING

Allen, R.G.D., *Mathematical Analysis for Economists*, (Macmillan, London, 1956).

BAumol, W.J., *Economic Theory and Operations Analysis*, (Englewood Cliffe, N.J.), 4th Edn.

ChiAnG, AlphA, *Fundamental Methods of Mathematical Economics*, (McGraw-Hill, N.Y.), 3rd Edn.

CHAPTER

4 Basic Tools of Economic Analysis and Optimization Techniques

CHAPTER OBJECTIVES

Some specific analytical tools are applied to analyse economic problems to find the optimum level of business activities. The objective of this chapter is to present a brief description of the following tools of economic analysis and optimization techniques:

• The meaning and kinds of economic variables often used in business decisions
• Method of determining the functional relationship between the interrelated variables
• The meaning and derivation of economic functions
• The methods of measuring relationship between the variables – the calculus method
• Techniques of finding the optimum value of target variable – production, total revenue, cost of production and profit

4.1 INTRODUCTION

We have so far discussed the nature and scope of managerial economics, objectives of business firms, and some basic economic concepts and their application in business decision making. As noted above, managerial economics is essentially the application of economic theories and ***tools of analysis*** in business decision making. In this chapter, we

discuss some basic tools of analysis applied to analyse business issues with the objective of finding a reasonable solution to the decision problem. The tools of economic analysis constitute a vast area of study, known as 'mathematical economics'. However, we confine our discussion to the basic concepts and tools of economic analysis generally applied to business decision making, as listed below.

(*i*) Functional relationship between economic variables,

(*ii*) Concept of slope and its application,

(*iii*) Optimization techniques,

(*iv*) Constrained optimization, and

(*v*) Regression analysis.

This chapter presents a discussion on the first four tools of analysis. Regression analysis, not discussed here, is a statistical tool widely used in measuring the relationship between any two or more interrelated economic variables. Regression method is a highly technical tool and requires a detailed treatment. Therefore, *regression technique* has been explained elaborately in the next Chapter. We begin our discussion on the analytical tools by explaining the meaning and kinds of ***economic variables*** generally used in economic analysis and also in most business decision issues.

4.2 ECONOMIC VARIABLES: MEANING AND KINDS

4.2.1 Meaning of Economic Variables

By definition, any thing that is subject to change is called ***variable.*** In economic sense, any quantity, value, price and rate of any thing which changes on its own or due to change in its determinants is called ***economic variable.*** For example, the terms like demand, supply, price, cost, sales, revenue, profit, capital, labour, money demand and supply, interest rate, advertisement spending, etc. are all economic variables. Most of these variables are interrelated and interdependent. The interrelatedness and interdependence of the variables means that change in one variable causes change in other related variable(s). For example consider the interrelationship between and interdependence of some the important economic variables.

- Demand depends on price and price depends on demand
- Supply depends on price and price depends on supply
- Production depends on cost of production and production on cost
- Market price depends on both demand and supply and *vice versa*
- Borrowing depends on the rate of interest and interest rate on borrowing
- Profit depends on price, cost, and sales.

4.2.2 Kinds of Economic Variables

Economic variables are generally classified under two broad categories depending on their role in economic analysis.

(*i*) Dependent and Independent variables and

(*ii*) Endogenous and Exogenous variables.

Dependent and Independent Variables Depending on the purpose of analysis, any of the economic variables can be treated as *dependent variable* or as *independent variable.* A *dependent variable* is one whose value depends on the value of other variables. An *independent variable* is one whose value changes on its own, or is assumed to change due to certain exogenous factors—the factors outside the model. For example, if demand for computers is assumed to depend on its price, then 'demand for computer' is the dependent variable and computer price is the independent variable. Consider another example. Petrol price in India has been increasing due to increase in import oil price. In this case, domestic oil price is a dependent variable and international oil price is an independent variable.

Endogenous and Exogenous Variables Economic variables are also classified as *endogenous variables* and *exogenous variables.* An endogenous variable is one whose value is determined within the model or framework of the analysis and an *exogenous variable* is one whose value is determined outside the model. In our petrol-price example, domestic oil price is endogenous variable and international oil price as exogenous variable.

Endogenous and exogenous categories of variables will be explained further and illustrated in the following section.

4.2.3 Methods of Showing Relationships Between Variables

A variety of functions are used in economic theory and many of them are applied to analyze business problems and in business decisions making, e.g., demand function, supply function, production function, cost function, total revenue function, profit function, and so on. These functions are used to explain various economic theories which constitute managerial economics. In this sections, we introduce the concept of functions and show the use of functions in economic analysis.

As stated above, most economic variables are interrelated and interdependent. In most cases, economic variables have ***cause-and-effect relationship.*** The relationship between any two or more related economic variables can be expressed in a ***tabular***, a ***graphical*** and a ***functional*** form. Tabular and graphical forms serve the purpose only when the number of variables and the number of observations (data) are small. In this section, *tabular* and *graphical methods* are discussed briefly. The *functional method* has been discussed elaborately in the next section.

Table 4.1 *Price of Pizza and Weekly Pizza Sale*

Pizza Price	*No. of Pizzas sold*
100	00
80	100
60	200
40	300
20	400
00	500

Tabular Method When the relationship between two variables is expressed by sequential data in a table, it is called a *tabular form*. For example, consider a simple case of relationship between price of pizza and the number of pizzas sold per week in your college canteen. Suppose weekly sale of pizza by the canteen is given in Table 4.1.

The data given in Table 4.1 shows that there exists a relationship between the price of pizza and its quantity demanded per week. It shows that as price of pizza decreases, its demand increases, or conversely, as pizza price increases, demand for it decreases. A little deeper observation shows that each fall in pizza price by `20 per piece results in increase in pizza demand by 100, or a one-rupee decrease in pizza price causes an increase in sale by 5 pizzas. This is tabular presentation of relationship between two economic variables—price and demand. One can imagine a number of other such relationships between consumer goods and their prices.

Graphical Method The data given in Table 4.1 can be presented graphically as shown in Fig. 4.1. By plotting the data, we get a line marked *PQ*. A curve or line so generated is called *demand curve*. The line *PQ* shows the nature of relationship between the price of pizza and its quantity demanded, graphically.

Both Table 4.1 and Fig. 4.1 show that
as the price of pizza decreases, its demand
increases and *vice versa, all other things*
Fig. 4.1 *Pizza Price and Demand for Pizza*
remaining the same. One can easily find

the relationship between pizza price and its quantity demanded per week in the canteen. This gives the law of demand. The relationship between the price of pizza and its weekly demand in the college canteen can be stated as follows.

(*i*) There is an *inverse* relationship between the pizza price and demand for it, and

(*ii*) For each fall in pizza price by `20, weekly demand increases by 100 or for each one-rupee fall in the price of pizza, its weekly demand increases by 5.

Limitations The tabular and graphical methods have serious limitations. The example used to illustrate these methods is a very simple case of showing relationship between two economic variables— price and demand. Now consider the case of a commodity with a long series of its prices running into pages and both price and demand changing at different rates. In this case, tracing the relationship between the price of a commodity and its demand would be a cumbersome process. The conclusion drawn may not be as precise and accurate as in our pizza example above.

In reality, most economic problems are complex and involve large number of variables with a long series of data or number of observations. In case of such economic problems, expressing relationship between two or more economic variables by tabular and graphical methods does not present a clear picture, nor is it easily comprehensible and precisely measurable. In such cases, therefore, economists use a mathematical tool, called "function", to express the relationship between the economic variables. Let us now discuss and illustrate the use of 'functions' in economic analysis.

4.3 FUNCTIONAL RELATIONSHIP BETWEEN INTERRELATED VARIABLES

A function is a mathematical technique of stating symbolically the relationship between two or more interrelated and interdependent variables. The variables may have cause-and-effect relationship. When there are only two variables in a function – a dependent and an independent variable – it is called ***bivariate function.*** And, when a function involves more that two variables – one dependent and several independent variables – it is called ***multivariate function.*** The formation and the forms of the bivariate and multivariate functions are discussed here briefly.

Bivariate Function**:** As noted above, a bivariate function involves two interrelated variables – one dependent and one independent variable. Recall that, in our pizza example, there are only two variables – pizza price and pizza demand. It may be concluded from the data given in Table 4.1 that demand for pizza depends on its price, implying the assumption that *all other variables affecting pizza demand remain constant*. This conclusion may be re-expressed in mathematical terms as "Demand for pizza is the function of pizza price". This statement is expressed in functional form as follows.

$$Dp = f(Pp) \quad (4.1)$$

where Dp= demand for pizza, and Pp = pizza price per unit.

In its ***general form,*** a bivariate demand function is generally expressed as

$$Q = f(P) \quad (4.2)$$

In Eq. (4.2), the variable Q denotes demand for a commodity – the dependent variable – and variable P the price of the commodity. Similarly, in case of any two interrelated variables1 their relationship can be expressed in the form of a bivariate function.

Multivariate Function: When a particular variable is found to be dependent on several independent variables, the function takes the form of a *multivariate function*. In pizza demand function, all other factor are assumes to remain constant. This may be a reality in the *short run*. In the long run, however, variables other than price affecting demand for a commodity (be it pizza or any other commodity) may change and may affect the demand for the product. In that case, demand function for the product is written in the form of a *multivariate function.* For instance, the long-run demand for a consumer good depends on:

- Commodity price (*P*)
- Advertisement (*A*)
- Consumers' income (*Y*)
- Taste and fashion (*T*)
- Price of the substitutes (*P* s)
- Price expectations (*PX*)
- Price of complements (*P* c)
- Unexpected factors (*U)*

Then the *long-run demand function* for a commodity, say *X,* is expressed in the form of a *multivariate demand functions* as given below.

$Dx = f\ (Px, Y, Ps, Pc, A, T, E, U)$

…(4.3)

In demand function (4.3), *Dx* is the *dependent variable* and all other variables are treated as *independent variables.*

Given the scope of this book, we will use mainly the two-variable functions, i.e., those involving only one *dependent variable* and one *independent variable,* like pizza demand function given in Eq. (4.1). Therefore, the bivariate function has been discussed here in detail.

4.3.1 Derivation of a Bivariate Function

A function in its general form, as given above, shows only that 'there is a relationship between the two variables'. For example, Eq. (4.1) indicates that 'demand for pizza depends on its price'. It does not reveal the following two important aspects of the relationship between the two variables.

(*i*) The *nature of relationship,* i.e., whether pizza price (*PP*) and pizza demand (*DP*) are related positively or negatively, or directly or inversely, or in simple words, whether a rise in pizza price leads to a rise or fall in pizza demand; and

(*ii*) The *quantitative measure of relationship* or the *degree of relationship,* i.e., what change in the independent variable (pizza price, *PP*) brings what change in the independent variable (pizza demand, *DP*), whatever the nature of relationship.

The nature and degree of the relationship between the variables *must* be determined *a priori* and known for the use of a function in economic analysis and its application in business decision-making. In other words, *the relationship between the dependent and independent variables must be specified and quantified for a meaningful use of economic functions in business decisions.* A function which reveals the nature and measure of relationship is called an emperial or *specific function.* Once functional relationship between the variables is specified, the function can then be used to predict the value of the dependent variable.

1. The two variables may represent any two interrelated variables, e.g., price of a product and its supply; consumption expenditure and income; wage rate and labour demand; interest rate and funds borrowed; input (labour) and output; distance travelled and airfare; and so on.

The relationship between the variables is specified and quantified by using a *statistical technique* called *regression.* The application of *regression technique* for measuring the relationship between the dependent and independent variables is discussed in detail in the next chapter. Here we use some simple functions involving two variables to show the use of functions in economic analysis and business decisions.

When regression technique is applied to data given in Table 4.1, it produces a pizza demand function of the following form.

$DP = 500 – 5\ PP$

…(4.4)

Equation (4.4) gives the weekly demand function for pizza. This function reveals the following facts: (*i*) at zero price, i.e., at *PP* = 0, pizza demand equals 500 units; (*ii*) the minus (–) sign indicates inverse relationship between pizza price and pizza demand, and (*iii*) number 5 gives the measure of relationship between pizza price and demand – it means that for each one-rupee change in price, demand changes by 5 units.

Application of demand function for predicting demand Once pizza demand function is estimated, the canteenwala can use the function for predicting the weekly demand for his pizza at a given price. For example, if he fixes pizza price at `60 per piece, he can find the weekly demand for pizza *(DP)* simply by substituting `60 for *PP* in Eq. (4.4) as shown below.

$DP = 500 – 5\ (60) = 200$

This shows that if the canteenwala fixes pizza price at `60, he would be able to sell 200 pizzas per week. This can be verified from the data given in Table 4.1. Similarly, if he fixes the price at `40 per piece, his weekly sale would rise to *DP* = 500 – 5 (40) = 300 pizzas.

Application The application of demand function in business decision can be illustrated with reference to Eq. (4.4). By using the demand function (4.4), the canteen owner can generate his total revenue (*TR*) at different prices. He finds that his *TR* at price of `60 and `40 is the same. So what price he would charge. Note that if he charges `40, he will have to increase pizza production. This requires extra cost of production, *TR* remaining the same. So his total profit will decrease. Therefore, he will chargè60 for maximizing his total profit.

4.3.2 General Forms of Economic Function

We have illustrated the formation and use of ***function*** in business decision-making with the help of an example of simple demand function. Let us now show the formation of same economic functions in their general form.

Demand Function Consider, for example, the *demand function*—one of the simplest and most widely used functions in demand analysis. A demand function expresses the relationship between the price of a commodity, say *X*, and its quantity demanded. Let the price of commodity *X* be denoted by *Px* and its quantity demanded by *Qx*. Here, *Qx* is the dependent variable and *Px* is the independent variable. The law of demand states that there is an inverse relationship between the demand for a normal good, *X*, and its price, *Px*. Assuming a constant inverse relationship between *Qx* and *Px*, the demand function for commodity *X* is written as

$$Qx = a - bPx \quad \ldots(4.5)$$

In demand function (Eq. 4.5), *a* and *b* are constants. Such constants in a function are called the *parameters* of the function. The *parameters* of a function specify the extent

of relationship between the *dependent* and the *independent variables*. The constant ‘ *a*’ gives the limit of quantity demand (*Qx*) when *Px* = 0. The constant ‘ *b*’ is the coefficient of variable *Px*, i.e., *b* measures the change in *Qx* as a result of a change in *Px* expressed as D *Q*/D *P*. Note that constant *b* carries a minus sign (–) which specifies the inverse relationship between the dependent variable, *Qx*, and the independent variable, *Px*.

Supply Function There is a positive relationship between the supply of a commodity, say *X*, and its price, *Px*. Therefore, the supply function of commodity *X* is expressed as

$$Qx = \alpha + \beta Px \quad \ldots(4.6)$$

where α = supply at zero price, and β = Δ *Qx*/Δ *Px*, i.e., the ratio of change in supply to change in price.

Some Other Functions Apart from *demand* and *supply functions*, economic theories or economic analyses, in general, trace the quantitative relationship between many other sets of dependent and independent economic variables having a *cause-and-effect kind of relationship* between them. The other important economic functions that will be used in the subsequent chapters are:

(*i*) *Production function*. Production function relates to the relationship between inputs (labour and capital) and output. Different types of production functions reveal the different *laws of production*.

(*ii*) *Cost function*. Cost function specifies the relationship between output and cost of production, i.e., the behaviour of cost in response to a change in output.

(*iii*) *Total revenue function*. Total revenue function is a combined function of quantity produced and price function derived from the demand function. It shows the behaviour of total revenue given the demand function.

(*iv*) *Profit function*. Profit function brings out the possible profit levels given the total revenue and total cost function. The profit function is obtained, in fact, by subtracting the total cost function from the total revenue function.

These functions will be discussed further in the subsequent sections and chapters. Let us now return to various forms of functions used in economic analysis and managerial decisions.

4.3.3 Kinds of Function

The kind of a function is known by the nature of relationship between the dependent and independent variables. Given the nature of relationship between the dependent and independent variables, the following kinds of functions are generally used in economic analyses:

1. Linear functions,
2. Nonlinear functions, and
3. Polynomial functions.

These kinds of functions are described here briefly and illustrated in respect of demand for a commodity.

Linear Functions

When a function produces a straight line, it is called a *linear function*. A linear function shows a constant relationship between the dependent and independent variables. The

relationship is *linear* in the sense that the change in the dependent variable remains constant throughout for a one unit change in the independent variable, irrespective of the level of the dependent variable. For example, suppose that demand function (4.5) is estimated as

$Qx = 20 – 2\ Px$

...(4.7)

This function implies that for each one rupee change in price (*Px*), the demand for commodity *X* changes by 2 units. This demand function can be converted into a table by assigning a numerical value (from 0–10) to *Px*. When a table is plotted, this demand function produces a line as shown in Fig. 4.2. Therefore, demand function (4.7) gives a ***linear demand function.***

The linear function (4.7) has only two variables: a dependent variable (*Qx*) and an independent variable (*Px*). But it does not mean that a linear function involves only two variables or it states the relationship between only two variables. A linear function may involve more than one independent variable. Such functions are called *polynomial functions*. For example, suppose the long-run demand function given in Eq. (4.3), is estimated and expressed (excluding A, T, E and U) as

$Dx = a – b\ 1\ Px + b\ 2\ Y + b\ 3\ Ps – b\ 4\ Pc$

In this case too, the demand function is

Fig. 4.2 *A Linear Demand Function*

a linear function.

Nonlinear Functions

A *nonlinear* or *curvilinear* function is one that produces a curvilinear relationship between the dependent and the independent variables. In a nonlinear function, the quantitative relationship between the dependent and the independent variables does not remain constant. It changes with the change in the level of the independent variable. A *nonlinear demand function* takes the form of a *power demand function* as follows.

D

– b

$x = aPx$

...(4.8)

The difference between a linear function and a nonlinear (or power) function can be understood better by comparing the two types of demand functions as given in Eqs. (4.5) and (4.8). Note that in the linear demand function (4.5), the *coefficient b* is the *coefficient of variable Px*, whereas in nonlinear demand function (4.8), power *b* is the exponent of variable *Px* and constant ‘ *a*’ is the coefficient of variable *Px*. A demand function as given in Eq. (4.8) produces a nonlinear or a curvilinear demand curve, as shown in Fig. 4.3. Suppose demand function (4.8) is given as

Fig. 4.3 *A Nonlinear Demand Function*

D

–2

x = 32 *Px*

…(4.9) **Table 4.2** *A Nonlinear Demand Schedule*

= 32/ *P* 2

x

Px

Dx (approx.)

The demand function (4.9) produces a

1

32.00

demand schedule as given in Table 4.2. The

2

8.00

demand schedule, presented graphically, produces

3

3.50

a demand curve as given in Fig. 4.3.

4

2.00

Some other forms of nonlinear demand

5

1.33

functions are following:

6

0.90

a

(*i*) *D*

a − *P*

x =

;

(*ii*) *D*

;

(*iii*)

bPx

P

D

ae−

=

+ *c*

x =

x

x

x

b

Polynomial Functions

In economic analysis, one encounters functions which contain many 'terms' of the same independent variable. Functions containing many 'terms' of the same independent variable are called *polynomial functions*—'poly' meaning 'many' and 'nomial' meaning 'names' or 'terms'. For example, consider the short-run production function (capital held constant) which is expressed in its general form as

$Q = f(L)$

where Q = output, L = labour input.

This short-run *production function* can take any of the following polinomial forms of the function producing curvilinear relationship between labour and output forms.

(*a*) A quadratic function:

$$Q = a + bL - cL^2 \quad \ldots(4.10)$$

(*b*) A cubic function: $Q = a + bL + cL^2 - dL^3$ ...(4.11)

(*c*) A power function: $Q = aL^b$ ...(4.12)

where Q = output, L = labour input, and *a, b, c* and *d* are constants coefficients.

The kind of total production (*TP*) curves these production functions produce are given in panels (*a*), (*b*), and (*c*) respectively, of Fig. 4.4.

Fig 4.4 *Nonlinear Production Functions*

The Degree of a Polynomial Function

The degree of a function is given by the highest *power* (the exponent) appearing in the function. Thus, the degree of a polynomial function is given by the highest *power* that appears in the polynomial equation.

The three types of production *functions* given above can be described as follows.

(1) Production function (4.10) has the highest power 2. It is, therefore, called a polynomial function of degree 2. A polynomial function of power 2 is also called a ***quadratic function.***

(2) Production function (4.11) has the highest power 3. It is, therefore, called, a polynomial function of degree 3. A function of power 3 is also called a ***cubic function***.

(3) Production function (4.12) is a power function. The range of power b in Eq. (4.12) may be expressed as $b < 1$, $b = 1$, and $b > 1$. That is, except zero, it can take any power.

Incidentally, a production function can also be a *linear* function of the form $Q = a + bL$. A linear function is also called a *polynomial of degree* 1. All linear functions are of a special type of polynomial function of degree 1. In economic analysis, however, polynomials of degrees 1, 2 and 3 are of greater significance.

Quadratic and cubic functions belong to the category of *nonlinear* or *curvilinear functions*.

The quadratic and cubic functions will be used in the analysis of production costs in Chapter 11.

4.3.4 Solving Quadratic and Cubic Functions

Quadratic and cubic functions can be solved by the *factoring method* and quadratic equations by *quadratic formula*. An interesting property of quadratic and cubic functions is that they have more than one solution. A quadratic equation has two solutions and a cubic function has three solutions.

Solution of Quadratic Function Factoring Method. Let us suppose that a quadratic function of the following form is given.

$$y = x^2 + x - 12 \quad \ldots(4.13)$$

The *first step* is to set the Eq. (4.13) equal to zero (0). Thus,

$$x^2 + x - 12 = 0 \quad \ldots(4.14)$$

The *second step* is to factor out the equation. Eq. (4.14) can be factored by expressing variable x as $4x - 3x$. By substituting $4x - 3x$ for x, Eq. (4.14) can be expressed as

$$x^2 + 4x - 3x - 12 = 0$$

$$(x + 4)(x - 3) = 0 \quad \ldots(4.15)$$

Equation (4.15) can hold only if any of the two bracketed terms equals zero. If $(x + 4) = 0$, then $x = -4$ and if $(x - 3) = 0$, then $x = 3$. Thus, there are two values of x that satisfy Eq. (4.14): these values are $x = -4$ and $x = 3$. Since a negative value ($x = -4$) has no meaning in economic analysis, the positive value ($x = 3$) is accepted as the answer.

Solution of Quadratic Function by Quadratic Formula. Equation (4.13) can also be solved by the quadratic formula given below.

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

Note that coefficients of x in Eq. (4.13) are not given. Suppose coefficient are given as *a*, *b* and *c* and $a = 1$, $b = 1$, and $c = -12$. Substituting these values in the quadratic formula, we get

2

$- 1 \pm 1 - 4(1)\ (-12)$

$x =$

2(1)

$- 1 \pm 7$

$x = - 1 \pm 49$ or $x =$

2

2

This solution of Eq. (4.13) implies that there are two values for *x* that satisfy the quadratic function:

6

(*i*) if $x = - 1 + 7$,

then

$x = = 3$

2

2

(*ii*) if $x = - 1 - 7$,

then

$x = - 8 = -4$

2

2

Note that the quadratic formula also provides the same two solutions as the factor method.

4.4 CONCEPT OF SLOPE AND ITS USE IN ECONOMIC ANALYSIS

The concept of slope is a very important tool of economic analysis. It is used for measuring relationship between marginal changes in two related variables and more importantly in the problems of optimization.

Geometrically, the term 'slope' refers to steepness of a line or curve showing the relationship between two variables. From economic analysis point of view, the **slope** gives the measure of *the rate of change in the dependent variable as a result of a change in the independent variable*. To understand the concept of slope, recall that when the relationship between a dependent and an independent variable is presented graphically, it produces a line or a curve. The slope of the line or the curve shows how strongly or weakly are the two variables related: the steeper the curve or line, the weaker the relationship and the flatter the curve or line, the stronger the relationship.

However, economic analysis of relationship between interrelated variables requires an exact measure of the slope of the line or the curve. With respect to the demand curve, slope is the ratio of change in the dependent variable (*D*) to the change in the independent variable (*P*). The movement down the demand line or the demand curve gives the decrease in price ($-\Delta P$) and the consequent increase in demand (ΔD). The ratio $-\Delta P/\Delta D$ gives the slope of the demand curve.2 The slope of a line remains constant throughout whereas slope of a curve changes all along the curve. Let us now look at the slopes of a *linear* and a *nonlinear function* in some detail.

2. For measuring the slope of an economic function, however, the change in dependent variable is used as numerator and the change independent variable used as demonimator. For example the slope of demand curve is measured as $\Delta D/\Delta P$.

4.4.1 Linear Function

Recall the linear demand function given in Eq. 4.7 given as

$D_x = 20 - 2P_x$

The demand curve arising out of this demand function is reproduced in Fig. 4.5. It shows a change in price (P_x) and the consequent change in demand (D_x). Suppose that the price is initially given at $P_x = 6$ and quantity demanded of commodity *X*, is 8, i.e., $D_x = 8$. Now let P_x decrease to 5. Then D_x increases to 10. With these changes in P_x and D_x, the change in price, $(-)\Delta P_x = 5 - 6 = (-)1$, and the resulting change in demand, $\Delta D_x = 10 - 8 = 2$. Note that the negative sign (–) denotes the inverse relationship between the dependent variable (*Dx*) and the independent variable (*Px*). Given the values of D *Px* = 1 (ignoring the minus sign) and D *Dx* = 2, the slope of the straight-line demand curve between points *J* and *K* can be obtained as follows.

Δ *Dx* 2

Slope of the demand curve =

= = 2

Δ *x*

P

1

Fig. 4.5 *The Slope of the Linear Demand Curve*

This means that the slope of the demand curve between points *J* and *K* equals 2.

Note that the slope of a straight-line is the same throughout the length of the line. For example, when *Px* decreases from 5 to 3, then *Dx* increases from 10 to 14. With these changes in *Px* and *Dx*, the change in price, (–)D *Px* = 3 – 5 = (–)2, and the resulting change in demand, D *Dx* = 14 – 10 = 4. Thus,

Slope = Δ *Dx* 4

= = 2

Δ *x*

P

2

4.4.2 Nonlinear Demand Curve

The slope of a nonlinear curve changes at each point of the curve. From analysis point of view, the slope of a curve needs to be measured (*i*) between any two points, and

(*ii*) at any two points and compared. For this purpose, let us recall our nonlinear demand function (4.9) given as

D

–2

2

x = 32 *Px* = 32/ *Px*

The nonlinear demand curve arising out of this demand function is presented graphically in Fig. 4.6 with some changes in *Px* and *Dx* for the purpose of measuring the slope of the curve. Figure 4.6 shows the variation in the slope between any two points all along the nonlinear demand curve.

Let us compare, for example, the slopes of a nonlinear demand curve *Dx* between two sets of different points: (*i*) *A* and *B*, and (*ii*) *C* and *D*. The slopes of the nonlinear demand curve between the two sets

of points are given below.

(*i*) Slope between points *A* and *B*

Δ *Dx* 2 − 1.3 0.7

=

=

=

Δ

= (–) 0.7

x

P

4 − 5

(−)1

(*ii*) Slope between points *C* and *D*

= Δ *Dx* 8 − 3.5

4.5

=

=

= (–) 4.5

Δ *x*

P

2 − 3

(−)1

Note that the slope of the nonlinear demand curve between points *A* and *B* and between points *C* and *D* is different from one another. The slope the curve between points *A* and *B* is

Fig. 4.6 *The Slope of a Nonlinear Demand Curve*

0.7 and between points *C* and *D*, it is 4.5. Thus, the slope of a nonlinear demand curve is different between any set of two points. A downward movement on the nonlinear demand curve in Fig. 4.6 shows that its slope goes on increasing.

More importantly, the slope of a nonlinear demand curve is different not only between a set of two points but also at each point on it. Consider, for example, points *B* and *C* on demand curve *Dx* in Fig. 4.6. The slope of the demand curve at point *B* can be measured by drawing a line tangent to the demand curve *Dx* at point *B* as shown in Fig. 4.6. The slope of the demand curve at point *B* is the same as that of the tangent. The slope of the tangent *equals* Δ *Dx*/Δ *Px* = 6/–6 = – 1. Therefore, the slope of the demand curve *Dx* at point *B* equals 1. By the same method, the slope of the demand curve at point *C equals* 10/4.5 = 2.22. This exercise can be repeated for any number of points on the demand curve, *Dx* and in each case the slope will be different. Thus, *a nonlinear curve, whether it has a negative or a positive slope, has a different slope throughout*.

4.4.3 Measuring Slope at a Point on a Curve

We have shown above how slope between any two points of a demand curve is measured. This kind of slope measurement has certain *limitations*.

First, the method of measuring slope of a curve between any two points is not very reliable if change in the independent variable (price) is *large* because slope is different between any set of two points and at each point within the chosen two points of the curve. For example, in Fig. 4.6 the slope of the demand curve between points *B* and *D* equals (8 – 2)/(2 – 4) = – 3. But the slope of the demand curve varies between any two other points within the range of *B* and *D*. For example, slope between points *B* and *C* equals – 1.5 and between points *C* and *D*, it equals – 4.5. Although the slope of demand curve at any point can be measured by drawing a tangent, as shown above, it is not an efficient method.

Second, the method of measuring slope, as given above, is not of much help in case an optimum solution to a business problem has to be found because an optimization problem may involve polynomial functions.

Therefore, a mathematical technique called ***differential calculus*** is used to measure the slope of a function. Some basic rules of differential calculus are given in the following section. Here we show the use of differential calculus for estimating the slope of the demand function given in Eq. (4.9) reproduced below.

D

–2

x = 32 *Px*

This demand function is shown graphically in the form of a demand curve in Fig. 4.6. The demand curve, as already mentioned, has a different slope at each point on it. The slope of the demand curve at any point can be measured by applying the ***differential calculus*** by the following method. Let us first take the first derivative of the demand function. Thus,

∂ *Dx*

3

−

64

= (2

−) 32 *x*

P = − 3

∂

…(4.15)

x

P

x

P

The Eq. (4.15) gives the slope of the demand curve drawn on the demand function (4.9).

Given the Eq. (4.15), the slope of the demand curve at any point can be easily obtained.

For example, suppose we want to measure slope at point *B* on the demand curve given in Fig. 4.6. At this point, price (*Px*) equals 4. By substituting 4 for *Px* in Eq. (4.15) we get

∂Dx

64

64

=

=

= 1

−

3

∂x

P

− (4) − 64

Note that this measure of the slope of the demand curve (– 1) at point *B* by using the method of differential calculus is the same (– 1) as measured by the tangent method.

The slope of the demand curve at any other point can be similarly measured.

4.5 APPLICATION OF DIFFERENTIAL CALCULUS

In the preceding section, we have shown the application of differential calculus in measuring the slope of a demand curve. The same technique can be applied to measure the slope of any curve based on any other function. Differential calculus is applied to analyze and to find solutions to a wide range of economic problems and to business decision-making, especially where an analyst or business decision-maker has to find an optimum solution to a problem. Therefore, before we discuss the tools and techniques of optimization, let us equip ourselves with basic rules of differential calculus.

4.5.1 Concept of Derivative

To explain the concept of derivative, let us recall the concept of *marginal cost* (*MC*).

We have so far defined *marginal cost* (*MC*) in terms of change in total cost (*TC*) as a result of a one unit change in total output (*Q*). The *MC*, so defined, can be expressed as

Change in *TC*

Δ *TC*

MC = Change in *Q* = Δ *Q*

…(4.16)

Here, D *Q* is assumed to be equal to 1. This method can be used, as already mentioned, only if the cost of producing each successive unit is precisely known or calculable.

This method cannot be used to measure the marginal change in *TC when* D *Q becomes increasingly small, so small that it is close to zero*. But business decision-makers are often faced with the problem as to *how to measure the marginal change in TC (or TR or profit) when* D *Q becomes increasingly small tending to zero*.

This problem is solved by a mathematical technique called *differential calculus*.

Suppose a function is given as

$Y = f(X)$

Given the function, differential calculus provides a technique of measuring the marginal change in the dependent variable, *Y*, due to a change in the independent variable, *X, when the change in X approaches zero*. The measure of such a marginal change is called *derivative*.

The concept of derivative can be defined as follows: *The derivative of a dependent variable Y, is the limit of change in Y when the change in the independent variable X, approaches zero*. This definition of derivative can be expressed more precisely as

δY

limit ΔY

δX =

…(4.17)

X

$\Delta \to 0 \; \Delta X$

The derivative expressed in Eq. (4.17) is read as the derivative of Y with respect to X equals the limit of the ratio D Y/D X as D X approaches zero. It is important to note that the *derivative* measures the *slope* of a function—linear or nonlinear— *at a point*.

The concept of 'derivative' can be further clarified through an example. Suppose a function is given as $Y = f(X)$ and the relation between Y (the dependent variable) and X (the independent variable) is such that when graphed, it produces a nonlinear curve as shown in Fig. 4.7. Let us begin assuming a big change in X and Y, e.g., movement from point A to B. As the figure shows, at point A, $X = X_1$ and $Y = Y_1$. Now let X, the independent variable, increase from X_1 to X_2. Consequently, Y increases from Y_1 to Y_2 and point A shifts to point B. As Fig. 4.7 shows, D $X = OX_2 – OX_1 = X_1X_2$ and D $Y = Y_2 – Y_1 = Y_1Y_2$. This is a big change, of course.

Given the changes in X and Y, i.e., ΔX and ΔY respectively, the slope of the function, $Y = f(X)$, can now be expressed as

Slope = ΔY

$Y\ Y$

1 2

=

...(4.18)

ΔX

$X\ X$

1

2

Recall that in Eq. (4.18), D $X = X_1X_2$ and the consequent D $Y = Y_1Y_2$ are large changes in X and Y. Therefore, the slope measured in Eq. (4.18) is the slope for large changes. But, what will be the slope of the function, $Y = f(X)$, if D $X \to 0$? This situation is shown in

Fig. 4.7. Note that when point X_2 shifts toward point X_1, as indicated by the arrow, then D X decreases and approaches zero. And then point B on the curve tends to reach point A. The problem here is, in fact, how to measure the slope of a function at a point or, realistically, when the change in independent and the dependent variables is *infinitesimally* small. The measure of slope of a function for an *infinitesimally small* change in the independent variables is called *derivative*. The derivative is computed by the method of *differentiation*

Fig. 4.7 *The Slope of a Nonlinear Function*

described below.

4.5.2 Rules of Differentiation

The kinds of functions that are encountered in managerial decisions are:

(*i*) constant function

(*ii*) power function

(*iii*) function as a sum or difference of two functions

(*iv*) function as a product of two functions

(*v*) function of quotients, and

(*vi*) function of a function.

We will confine the discussion here to the rules for differentiating these kinds of functions, i.e., rules of finding derivatives of these functions. For describing the rules of differentiation, we will use Y as the dependent variable, X as the independent variable, and a, b and c as constants.

1. *Derivative of a Constant Function.* The derivative of a constant function equals zero. For example, if

$Y = f(X) = a$ (where a is constant)

…(4.19)

δ

then

Y

$\delta X = 0$

The reason is that a constant function implies that whatever the value of X, the value of Y remains constant. That is, even if the value of X changes, the value of Y does not change. For example, if the optimum level of capital-labour ratio has been reached and capital is constant, then the production function can be written as

$Y = f(X) = 500$

where Y is output and X is labour.

Given this function, output will remain constant whatever the number of workers employed. This kind of production function produces a horizontal straight line when presented graphically. A production function of this kind is often found in Indian agriculture where there is disguised unemployment.

2. *Derivative of a Power Function.* A power function takes the following form.

$Y = f(X) = aX^b$

…(4.20)

where a and b are constants, a being the coefficient of X and b the power of X.

Given a power function in Eq. (4.20), the first derivative of Y with respect to X is obtained as follows.

δY

$\delta X = baX^{b-1}$

…(4.21)

Some examples of power functions and their derivatives are given below.

(*a*) If $Y = 5X^3$

δY

then

$\delta X = 3 \times 5 \times X^{3-1} = 15X^2$

…(4.22)

(*b*) If $Y = 4X^2$

δY

then

$= 2 \times 4 \times X^{2-1} = 8X$

…(4.23)

δX

(*c*) If $Y = 2X$

δY

then

$\delta X = 1 \times 2 \times X^{1-1} = 2X^0 = 2$

…(4.24)

(*d*) If $Y = X$

then

$\delta Y = X^{1-1} = 1$

…(4.25)

δX

(*e*) If $Y = 5X^{-2}$

δY

then

$\delta X = -2 \times 5 \times X^{-2-1} = -10X^{-3}$

…(4.26)

3. *Derivative of Functions of Sum and Difference of Functions.* A dependent variable Y may be the function of the sum (or difference) of two different functions of the same independent variable X or of a sum (or difference) of two other variables which are functions of X. The derivatives of such functions are given below.

Suppose total production (Y) of firm is a function of the sum of two different functions of its two production plants using the same independent variable, input labour (X). That is,

$$Y = f(X) + g(X) \qquad \ldots(4.27)$$

where $f(X)$ and $g(X)$ denote two different functional relationships between Y and X.

If a function is given as in Eq. (4.27), then the derivative of Y with respect to X is the sum of the derivative of the individual terms. That is,

$$\frac{\delta Y}{\delta X} = \frac{\delta f(X)}{\delta X} + \frac{\delta g(X)}{\delta X} \qquad \ldots(4.28)$$

similarly, if $Y = f(X) – g(X)$

(where $f(X)$ and $g(X)$ denote two different functions)

then $$\frac{\delta Y}{\delta X} = \frac{\delta f(X)}{\delta X} - \frac{\delta g(X)}{\delta X} \qquad \ldots(4.29)$$

For example, consider the following functions and their derivatives.

(i) If $Y = 5X + 2X^3$

then $$\frac{\delta Y}{\delta X} = 5X^{1–1} + 2 \times 3X^{3–1} = 5 + 6X^2 \qquad \ldots(4.30)$$

(ii) If $Y = 5X^2 – 2X^4$

then $$\frac{\delta Y}{\delta X} = 2 \times 5X^{2–1} – 4 \times 2X^{4–1} = 10X – 8X^3 \qquad \ldots(4.31)$$

(iii) If $Y = 4X^3 – 3X^2 + 3$

then $$\frac{\delta Y}{\delta X} = 3 \times 4X^{3–1} – 2 \times 3X^{2–1} + 0 = 12X^2 – 6X \qquad \ldots(4.32)$$

4. *Derivative of a Function as a Product of Two Functions.* *The derivative of a function as a product of two functions is equal to the first term (or function) multiplied by derivative of the second function plus second term (or function) multiplied by the derivative of the first function*. For example, suppose

$$Y = f(X) \times g(X) \qquad \ldots(4.33)$$

where $f(X)$ and $g(X)$ denote two different functional relationships between Y and X.

The derivative of function (4.33) can be expressed as

$$\frac{\delta Y}{\delta X} = f(X) \times \frac{\delta g(X)}{\delta X} + g(X) \times \frac{\delta f(X)}{\delta X} \qquad \ldots(4.34)$$

δX

δX

For example, if

$Y = 5X^2(4X + 3)$

…(4.35)

then $\delta Y = 5X^2(4) + (4X + 3)(10X) = 20X^2 + 40X^2 + 30X$

δX

$= 60X^2 + 30X$

…(4.36)

Alternative Method The derivative of function (4.35) can also be obtained by multiplying the function and taking its derivative. By multiplying the two terms in function (4.35), we get

$Y = 20X^3 + 15X^2$

…(4.37)

Note that after the two terms are multiplied, the function (4.35) gets transformed into a *power function* (4.37). So the derivative of function (4.37) can be obtained by the rule of differentiating a power function. Thus, the derivative of function (4.37) is give as

δY

$= 3 \times 20X^{3-1} + 2 \times 15X^{2-1}$

δX

$= 60X^2 + 30X$

…(4.38)

Note that the derivative given in Eq. (4.38) is the same as the one given in Eq. (4.36).

Consider another example. Suppose a function is given as follows.

$Y = (X^3 + 2X^2 + 3)(2X^2 + 5)$

…(4.39)

Then

$\delta Y = (X^3 + 2X^2 + 3)(4X) + (2X^2 + 5)(3X^2 + 4X)$

δX

$= (4X^4 + 8X^3 + 12X) + (6X^4 + 8X^3 + 5X^2 + 20X)$

$= 10X^4 + 16X^3 + 5X^2 + 32X$

…(4.40)

5. ***Derivative of a Quotient.*** If a function is in the form of a *quotient*, then *the derivative of the function is equal to the denominator times the derivative of the numerator minus the numerator times the derivative of the denominator*, whole divided by the square of the denominator. For example, suppose

$Y = f(X)$

…(4.41)

$g(X)$

$\delta f(X)$

$\delta g(X)$

δY

$g(X) \cdot$

$- f(X) \cdot$

Then

$=$

δX

δX

…(4.42)

δX

2

$[g(X)]$

Consider another example. Suppose

$Y = 5X + 4$

…(4.43)

$2X + 3$

Then the derivative of the function (4.43) is given as

$$\frac{\delta Y}{\delta X} = \frac{(2X+3)(5)-(5X+4)(2)}{(2X+3)^2} = \frac{(10X+15)-(10X+8)}{(2X+3)^2} = \frac{7}{(2X+3)^2} \quad \ldots(4.44)$$

6. *Derivation of a Function of a Function.* Sometimes business analysts come across a function in which a variable is a function of another variable which in itself is the function of yet another variable. For example, suppose $Y = f(U)$ and $U = f(X)$, and the analysis is required to find the derivative of Y with respect to X. In this case, the derivative Y with respect to X is equal to the derivative of Y with respect to U multiplied by the derivative of U with respect to X.

That is,

$$\frac{\delta Y}{\delta X} = \frac{\delta Y}{\delta U} \times \frac{\delta U}{\delta X} \quad \ldots(4.45)$$

For example, suppose that

$Y = U^3 + 5U$ and $U = 2X^2$

Here, the process is to find the derivative of Y with respect to U and the derivative of U with respect to X and then to multiply the two derivatives. Thus,

$$\frac{\delta Y}{\delta U} = 3U^{3-1} + 5 = 3U^2 + 5 \quad \ldots(4.46)$$

Since $U = 2X^2$, by substitution, Eq. (4.46) can be writtn as

$$\frac{\delta Y}{\delta U} = 3(2X^2)^2 + 5 = 3(2X^2)(2X^2) + 5 = 12X^4 + 5 \quad \ldots(4.47)$$

Let us now find the derivative of U with respect o X.

$$\frac{\delta U}{\delta X} = 2 \times 2X^{2-1} = 4X \quad \ldots(4.48)$$

We can now obtain $\delta Y/\delta X$ by multiplying the two derivatives given in Eqs. (4.47) and (4.48).

Thus,

$$\frac{\delta Y}{\delta X} = (12X^4 + 5)(4X) = 48X^5 + 20X \quad \ldots(4.49)$$

4.5.3 Partial Derivative

Functions with Several Independent Variables We have so far described the rules of differentiating functions with one independent variable. Many functions used in economic and business analyses have more than one independent variable. Some common examples of such functions are demand function, production function and cost function. For example, consider some functions with several independent variables.

Demand function: $Dx = f(Px, Ps, Pc, Rd, A, T, \ldots \text{etc.})$

where Dx = demand for commodity X, Px = price of X, Ps = price of substitutes, Pc = price of complements, Rd = disposable resources (income) A = advertisement expenditure by producers, T = tastes and preferences, etc.

Production function: $Q = f(K, L)$

where Q = quantity produced, K = quantity of capital, and L = number of workers.

Cost function: $C = f(K, r, L, w)$

where C = total cost, K = capital, r = rental rate, L = number of workers, w = wage rate.

Rules of Partial Differentiation We have noted some functions above which involve two or more independent variables. We will, however, describe the rules of partial differentiation in terms of Y as dependent variable and X and Z as two independent variables. Suppose $Y = f(X, Z)$ and the functional relationship between Y (the dependent variable) and the independent variables, X and Z, is given as

$$Y = X^3 + 4XZ + 5Z^2 \quad \ldots(4.50)$$

The rules of partial differentiation can be stated as follows:

(*i*) Only one of the independent variables is allowed to change at a time and all other independent variables are held constant.

(*ii*) For differentiating the dependent variable with respect to one independent variable, the rule of differentiation is followed.

Based on these rules, the derivatives of Y with respect to X and Z in Eq. (4.50) are given below.

(*i*) Derivative of Y with respect to X with Z held constant,

$$\frac{\delta Y}{\delta X} = 3X^{3-1} + 4Z = 3X^2 + 4Z \quad \ldots(4.51)$$

(*ii*) Derivative of Y with respect to Z with X remaining constant,

$$\frac{\delta Y}{\delta Z} = 4X + 2 \times 5Z^{2-1} = 4X + 10Z \quad \ldots(4.52)$$

In case a function with two (or more) independent variables is in the multiplicative form, such as

$$Y = aX^b Z^c$$

then the derivative of Y with respect to X is obtaind as

$$\frac{\delta Y}{\delta X} = baX^{b-1} Z^c \quad \ldots(4.53)$$

and the derivative of Y with respect to Z is

$$\frac{\delta Y}{\delta Z} = aX^b cZ^{c-1} = acX^b Z^{c-1} \quad \ldots(4.54)$$

To conclude, the tools of analysis discussed above equip us with basic analytical techniques that are often used in business analysis and business decisions. These analytical tools will be frequently used in the following chapters of this book. In the next section, we will discuss a more important decision tool that is, the techniques of optimization.

4.6 OPTIMIZATION TECHNIQUES

Having described the rules of decision and tools of analysis, especially the rules of differentiation, in the preceding section, we will describe in this section the more important techniques of managerial decision making, i.e., the **techniques of optimization.** Optimization means finding best results under the given conditions, given the objecitve. An optimization technique is a technique of maximizing or minimizing a function given the objecitve. In simple words, it is a technique of finding the value of the independent variable(s) that maximizes or minimizes the value of the dependent variable. For example, some firms may be interested in finding the level of output that maximizes their total revenue; some firms facing a constant price may want to find the level of output that would minimize the average cost; and most important of all, most firms may be interested in finding the level of output that maximizes their profits.

4.6.1 Technique of Maximizing Total Revenue

The total revenue (*TR*) of a firm is defined as

$TR = P \cdot Q$

...(4.55)

where P = price and Q = quantity sold.

Suppose a price function is given as

$P = 500 – 5 Q$

...(4.56)

By substituting Eq. (4.56) into Eq. (4.55), we get *TR* as follows.

$TR = (500 – 5 Q) Q$

$= 500 Q – 5 Q 2$

...(4.57)

Now the problem is to find the value of Q that maximizes total revenue.

The Rule of TR Maximization

By rule, total revenue is maximum at the level of sales (*Q*) at which $MR = 0$. That is, for total revenue to be maximum, the marginal revenue (*MR*), i.e. the revenue from the sale of the marginal unit of the product, must be equal to zero. *MR* is given by the first derivative of the *TR* function. So, to find the value *Q* that maximizes *TR*, we need to find the derivative of the *TR* function (4.57) with respect to *Q*; set it equal to zero and solve it for *Q*, as shown below.

Given the *TR* function as

$TR = 500 Q – 5 Q$

TR

δ

$MR = δ = 500 – 10 Q$

...(4.58)

Q

By setting Eq. (4.58) equal to zero and solving for *Q*, we get

$500 – 10 Q = 0$

$– 10 Q = – 500$

$Q = 50$

...(4.59)

Equation (4.59) implies that at $Q = 50$, $MR = 0$ and, therefore, it maximizes the total revenue.

The maximum *TR* can be obtained by substituting 50 for *Q* in the *TR* function (4.57). Thus,

$TR = 500 (50) – 5 (50)2$

...(4.60)

$= 25,000 – 12,500 = 12,500$

Let us now check the result. Whether *TR* = `12,500 is maximum can be checked by increasing and decreasing *Q* by one unit and then comparing *TR* at $Q = 51$ and at $Q = 49$ with *TR* at $Q = 50$.

TR (at $Q = 51$) = 500(51) – 5(51)2 = 25,500 – 13,005 = 12,495

TR (at $Q = 49$) = 500(49) – 5(49)2 = 24,500 – 12,005 = 12,495

These estimates of *TR* show that if sales are increased above 50 units or reduced below 50 units, *TR* decreases in both the cases. Thus, it is proved that $Q = 50$ maximizes *TR*.

4.6.2 Technique of Optimizing Output: Minimizing Average Cost

The optimum output of a firm is one that minimizes its average cost of production. The optimum output determines the most efficient size of the firm. A prior knowledge of the optimum size of the firm is very important for future planning under at least three conditions.

One, a businessman planning to set up a new production unit would like to know the optimum size of the plant for future planning. This problem arises because, as the theory of production tells us, the average cost of production in most productive activities decreases to a certain level of output and then begins to increase.

Two, the firms planning to expand their scale of production would like to know the most efficient level of the economies of scale so that they are able to plan the marketing of the product accordingly.

Three, firms working in a competitive market are often constrained in raising the price for raising their total revenue and profit. Their profit then depends on their ability to reduce their unit cost of production. Given the technology and input prices, the prospect of reducing the average cost of production depends invariably on the level of production. The problem that decision makers might face in this regard is how to find the *optimum level* of output, i.e., the level of output that minimizes the average cost of production.

As already mentioned, under the general production conditions, the optimum level of output is the one that minimizes the average cost. The average cost (AC) can be obtained as:

$$AC = \frac{TC}{Q} \quad \text{...(4.61)}$$

Suppose total cost (TC) function of a firm is given as

$$TC = 400 + 60Q + 4Q^2 \quad \text{...(4.62)}$$

Then

$$AC = \frac{400 + 60Q + 4Q^2}{Q} = 400/Q + 60 + 4Q \quad \text{...(4.63)}$$

Now the problem is how to find the value of Q that minimizes AC.

The Rule of Minimization. Like the rule of maximization, *the rule of minimizing a function is to find and equalize its first derivative to zero*. Thus, the value of Q that minimizes AC can be obtained by finding the first derivative of the AC function (4.63) and setting it equal to zero and solving it for Q.

The derivative of the AC function (4.63) is given as

$$\frac{\delta AC}{\delta Q} = -400/Q^2 + 4 \quad \text{...(4.64)}$$

By setting Eq. (4.64) equal to zero, we get

$$-400/Q^2 + 4 = 0 \quad -400/Q^2 = -4$$

$$Q^2 = -400/-4 = 100$$

$$Q = 10$$

The result shows that $Q = 10$ minimizes the average cost. In other words, the optimum size of the output is 10 units. Any other output will increase the average cost of production.

4.6.3 Profit Maximizing Conditions

Total profit (Π) is defined as

$$\Pi = TR - TC$$

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* or the *first order condition,* and (*ii*) *secondary or supplementary condition*.

(i) The necessary or the first-order condition requires that marginal revenue (MR) must be equal to marginal cost (MC), i.e., $MR = MC$. By definition, marginal revenue is the revenue 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.

(ii) The secondary or the second-order condition requires that the necessary or first-order condition must be satisfied under the condition of decreasing MR and rising MC. The fulfilment of the two conditions makes it the *sufficient condition*.

The profit maximizing conditions can also be presented algebraically as follows. We know that a profit maximizing firm seeks to maximize

$$\Pi = TR - TC \qquad \ldots(4.65)$$

Let us suppose that the total revenue (TR) and total cost (TC) functions are, respectively, given as

$TR = f(Q)$ and $TC = f(Q)$

where Q = quantity produced and sold.

By substituting total revenue and total cost functions in Eq. (4.65), the profit function may be written as

$$\Pi = f(Q)_{TR} - f(Q)_{TC}$$

This can now be manipulated to illustrate the *first* and *second* order conditions of profit maximization as follows.

First-order condition The first-order condition of maximizing a function requires 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. (4.65) must be equal to zero. Differentiating the total profit function and setting it equal to zero, we get

$$\frac{\partial \Pi}{\partial Q} = \frac{\partial TR}{\partial Q} - \frac{\partial TC}{\partial Q} = 0$$

This condition holds only when

$$\frac{\partial TR}{\partial Q} = \frac{\partial TC}{\partial Q}$$

In this equation, the term $\partial TR/\partial Q$ gives the slope of the TR curve which gives the marginal revenue (MR). Similarly, the term $\partial TC/\partial Q$ gives the slope of the total cost curve which 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 necessary condition is one that *must* be satisfied for an event to take place. In other words, the condition that $MR = MC$ must be satisfied for profit to be maximum.

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 Fig. 4.8. The MC and MR curves are the usual marginal cost and marginal revenue curves respectively. Incidentally, MC and MR curves are derived from TC and TR functions, respectively. (The method of derivation has been shown later in the book). MC and MR curves intersect at two points, P_1 and P_2. Thus, the firstorder 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

Fig. 4.8 *Marginal Conditions of Profit Maximization* that the second derivative of the profit function is negative. The second derivative of the total profit function is given as

$$\frac{\partial^2 \Pi}{\partial Q^2} = \frac{\partial^2 TR}{\partial Q^2} - \frac{\partial^2 TC}{\partial Q^2} \qquad \ldots(4.66)$$

The second-order condition requires that the second derivative of the second derivative of the *TR* function is *less* than that of the *TC* function. That is,

$$\frac{\partial^2 TR}{\partial Q^2} - \frac{\partial^2 TC}{\partial Q^2} < 0$$

or

$$\frac{\partial^2 TR}{\partial Q^2} < \frac{\partial^2 TC}{\partial Q^2}$$

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 may also be written as

Slope of *MR* < Slope of *MC*

It implies that the slope of *MC* must have higher than that of and *MC* must intersect the *MR* from below.

To conclude, profit is maximized where both the first and second order conditions are satisfied.

4.6.4 Algebra of Profit Maximization

We may now apply the profit maximization conditions to a hypothetical example and compute profit maximizing output.

Suppose *TR* and *TC* functions are given, respectively, as follows.

$$TR = 600Q - 3Q^2 \qquad \ldots(4.67)$$

$$TC = 1000 + 100Q + 2Q^2 \qquad \ldots(4.68)$$

Given the *TR* and *TC* functions as in Eq. (4.67) and Eq. (4.68), respectively, *MR* and *MC* can be obtained as follows.

$$MR = \frac{\delta TR}{\delta Q} = 600 - 6Q \qquad \ldots(4.69)$$

δ

and

MC = *TC* = 100 + 4 *Q*

…(4.70)

δ *Q*

By applying the *first* ***order condition*** of profit maximization, we get maximum profit

where

MR = *MC*

600 – 6 *Q* = 100 + 4 *Q*

…(4.71)

500 = 10 *Q*

Q = 50

The first order condition of profit maximization reveals that given the *TR* and *TC*

functions, the total profit is maximum at *Q* = 50.

Let us now apply the ***second order*** condition. Given the first order derivative of the

TR function in Eq. (4.69) and that of the *TC* function in Eq. (4.70), the *second derivatives*

of the *TR* and *TC* functions are given below.

2

δ *TR* δ *MR*

2

=

δ *TC* δ *MC*

=

2

δ

= – 6 and

= 4

Q

δ *Q*

2

δ *Q*

δ *Q*

Note that the second derivative of the *TR* function equals – 6 and the second derivative

of the *TC* function equals 4. The second order condition [Eq. (4.66)] requires that the

sum of the derivatives of *MR* and *MC* functions must be less than zero. As shown above,

the sum of the two second derivatives, i.e., – 6 + 4 = – 2 and – 2 < 0. So the second

order condition of profit maximization is also satisfied at *Q* = 50.

Is Total Profit Maximum at Q = 50? Whether total profit (Π) at *Q* = 50 is maximum

can be checked by comparing profits at *Q* = 50, at *Q* > 50 and at *Q* < 50. By substituting

these numbers by turn into the profit function, we can get the total profit at three levels

of output. Let us first work out the profit at *Q* = 50.

Total profit (Π) at *Q* = 50 = *TR* – *TC*

By substituting Eq. (4.67) for *TR* and Eq. (4.68) for *TC*, we get

= (600 *Q* – 3 *Q* 2) – (1000 + 100 *Q* + 2 *Q* 2)

= [600(50) – 3 (50)2] – [1000 + 100(50) + 2(50)2]

= 22,500 – 11,000 = 11,500

Let us now work out profit at a higher and lower output. Total profit P (at *Q* = 51)

= [600(51) – 3 (51)2] – [1000 + 100 (51) + 2(50)2]

= 22,797 – 11,302 = 11,495

Total profit Π (at *Q* = 49)

= [600(49) – 3(49)2] – [(1000 + 100(49) + 2(49)2]

= 22,197 – 10,702 = 11,495

The foregoing calculations show that at *Q* = 50 total profit equals `11,500. And,

when *Q* is increased or decreased even by a single unit, the total profit decreases by `5

in either case. This proves that profit is maximized at 50 units of output.

4.6.5 Optimization of Output by Profit Function: An Alternative Method

We have shown above how profit maximizing output can be obtained by using derivatives of *TR* and *TC* functions. Alternatively, profit maximizing output can be obtained by finding out profit function and maximizing it with respect to output. Profit function can be derived from *TR* and *TC* functions as shown below.

Given the *TR* and *TC* functions as in Eqs. (4.67) and (4.68), respectively, the profit function may be specified as follows

$\Pi = 600Q - 3Q^2 - (1000 + 100Q + 2Q^2)$

$= 600Q - 3Q^2 - 1000 - 100Q - 2Q^2$

$= -1000 + 500Q - 5Q^2$

...(4.72)

Going by the profit maximization rule, the derivative of the profit function (4.72) must be equal to zero. The derivative of the profit fuction is

$\frac{\delta\Pi}{\delta Q} = 500 - 10Q$

...(4.73)

For profit to be maximum, the *first derivative* of the profit function given in Eq. (4.73) must be equal to zero. That is,

$500 - 10Q = 0$

$Q = 50$

Note that the profit maximizing output obtained by the alternative methods is the same, i.e., $Q = 50$

Optimization of a Multivariate Profit Function

We have so far discussed the output optimization by using a profit function with one independent variable. In reality, however, decision-makers deal with functions with more than one independent variable. For example, output is the function of two independent variable inputs —labour and capital; total revenue is the function of quantity sold, price and advertisement expenditure. In case of firms producing more than one commodity, i.e., multi-product firms, profit is function of all the products. In this section, we explain the technique of optimization of a multivariate function assuming a simple case of two independent variables. One common case is that of profit maximization by a firm producing two commodities. We explain here the profit maximization technique assuming a case of two products. In this case, the problem is to find the optimum outputs of both the products that maximize the profit.

Suppose profit function with two products, say *X* and *Y*, is given

$\Pi = f(X, Y)$

and that the profit function is given as follows.

$\Pi = 100X - 2X^2 - XY + 180Y - 4Y^2$

...(4.74)

Maximization of a multivariate profit function (4.74) requires that *partial derivative of* P function with respect to *X*, and *partial derivative* of P with respect to *Y* are each set equal to zero and solved for *X* and *Y*.

The partial derivative of P function with respect to *X*, holding *Y* constant, is derived as

$\frac{\delta\Pi}{\delta X} = 100 - 4X - Y$

...(4.75)

and partial derivative of Π function with respect to *Y*, holding *X* constant, is derived as

$\frac{\delta\Pi}{\delta X} = 180 - X - 8Y$

...(4.76)

Setting each of the partial derivatives given in Eqs. (4.75) and (4.76) equal to zero, we get

(*i*) $100 - 4X - Y = 0$

(*ii*) $180 - X - 8Y = 0$

Note that setting each partial derivative equal to zero results in two simultaneous equations. By solving these equations we can find the values of *X* and *Y* that maximize the profit function. To solve the equations, we need to eliminate one of the variables (say, X). For this, let us multiply Eq. (*ii*) by 4 and subtract the product from Eq. (*i*). Thus, we have

(*iii*) $100 - 4X - Y = 0$ [Eq. (*iii*) is the same as Eq. (*i*)]

(*iv*) $4(180 - X - 8Y) = 0 = 720 - 4X - 32Y = 0$

By subtracting Eq. (*iv*) from Eq. (*iii*), we get

$-620 + 31Y = 0$

$Y = 20$

By substituting 20 for Y in Eq. (*i*) we can obtain the value of X.

$100 - 4X - 20 = 0$

$-4X = -80$

$X = 20$

The foregoing calculations show that the firm can maximize its profit function by producing 20 units each of its products X and Y. The maximum profit can be worked out by substituting the values of X and Y in the profit function. The estimate of total profit is shown below by using profit function (4.74). To work out profit, variables X and Y are replaced by their respective values (20). Thus,

$\Pi = 100(20) - 2(20)^2 - (20)(20) + 180(20) - 4(20)^2$

$= 2000 - 800 - 400 + 3600 - 1600$

$= 2800$

Any other combination of X and Y will reduce the profit. Whether the profit is maximum can be checked by substituting any value other than 20 for X and Y.

4.7 CONSTRAINED OPTIMIZATION

The optimization techniques discussed above can be called *unconstrained* optimization techniques, as these techniques are based on the assumption that firms operate under no *constraints* on their activity. For example, in the case of output maximization, it is implicily assumed that firms face no resource constraints; they possess unlimited resources; and that they can acquire all the inputs, finance, capital equipment, labour and raw materials that they need to maximize output. Same is the case with cost minimization technique. The firms are supposed to have all the resources to carry out production activity until *average cost* is minimized or cost for a given output is minimized. In real business world, however, managers face serious resource *constraints*. For example, they need to maximize output with given quantity of capital and labour time. The technique that is used to optimize the business objective(s) under resource constraints are called *constrained optimization techniques*. There are three very common techniques of constrained optimization, *linear programming*, *constrained optimization by substitution* and *Lagrangian multiplier*. The linear programming technique has a wide range of application and requires a detailed discussion. It has been discussed in detail in chapter 12 with reference to constrained output maximization. Here, we discuss the two other techniques of constrained optimization.

4.7.1 Optimization by Substitution Technique

We will illustrate the application of substitution method of constrained optimization problem to a profit maximization and a cost minimization problem.

(i) Constrained Profit Maximization. Let us recall our earlier example of profit maximization by using a profit function given in Eq. (4.74), reproduced below as Eq. (4.77) for ready reference.

$$\Pi = 100X - 2X^2 - XY + 180Y - 4Y^2 \quad \ldots(4.77)$$

In the preceding section, we have illustrated the maximization of this profit function *without any constraint*. That is, the independent variables, X and Y, could take any value in the profit maximization solution.

Here we illustrate the maximization of the same profit function with a constraint on output that the sum of X and Y must be equal to 30 (instead of the sum of X and Y being 20 + 20 = 40 as found in the solution without constraints). The constraint is expressed as

$$X + Y = 30 \quad \ldots(4.78)$$

A constrained problem of this kind can be solved by *substitution method* as illustrated below. The process of solution involves two steps: (*i*) express one of the variables in terms of the other and solve the constraint equation for one of the variables (*X* or *Y*) and (*ii*) substitute the solution into the objective function to be maximized and solve it for the other variable.

Given the constraint Eq. (4.78), the values of *X* and *Y* can be expressed in terms of one another as follows.

$$X = 30 - Y \text{ or } Y = 30 - X$$

We can now substitute the value of *X* (or *Y*) into Eq. (4.77) and find the maximization solution.

By substituting the value of *X*, the profit function (4.77) can be expressed as

$$\Pi = 100(30 - Y) - 2(30 - Y)^2 - (30 - Y)Y + 180Y - 4Y^2$$

$$= 3000 - 100Y - 2(900 - 60Y + Y^2) - 30Y + Y^2 + 180Y - 4Y^2$$

$$= 3000 - 100Y - 1800 + 120Y - 2Y^2 - 30Y + Y^2 + 180Y - 4Y^2$$

$$= 1200 + 170Y - 5Y^2 \quad \ldots(4.79)$$

Note that the *substitution method* converts a constrained problem into an unconstrained one. Equation (4.79) can now be maximized by obtaining its derivative and setting it equal to zero and solving it for *Y*.

Since

$$\Pi = 1200 + 170Y - 5Y^2,$$

$$\frac{\delta \Pi}{\delta X} = 170 - 10Y \quad \ldots(4.80)$$

By setting Eq. (4.80) equal to zero, we get

$$170 - 10Y = 0 \text{ and } Y = 17$$

By substituting 17 for *Y* in constraint Eq. (4.78), we get

$$X + 17 = 30 \text{ and } X = 13$$

Thus, the optimum solution of the profit maximization problem lies in producing $X = 13$ and $Y = 17$. These values of *X* and *Y* satisfy the constraint. In simple words, we get the optimum solution that the firm maximizes its profit by producing 13 units of *X* and 17 units of *Y*. The answer will be the same if we substitute $30 - X$ for *Y* in Eq. (4.77) and solve the equation for *X*.

Now let us compute the maximized profit under constraints. This can be done by substituting the values of *X* and *Y* into the profit function (4.77). By substitution, we get Eq. (4.77)

$$\Pi = 100(13) - 2(13)^2 - (13)(17) + 180(17) - 4(17)^2 = 2645$$

Note that maximum profit (2645) under constraint is less than the maximum profit under no constraint (2800).

(ii) Constrained Cost Minimization. Let us now apply the *substitution method* of optimization to a problem of constrained cost minimization. Suppose that cost function of a firm producing two goods, *X* and *Y*, is given as

$$TC = 2X^2 - XY + 3Y^2$$

and the firm has to meet a combined order of 36 units of the two goods. The manager's problem is to find an optimum combination of goods *X* and *Y* that minimizes their cost of production. The problem can be restated formally as

Minimize

$$TC = 2X^2 - XY + 3Y^2 \quad \ldots(4.81)$$

Subject to

$$X + Y = 36 \quad \ldots(4.82)$$

Substitution method requires that the constraint Eq. (4.82) is expressed in terms of any one of the two goods and then substituted into the *objective function* (4.81). By expressing *X* in terms of *Y*, we get

$$X = 36 - Y \quad \ldots(4.83)$$

By substituting Eq. (4.83) for *X* in the objective function (4.81), we get

$TC = 2(36 – Y)2\ (36 – Y)\ Y + 3\ Y\ 2$

$= 2(1296 – 72\ Y + Y\ 2) – 36\ Y + Y\ 2 + 3\ Y\ 2$

$= 2592 – 144\ Y + 2\ Y\ 2 – 36\ Y + Y\ 2 + 3\ Y\ 2$

$= 2592 – 180\ Y + 6\ Y\ 2$

...(4.84)

For the objective function (4.84) to be minimized, its first derivative must be set to zero. Thus,

δ *TC*

δ $Y = – 180 + 12\ Y = 0$ ...(4.85)

Solving Eq. (4.85) for *Y*, we get

$12\ Y = 180$ and $Y = 15$

By substituting the value of *Y* in the constraint Eq. (4.82), we get

$X + 15 = 36$ and $X = 21$

Thus, we get the optimum solution that $X = 21$ and $Y = 15$ minimize the cost of meeting the order. The minimum cost of producing 21 units of *X* and 15 units of *Y* can be obtained by substituting these values in cost function (4.81).

Minimum $TC = 2(21)2 – (21)\ (15) + 3\ (15)2 = 882 – 315 + 675 = 1242$

4.7.2 Lagrangian Multiplier Method

A sophisticated method of solving constrained optimization problems is the Lagrangian multiplier method. This method is used to solve the optimization problems of a complex nature and those which cannot be solved by the substitution method. We will, however, illustrate the Lagrangian multiplier method in respect of two optimization problems of general nature.

(*i*) a constrained profit maximization problem and

(*ii*) a constrained cost minimization problem.

(i) Constrained Profit Maximization. To illustrate the application of the Lagrangian multiplier to solve a profit maximization problem, we will use the profit function given in Eq. (4.77) with the same constraint imposed.

Let us restate the problem as

Maximize

$\Pi = 100\ X – 2\ X\ 2 – XY + 180\ Y – 4\ Y\ 2$

...(4.86)

Subject to the constraint

$X + Y = 30$

...(4.87)

The basic approach of the Lagrangian multiplier method is to form a Lagrangian function by combining the objective function and the constraint equation and then solving it by the partial derivative method. There is a simple technique of formulating the Lagrangian function.

First, set the constraint Eq. (4.87) equal to zero, i.e.,

$X + Y – 30 = 0$

Second, multiply the resulting equation by l (the Greek letter ‘lambda’ called *Lagrangian muliplier*) as shown below.

$\lambda(X + Y – 30)$

This equation gives the **Lagrangian** constraint. By adding the Lagrangian constraint to the objective function (4.86), one gets the **Lagrangian function**.

Thus, a **Lagrangian function** is formed as

$L\Pi = 100\ X – 2\ X\ 2 – XY + 180\ Y – 4\ Y\ 2 + \lambda(X + Y – 30)$

...(4.88)

Equation (4.88) is the unconstrained Lagrangian function with three unknowns, *X*, *Y* and l. The values of *X*, *Y* and l that maximize the degrangian function (*L* P) maximize also the profit function (p). The Greek letter l is the *Lagrangian multiplier*. It gives the measure of a small change in the constraint on the objective function.

In order to maximize the *L* P function (4.88), partial derivative of *L* P is obtained with respect to *X*, *Y* and l and each of them, is set equal to zero. This gives the first order condition of profit maximization in the form of three simultaneous equations, as shown below.

δ L

Π = 100 – 4 X – + l = 0

...(4.89)

δ X

δ L

Π = – X + 180 – 8 + l = 0

...(4.90)

δ Y

δ TC = X + Y – 30 = 0

...(4.91)

δ λ

By solving these simultaneous equations, we get the values of X, Y and l that maximize the objective function (4.88). In order to solve these equations for X, Y and l, methodologically three simultaneous equations (4.89), (4.90) and (4.91) have to be reduced to two equations. To do this, let us rearrange the terms of Eq. (4.90) and subtract it from Eq. (4.89). By subtracting we get,

100 – 4 X – Y + l = 0

(–) 180 – X – Y + l = 0

– 80 – 3 X + 7 Y + 0 = 0

...(4.92)

Now we have two simultaneous equations, i.e., Eqs. (4.91) and (4.92) which can now be solved to find the value of X and Y. Let us first rearrange the values of Eq. (4.92) to match with Eq. (4.91) and set them together. Now by using the method of solving the simultaneous equations, we multiply Eq. (4.91) by 3 and add it to Eq. (4.92). By solving the simultaneous equations we get,

3 X + 3 Y – 90 = 0

– 3 X + 7 Y – 80 = 0

---

10 Y – 170 = 0

Y = 17

By substituting 17 for Y in the constraint Eq. (4.87), we get the value of X as

X + 17 = 30 or X = 13

Note that the values of X and Y are the same as computed by substitution method.

The value of l can be obtained by substituting the values of X and Y in Eq. (4.89) or in Eq. (4.90). Using Eq. (4.89), we get,

– 13 + 180 – 8(17) + λ = 0 and λ = – 31

The value of l (minus sign ignored) has an important economic interpretation. It gives the measure of the change in the total profit when the output constraint is changed by 1 unit. For example, if output is increased by 1 unit, i.e., from 30 to 31 units, profit will increase by about 31 and if output is decreased by 1 unit, i.e., from 30 to 29 units, the profit will decrease by about 31.

(ii) Constrained Cost Minimization. Suppose an automobile firm has to supply 500 cars of two brands, X and Y to a German buyer. The joint cost function for the two varieties of carpets is given as

C = 100 X 2 + 150 Y 2

...(4.93)

The quantity of X and Y are not specified in the order. So the firm is free to supply X and Y in any combination. The firm’s problem is to find the combination of X and Y that minimizes the cost of production subject to the constraint X + Y = 500. The problem can be restated formally as

Minimize C = 100 X 2 + 150 Y 2

Subject to

X + Y = 500

...(4.94)

In order to solve the cost minimization problem by Lagrangian multiplier method, the problem has to be converted into a Lagrangian function. The procedure is to set the constraint Eq. (4.94) equal to zero, multiply it by l and add the result to the objective function. The cost minimization problem converted into the Lagrangian function is given below.

Minimize

$Lc = 100X^2 + 150Y^2 + (500 - X - Y)$

...(4.95)

Subject to

$500 - X - Y = 0$

The objective here is to minimize Eq. (4.95) subject to $X + Y = 500$. The first order condition requires that the derivative of Lc with respect to X, Y and l is set equal to zero. Thus,

δL

c

$\delta X = 200X - \lambda = 0$

...(4.96)

δ

$Lc = 300Y - \lambda = 0$

...(4.97)

δY

δL

$c = 500 - X - Y = 0$

...(4.98)

$\delta \lambda$

By using the Lagrangian method, we subtract Eq. (4.97) from Eq. (4.96) and, we get

$200X - \lambda - (300Y - \lambda) = 0$

$200X - \lambda - 300Y + \lambda = 0$

Thus, we get $200X - 300Y = 0$. It means $200X = 300Y$

$X = 1.5Y$

...(4.99)

By substituting 1.5 Y for X in Eq. (4.98), we get

$500 - 1.5Y - Y = 0$. It gives $500 = 2.5Y$

$Y = 200$

By substituting the value of Y with 200 in the constraint Eq. (4.94), we get

$X + 200 = 500$ and $X = 300$

Thus, the solution to the cost minimization problem is that the automobile company should produce $X = 300$ and $Y = 200$ units. This will minimize the cost of producing 500 cars. The minimum cost can be worked out as follows

$C = 100X^2 + 150Y^2 = 100(300)^2 + 150(200)^2$

$= 9{,}000{,}000 + 6{,}000{,}000 = 15{,}000{,}000$

Thus the minimum cost of supplying 500 cars works out tò15 million. This is the minimum cost because any other combination of X and Y brands of car will make the cost exceed `15 million.

SUMMARY

• This chapter explains briefly certain concepts and tools of economic analysis and technique of optimization of a value.

• *Economic variables* refer to the quantity, value and rate of goods that are subject to variation automatically or due to change in its determinant variable. Economic variables are classified as dependent and independent variables.

• Most economic variable are interrelated and interdependent. The relationship between interrelate economic variables are expressed in the form a function. The function takes the form of an equation. In general, economic functions are estimated by measuring the nature and extent of relationship between the interdependent variable by using a statistical technique called regression.

• The relationship between the two variables can be shown by a line or a curve. The *slope* of a line or curve shows the nature and extent of relationship between the variable. For example, given the demand curve, if price of a commodity changes by some amount, what is the change in demand for the commodity.

• Calculus is a mathematical technique of measuring the marginal relationship from an estimated function. (For details, read the relevant section).

• The condition for profit maximization is Marginal Cost is equal to Marginal revenue, i.e., $MC = MR$.

• *Constrained maximization* refers to the condition of maximizing a variable under certain given conditions fixing the limits for the independent variables.

REVIEW QUESTIONS

1. What is meant by 'variables'? What are economic variables? Distinguish between (*a*) dependent and independent variables, and (*b*) endogenous and exogenous variables.

2. What are the methods of showing relationship between a dependent and an independent variable. Illustrate by tabular and graphical methods.

3. What is meant by a 'function'? How is an economic function formulated?

4. Suppose the price function for commodity X is given as $Px = 100 – 0.2\ Qx$. Find the quantity sold that maximizes total revenue (TR).

(**Hint:** $TR = Q \times P$)

5. Suppose the demand function for commodity X is given as $Qx = 100 – 5\ Px$. Find the total sale that maximizes total revenue (TR).

(**Hint:** Find first the price function from the demand function; $TR = Q \times P$)

6. The total cost (TC) function of a company has been estimated as $TC = 100 + 5Q + 10Q^2$ where Q = quantity produced. Find (*a*) marginal cost when $Q = 10$; (*b*) quantity (Q) at which average cost is minimum; and (*c*) Q at which average cost (AC) equals marginal cost (MC).

7. A transport company plans to ferry tourists between Delhi and Simla. It has its cost function estimated as given below:

$TC = 100 + 50N + 4N^2$

where N denotes number of passengers per day.

Find (*a*) the average cost function; (*b*) number of passengers per day that minimizes the average cost.

8. Find the derivative of the following functions.

(*i*) $Y = f(X) = 100$

(*ii*) $Y = f(X) = aX^b$

(*iii*) $Y = f(X)/g(X)$

9. Find the second derivative of the following functions.

(*a*) $Y = 5 + 10X + 5X^2$ (*b*) $Y = 5X(2 + X^2)$

(*c*) $Y = aX^b + X$

10. Suppose TR and TC functions are given as follows.

(*a*) $TR = 500Q – 2Q^2$

(*b*) $TC = 750 + 50Q + 2Q^2$

Find (*i*) the profit function, and

(*ii*) value of Q that maximizes profit.

11. Suppose profit function of a multi-product firm producing two goods, X and Y is given as follows.

$X = 50X – X^2 – XY + 60Y – 2Y^2$

Find the profit maximizing combination of X and Y subject to $X + Y = 30$ by using Lagrangian multiplier method.

12. Suppose a firm has its TR and TC functions estimated as follows:

$TR = 300Q – 3Q^2$ and $TC = 500 + 50Q + 2Q^2$

Find (*a*) profit function of the firm, (*b*) the quantity of output (Q) that maximizes the firm's profit.

13. A company produces two goods, X and Y. The profit function (Π) of the company is given as follows:

$\Pi = 100X – 2X^2 – XY + 180Y – 4Y^2$

The company is under obligation to produce a minimum combined output of X and Y equal to 30 units. Find the output of X and Y subject to a total of 30 units that maximize total profit by using both substitution and Lagrangian multiplier methods.

14. Suppose that the manager of a firm is planning to meet an order of 1000 units of two products, X and Y. The manager's problem is to find the combination of two goods that minimizes its cost. He has the firm's cost function of two goods estimated as $C = 5X^2 + 20Y^2$. By using the Lagrangian multiplier method, find the quantity of X and quantity of Y, subject to $X + Y = 1000$, that minimize the cost of meeting the order.

15. The Jaico Jeans Co. has been manufacturing and selling a branded jeans "TOUGH" over the past several years. The company has always faced tough competition from the rival firms in the business. Jaico Jeans has, however, faced the competition successfully. The secret of its success

has been its ability to innovate new designs of branded jeans, some imitated and some original.

Nevertheless, it finds carrying on business this way and making profit to be a tough proposition because other firms follow suit quickly with new designs. The company has planned to enter the new millennium with a new and more attractive branded jeans. It has decided to give the brand name "SNOB" to its new model. This model is intended to cater to the demand of upper class consumers.

On a preliminary assessment of the cost and price conditions the Jaico Company finds that:

(*i*) The production cost of the new brand is expected to be much higher than that of TOUGH. Therefore, the cost function of TOUGH is not relevant for SNOB.

(*ii*) Since the cost of production is expected to be significantly higher, the price of SNOB has to be correspondingly higher than TOUGH. Therefore, the demand function for the new brand is bound to be a different one.

The Jaico Jeans Co. uses its in-house resources to find the cost and demand functions for its new brand SNOB. The research conducted by the production department estimates the cost function as

$TC = 1000 + 100\,Q + 2\,Q^2$

The market research department conducts a market survey. On the basis of information collected through the survey, it has estimated the demand function as given below.

$Q = 200 – 0.25\,P$

The Jaico Jeans Co. supplies this information to you and seeks your opinion on the following aspects of its decision-making.

(*a*) Given the cost and demand functions for the new product SNOB, will the introduction of the new product be profitable independent of TOUGH?

(*b*) If it is profitable, then what is the profitable range for the production of jeans? This information is required for chalking out an appropriate pricing strategy in face of the competition.

(*c*) What is the level of output that will maximize the profit from the production and sale of SNOB?

FURTHER READING

Allen, R.G.D, *Mathematical Analysis for Economists*, (Macmillan, London, 1956).

BAumol, W.J., *Economic Theory and Operations Analysis*, (Englewood Cliffe, N.J.), 4th Edn., 1977.

ChiAnG, AlphA, *Fundamental Methods of Mathematical Economics*, (McGraw-Hill, N.Y.), 3rd Edn., 1984.

DoWlinG, eDWARD T., *Mathematics for Economists,* (McGraw-Hill, N.Y.), 3rd Edn., 2000.

ThomAs, ChRisTopheR R. and S. ChARles mAuRiCe, *Managerial Economics: Concepts and Application (Tata McGraw-Hill Publishing Company, New Delhi, 8th Edn. 2003), Chs. 3 and 4.*

YAmAne, Taro, Mathematics for Economists (Prentice-Hall of India, New Delhi, 1968), Chs. 4, 5 and 6.

CHAPTER

5 Method of Estimating a Function: The Regression Technique

CHAPTER OBJECTIVES

The objective of this chapter is to explain the method of how a function is formulated and estimated by statistical technique of regression. By going through this chapter, you will learn:

- Why the need for estimating a function arises
- The method of formulating a function
- Method of estimating a function by regression technique
- Method of testing the significance of estimated parameters

5.1 INTRODUCTION

In the preceding chapters, we have used some functions, e.g., demand function, cost function, reverence function, etc. Recall demand function: $Dx = a – bPx$. This method of stating functional relationship between two or more economic variables and the results they produce are often mystifying to a section of students. A question is often asked: how are these functions specified and formulated, and how are the parameters ' *a*' and ' *b*' estimated? Our aim in this chapter is to answer this question keeping in view the interest of the students with an elementary understanding of statistics and to give them an idea as to how the relationship between two variables is specified and estimated. Specifically, we will explain a statistical technique called regression analysis.

Regression is a statistical technique used to quantify the relationship between economic variables. Regression, as a tool of analysis, was first used by Francis Galton[1] in 1886. Since then it has emerged as a powerful tool of estimating the nature and extent of the relationship between two or more related variables. Regression technique is used to measure

1. The term regression was first used by Francis Galton in his paper "Family Likeness in Stature" published in *Proceedings of Royal Society* (1886). Galton carried out a study to find out whether tall parents always have tall children and short parents always have short children. He found that though there was a tendency for tall parents to have tall children and short parents to have short children, there was a tendency for the average height of both tall and short child ren to reach the average height of the entire population. Galton's friend, Karl Pearson, a well known name in statistics, confirmed Galton's findings (For details, see Karl Pearson and A. Lee, *Biometrica,* Vol. 2, 1903). Pearson and Lee established that average height of sons of tall fathers tended to go down and that of the sons of short fathers tended to go up vis-a-vis the average height of the population.

the *stochastic relationship* [2]—approximate relationship—between the two or more variables having a cause-and-effect relationship between them. Regression technique measures an *approximate* value of the dependent variable in response to a change in the independent variable(s) of a function. The regression technique performs two basic functions:

(*i*) It estimates a function and quantifies the extent of relationship between the dependent and the independent variables from the observed data, or in other words, it estimates the coefficients of the independent variables; and

(*ii*) It provides a measure of the reliability of the estimated coefficients.

Regression is a versatile technique of estimating a variety of functions. It is used in all physical and social studies where a problem of specifying the relationship between two or more interrelated variables is involved.

A comprehensive treatment of this statistical technique is out of the purview of this book. Our aim in this chapter is simply to introduce elementary regression[3] technique to show how a function is formulated and how its parameters and co-efficients are estimated. An elementary knowledge of regression technique is very essential to understand the logic and basis of the functions used in economic analysis and managerial decisions. We begin the regression analysis with a simple example of managerial problem.

5.2 THE NEED FOR ESTIMATING A FUNCTION: AN EXAMPLE

Suppose a manager has been spending money year after year on advertisement to promote the sales of his firm's product. The annual sales figures, as presented in Table 5.1, show that sales have been increasing but not consistently. In some years, they increased at a high rate; in some years at a low rate; in some years they remained

Table 5.1 *Ad-expenditure and Sales Data*

constant; and in some years, sales declined in

Year

Ad-expenditure

Sales figure

spite of advertisement. The manager's problem

(mill. `)

('000 units)

is to find an answer to the following questions.

(*i*) How strong or weak is the relationship

1995

5

45

between advertisement expenditure (ad-

1996

8

50

expenditure, for short) and total sales?

1997

10

55

1998

12

58

(*ii*) What is the measure of the extent

1999

10

58

of this relationship, i.e., what is the

2000

15

72

quantitative response of sales to an

2001

18

70

increase in ad-expenditure?

2002

20

85

An answer to question (*i*) can be obtained

2003

21

78

by a careful scrutiny of advertisement and

2004

25

85

sales data given in Table 5.1. The data does

indicate that there is a positive relationship

between ad-expenditure and sales, despite some exceptions.

2. The word “stochastic” comes from the Greek word *stokhos* meaning “a bull’s eye”. Throwing a dart to hit the

‘bull’s eye’ on the dartboard is not always accurate. Most outcomes of throws are random and are called stochastic.

In modern usage, a *stochastic relationship* refers to statistical or a probable relationship in contrast to ‘exact

relationship’ between two variables. The exact relationship between two variabels, say *X* and *Y*, is expressed, for

example, as $Y = 5X$. Here, if the value of *X* is known, the exact value of *Y* is also known. In case of a stochastic or statistical relationship between any two variables, one can estimate *only an approximate, but not exact*, value

of *Y* for any given value of *X*.

3. For a comprehensive treatment of regression technique, the students are advised to refer to some standard texts on statistics.

A more clear and certain answer to question (*i*) can be found by plotting the sales

data against the ad-expenditure. This has been illustrated in Fig. 5.1. Axis *Y* measures

the sales and axis *X* measures ad-expenditure. The ad-expenditure and sales data plotted

in Fig. 5.1 present a scatter diagram. The plotted points show clearly that there is a

positive relationship between ad-expenditure and sales. In most years, increase in ad-

expenditure resulted in increased sales. This finding confirms the conclusion drawn from

the observation of sales and advertisement data.

A more precise answer to question (*i*) is provided by the straight line drawn through

the centre of the scatter points.

However, question (*ii*) is much more important from managerial decisions point of

view. For, the managers need to know the exact relationship between ad-expenditure and

sales for future planning. But, the scatter diagram does not answer question (*ii*), i.e., it

does not reveal extent of relationship between ad-expenditure and sales. For example, it

does not give the increase in sales due to, say, `1 million increase in ad-expenditure. This

question is answered by the regression technique. To answer this question, the regression

technique postulates a linear function. The function, when estimated with observed data,

provides the answer to the question. Therefore, before we use the regression technique

to find the answer to question (*ii*), let us look at its method of formulating a function.

Fig. 5.1 *Scatter Diagram of Ad-Expenditure and Sales*

5.3 METHOD OF FORMULATING A FUNCTION

There are two important steps in the formulation of a function: (*i*) formulation of a hypothesis, and (*ii*) converting the hypothesis into a mathematical function. Let us now look briefly at the two steps in the formulation of the function.

5.3.1 Formulation of Hypothesis

The first step in the formulation of a function is to formulate a hypothesis. A hypothesis is a postulate, an untested proposition, regarding the relationship between any two or more variables of the real world phenomena. A hypothesis is formulated on the basis of the observed relationship between two or more facts or events of real life.4 One can find innumerable phenomena which show a functional relationship between two (or more) events. For example, consider the following statements with *ceteris paribus* assumption.

(*i*) Demand for a product is a negative function of its price;

(*ii*) Supply of a product is a positive function of its price;

(*iii*) Consumption is a function of income, other factors remaining constant;

(*iv*) Per hectare production of wheat is a function of fertilizer used and irrigation;

(*v*) The number of visitors to a hill station is a function of temperature in the plains;

(*vi*) Crime is the function of corruption in administration;

(*vii*) Tax evasion is a function of tax rate;

(*viii*) Illiteracy is a function of poverty;

(*ix*) Petrol consumption is a function of distance travelled;

(*x*) Marks secured in an examination are a function of study hours;

(*xi*) Quality of students is a function of quality of teachers, and so on.

One can add an endless number of such statements. Such statements, even if they are conjectural, form the basis of formulating a hypothesis. For instance, in our example of ad-expenditure-and-sales relationship, a hypothesis can be formulated as

"Sales growth is a positive function of ad-expenditure".

The formulation of a hypothesis in the form of a verbal statement shows only a probability of the event and serves as a guide for future action. But it cannot predict the result of an action. For example, the hypothesis that "Sales growth is a positive function of ad-expenditure" suggests that if a firm spends money on advertisement, its sales will most probably increase. But it does not convey even the approximate increase in the sales for a given increase in ad-expenditure. To put it in a general form, a verbal hypothesis cannot predict the change in Y due to change in X. For this, a hypothesis needs to be converted into a mathematical equation or stated in the form of an estimable *function* or *equation*. The method of formulating functions is described below.

5.3.2 Transforming a Hypothesis into a Function

As mentioned above, a hypothesis stated verbally cannot be estimated in quantitative terms. The second step is, therefore, to transform the verbal hypothesis into the form of an estimable function. To formulate an estimable function, the relationship between the dependent and independent variables needs to be specified and stated in the form of an equation.

The form of equation—whether *linear* or *nonlinear*—can be determined *a priori* by plotting the data on two related variables and free-hand drawing of a line through the centre of the plotted points. For example, if we draw a line through the center of the plotted points in

4. Hypotheses are formulated also on the basis of (*a*) tested and untested beliefs of the people, and (*b*) the existing theories.

Fig. 5.1, we get a straight line *RL*. The line *RL* establishes that there is a constant positive relationship between ad-spending and sales and that the relationship between ad-expenditure and sales can be presented through a *straight-line*. In case of a straight-line-relationship between any two variables, say X and Y, the hypothesis regarding their relationship can be expressed in the form of a **linear function**. Going by this rule, the hypothesis that 'sales growth is a function of ad-expenditure' can be translated into a mathematical function as

$$Y = a + bX \quad \text{...(5.1)}$$

In Eq. (5.1), Y = sales (the dependent variable), X = ad-expenditure (the independent variable) and a and b are constants. The constant a is the intercept, it gives the quantity of sales without advertisement, i.e., when $X = 0$. Constant b is the coefficient of Y in relation to X. It gives the measure of increase in sales due to a certain increase in ad-spending.

Once a hypothesis has been converted into a mathematical function, as given in Eq.

(5.1), the task of the analyst is to find the values of constants *a* and *b*. The values of constants *a* and *b* can be estimated by two methods.

(*i*) an elementary or rudimentary method and

(*ii*) by using the mathematical method—the regression technique.

Let us first look here at the rudimentary method that can be used to obtain the values of *a* and *b* pending the regression method till the next section.

The Elementary Method The values of *a* and *b* can be estimated, at least crudely, by using a rudimentary method and Eq. (5.1) converted into a numerically specific function. Note that the line *RL* in Fig. 5.1 extended to *Y*-axis meets the axis at point *R*. It means that at zero ad-expenditure, sales will be only 40 thousand units. Thus, intercept *a* = 40,000. The value of *b* can be determined by measuring the *slope* of the line *RL*, i.e., the change in *Y* in response to per unit change in *X*. The slope of the line *RL* can be easily measured. It is given by the triangle *RMJ*. It can be seen in the figure that

$JM = \Delta Y = 58{,}000 - 40{,}000 = 18{,}000$ and $RM = \Delta X =$ `10 million

Thus, the slope of line *RL* between points *R* and *J* = D *Y*/D *X* = *b* = 1,800/10 million = 1,800 per million. By substitution, we can now rewrite Eq. (5.1) as

$$Y = 40{,}000 + 1{,}800\, X \quad ...(5.2)$$

Equation (5.2) can be interpreted as 'total sale without any ad-expenditure is 40,000 units and an ad-expenditure of `1 million increases sales by 1,800 units.

How Reliable is this Estimated Function? The test of the reliability of the function (5.2) is its power to predict the average value of *Y* as accurately as possible in response to a change in the value of *X*. The predicting power of function (5.2) depends on the accuracy of the estimated value of constants *a* and *b*. The accuracy of the values of *a* and *b* depends on how close is the line *RL* to the plotted points. The closer the line to the plotted points, the better the fit of the line. The manual drawing of a line cannot make sure that it best fits the observed data. Besides, different analysts may draw different lines which would produce different results. Therefore, the reliability of function (5.2) is doubtful. *It is very important to note here that the rudimentary method of formulating a function, as described above, indicates only the visualization of the function, not the formulation of the actual function. For example, Eq. (5.2) is only a crude approximation of the function, not the actual function. The estimation of the actual function has been described in the following section* by using the regression method .

5.4 ESTIMATING A LINEAR FUNCTION: THE REGRESSION METHOD

In this section, we describe the mathematical and the formal *method of derivation and estimation of a linear function* given as $Y = a + bX$. The mathematical method used to estimate a function, i.e., to measure the average value of parameters, *a* and *b*, from a given data, is called ***regression analysis or regression technique***. The regression technique can be applied to estimate both bivariate and multivariate functions. Given our limited objective here, we will confine ourselves to a simple *bivariate linear function*.

5.4.1 Estimating the Error Term

The first step in estimating the regression function is to estimate the error term. The 'error term' refers to the deviation of the plotted points from the straight line drawn through the centre of plotted points. With reference to our about example, suppose ad-expenditure and sales data of a corporation when plotted produce a scatter diagram as shown by points marked in Fig. 5.2. It can be seen in Fig. 5.2 that *total change in Y is not explained by a change in X*. The regression line can explain the total change in *Y* in response to a change in *X* only if all the ad-sales combination points fall on the regression line. Then there is no error term. But, as is evident from Fig. 5.2, all ad-sales combination points do not fall on the regression line. Some points are placed above and some points below the regression line. It means that actual sales data deviate from the linear regression line. This means that the slope of the regression line, does not explain the total change in *Y* in response to a change in *X*. The unexplained part of *Y* is called the *error term*, the *residual* or the *disturbance*.

Given a linear function as $Y = a + bX$, the purpose of regression technique is to find average values for *a* and *b* which make the values of observed pairs of *X* and *Y*, i.e., (X_1, Y_1), (X_2, Y_2), etc., as close to the regression line as possible. The line so fitted is called the *best-fit regression line*. This objective is achieved by minimizing the *error terms*, i.e., the deviation of observed value of *Y* from its estimated value. For example, let the observed value of *Y* in year *t* be denoted by Y_t and its estimated value by Y_c. The error term can then be defined as

$e_t = Y_t - Y_c$

...(5.3)

The concept of *error term* (*e*) can be explained with the help of the Fig. 5.2 which reproduces the scatter diagram of Fig. 5.1. Suppose that by some ingenious method, the best-fit regression line is estimated as given by the function $Y_c = a + bX$. The regression line drawn on the basis of this function gives Y_c, for a given value of *X*. For example, in Fig. 5.2, for *X* = `8 million, say, in year *n*, $Y_n = NM$ and $Y_c = PM$. Thus,

$e_n = Y_n - Y_c = NM - PM = -PN$

...(5.4)

Since, as given above, $Y_c = a + bX$, the error term can be defined more precisely as

$e_t = Y_t - (a + bX)$

...(5.5)

The error term, e_t, accounts for the following kinds of errors.

(*a*) *Sampling error*, the error arising out of inaccurate recording or specification of sales data;

(*b*) *Specification error*, the error that arises because other factors influencing sales could not be specified and included in the function; and

(*c*) *Measurement error*, the errors that arise due to computational errors in measuring, sampling, coding and writing (or feeding) data.

Y

100

95

$Y = 34.54 + 2.15\ X_t$

90

85

80

75

70

65

60

Sales (’000 units)

55

P

50

N

45

40

M

X

O

2

4 5 6

8

10

12

14 1516

18

20 2122

242526

Advertisement Expenditure (mill.)

`

Fig. 5.2 *Fitting a Regression Line*

The objective of the regression technique is to minimize the error term with a view to finding the regression equation that best fits the data. The method used to minimize the error terms is called ***ordinary least square method***. This method is described below.

5.4.2 Ordinary Least Square (OLS) Method

As mentioned above, regression technique minimizes the error term with a view to finding the regression equation that best fits the observed data. So the problem is how to minimize the error term. There are two methods of minimizing the error term.

One way to minimize the error could be to find the sum of the error terms. Note that in Fig. 5.2, the errors in some years are positive (as plotted points fall above the trend line) and negative in some years. Under this method, positive and negative errors tend to get cancelled out. It would mean that error terms do not exist or there are no deviations from the estimated line. However, as can be seen in Fig. 5.2, the positive and negative error terms may not cancel out. Therefore, the sum of error terms cannot be used as the measure of deviation of the observed data from the estimated one.

This problem is resolved by using the *square of the error term*. Squaring error terms eliminates the minus signs, making the problem easier to handle. This technique that regression analysis uses to minimize the error term is called the ***ordinary least square (OLS)*** *method*. It is the sum of the square of the error terms that regression techniques

seeks to minimize and find the values of intercepts and coefficient of, *a* and *b,* that produce a best fit line. The process of finding the values of *a* and *b* by *OLS* method is described below.

The OLS Method Suppose we have *N* pairs of observations (X_1, Y_1), (X_2, Y_2), ... (X_n, Y_n) and we want to fit a regression line given by equation $Y = a + bX$. In doing so, the purpose is to find such values for *a* and *b* that minimize the sum of the square of the error terms.

$$\sum_{t=1}^{n} e_t^2 = E$$

Let the sum of the square of the error terms be defined, using Eq. (5.5), as

$$E = \sum_{t=1}^{n} (Y - a - bX)^2 \qquad ...(5.6)$$

By differentiating Eq. (5.6) with respect to *a*, we get

$$\frac{\delta E}{\delta a} = -2\, S\, (Y - a - bX) \qquad ...(5.7)$$

and with respect to *b*, $\frac{\delta E}{\delta b} = -2\, S\, X\, (Y - a - bX)$ ...(5.8)

For *E* to be minimum, $\frac{\delta E}{\delta a}$ and $\frac{\delta E}{\delta b}$ both must be equal to zero. That is

$$\frac{\delta E}{\delta a} = -2\, S\, (Y - a - bX) = 0 \qquad ...(5.9)$$

and

$$\frac{\delta E}{\delta b} = -2\, S\, X\, (Y - a - bX) = 0 \qquad ...(5.10)$$

Note that constant factor – 2 is common to both the equations. This factor can, therefore, be dropped without affecting the result. Thus, Eq. (5.9) and Eq. (5.10) can be written, respectively, as

$$S(Y - a - bX) = 0 \qquad ...(5.11)$$

and

$$X(Y - a - bX) = 0$$

...(5.12)

Equations (5.11) and (5.12) can be extended and written,5 respectively, as

$$\Sigma Y - Na - b\Sigma X = 0 \quad ...(5.13)$$

and

$$\Sigma XY - a\Sigma X - b\Sigma X^2 = 0 \quad ...(5.14)$$

Equations (5.13) and (5.14) can be rewritten, respectively, as

$$\Sigma Y = Na + b\Sigma X \quad ...(5.15)$$

and

$$\Sigma XY = a\Sigma X + b\Sigma X^2 \quad ...(5.16)$$

Equations (5.15) and (5.16) are *normal equations* which can be solved for determining the values of constants a and b. By solving these equations, we get

5. When we expand Eq. (5.11), we get $\Sigma Y - \Sigma a - \Sigma bX$. But $\Sigma a = a + a + a + ... n$ times $= Na$. Similarly, $\Sigma bX = bX + bX + ... + bX$ times $= b\Sigma X$. By substituting these terms in Eq. (5.11), we get LHS of Eq. (5.13).

$$a = \frac{(\Sigma X^2)(\Sigma Y) - (\Sigma X)(\Sigma XY)}{N\Sigma X^2 - (\Sigma X)^2} \quad ...(5.17)$$

$$b = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2} \quad ...(5.18)$$

Once the numerical values of a and b are determined, by substituting the values of a and b in Eq. (5.1) given as $Y = a + bX$, we get the regression equation. Let us now find the numerical values of a and b from the data given in Table 5.1.

Application To find the values of constants a and b and the regression equation for our ad-expenditure and sales problem, what we need to do is to find the values of N, ΣX, ΣY and ΣXY and substitute these values in Eq. (5.17) to find the value of a and in Eq. (5.18) to find the value of b. The values of N, ΣX and ΣXY are computed in Table 5.2

Table 5.2 *Regression of Sales on Advertisement Expenditure*

Year	*Advertisement Expenditure (X)*	*Sales*	*X^2*	*XY*

(Y)

2001

5

45

25

225

2002

8

50

64

400

2003

10

55

100

550

2004

12

58

144

696

2005

10

58

100

580

2006

15

72

225

1080

2007

18

70

324

1260

2008

20

85

400

1700

2009

21

78

441

1638

2010

25

85

625

2125

$N = 10$

$\Sigma X = 144$

$\Sigma Y = 656$

$\Sigma X^2 = 2448$ $\Sigma XY = 10254$

By substituting the values of N, S X, S Y, S X^2 and S XY in Eq. (5.17) and Eq. (5.18),

we get the values of the constants a and b, respectively, as follows.

$1605888 - 1476576$

a = (2448) (656) − (144) (10254) =

2

10(2448) − (144)

24480 − 20736

129312

a =

= 34.54

3744

102540 − 94464

and

b = 10(10254) − (144) (656) =

2

10(2448) − (144)

24480 − 20736

8076

b =

= 2.15

3744

The constant a = 34.54 is the intercept which give the total expected sales at zero level of ad-expenditure. Constant b = 2.15 gives the coefficient—the slope of the regression

line. It means that an increase of `1 million in ad-expenditure will bring an increase of 2.15 thousand units in sales.

Now that we have got the numerical values of intercept as a = 34.54 and coefficient as b = 2.15, we can write the estimated regression equation of the best fit by substituting these values in hypothetical function $Y = a + bX$. Thus,

Y = 34.54 + 2.15 X

...(5.19)

Equation (5.19) is the estimated regression equation. By substituting the value for X from the observed data (given in Table 5.1), we can obtain the regression line as shown in Fig. 5.2.

Given the data in Table 5.1, the regression equation (5.19) serves as an important and fairly reliable tool for predicting the approximate sales response to a given amount of advertisement expenditure. For example, if managers decide to spend `30 million in year 2011, they can predict the approximate sale in that year, assuming the response of the buyers is instant, as follows:

Sales2005 = 34.54 + 2.15 (30) = 34.53 + 64.50

= 99 thousand (approximately)

This means that if managers decide to spend `30 million on advertisement in year 2011, they can expect a total sale of approximately 99 thousand units in that year.

It is, however, very ***important*** for the users of regression technique to bear in mind that this technique shows only a probable tendency, i.e., it predicts only a probable result, not the exact result, assuming that what happened in the past years will hold in the subsequent years also. Since this technique does not take into account the effect of predictable and unpredictable events that might affect the predicted result, the regression technique remains only a mechanical device.

5.4.3 Test of Significance of Estimated Parameters6

We have estimated the parameters a and b in the regression equation expressed as

$Y = a + bX$

and, we have also illustrated the use of the estimated regression to estimate the value of Y (sales) for a given amount of advertisement expenditure (X). The question that now arises is 'how reliable is the estimated value of coefficient b or how well does the estimated regression line fit to the observed data?' For example, as worked out above, b = 2.15. It implies that an increase of `1 million in advertisement expenditure will cause an increase in sales of approximately 2.15 thousand units. How reliable is this conclusion? The technique that is used to answer this question is called ***test of statistical significance.***

The process of testing reliability or statistical significance of an estimated value begins with making a 'null hypothesis'. The 'null hypothesis' with reference to our example can be stated as b = 0. It means assuming that there is no relationship between Y and X. The problem is now to accept or reject the hypothesis. If null hypothesis is accepted, it means that there is no relationship between Y and X or, in other words, the variation in

Y (sales) *is not* explained by the variation in X. On the contrary, if the null hypothesis is rejected, it means that estimated b ¹ 0 and that $b > 0$ significantly.

6. Our objective in this section is to indicate that estimating a regression equation is not sufficient. Testing the reliability of the estimate is equally important. Where multivariate regression analysis is involved, there are several other tests, viz., *standard deviation, F ratio test and Durbin-Watson statistics*. We give here only a simple test of statistical significance.

The next step is, therefore, to test the null hypothesis. In fact, the task is to find the probability of rejecting the null hypothesis. The probability of rejecting a hypothesis is known as finding the level of *significance*. There is a statistical rule in this regard. The rule is that if the level of significance is 10 per cent or less, then the hypothesis is rejected. It implies that if the level of significance is 10 per cent or higher, then the estimated coefficient b is *statistically significant*. That is, if estimated coefficient b is *statistically significant* at 10 per cent level of significance, then it is concluded that X (ad-expenditure) is a significant determinant of Y (sales).

How is the level of significance determined? The level of significance is determined on the basis of the *standard error* and *t-ratio* to t statistic. We will first describe the precise method of calculating the standard error and the t-ratio.

The 'standard error' is the standard deviation of the estimated value from the sample values. Therefore, to test the hypothesis that there is a statistically significant relationship between Y (sales) and X (ad-expenditure), we need to calculate *standard error of coefficient b*, generally denoted by s_b. The formula for estimating s_b is given below.

$$s_b = \sqrt{\frac{\Sigma (Y_t - Y_{et})^2}{(N-k)\,\Sigma (X_t - \bar{X})^2}} = \sqrt{\frac{\Sigma e_t^2}{(N-k)\,\Sigma (X_t - \bar{X})^2}} \quad \ldots(5.20)$$

where X_t and Y_t are the actual sample values for year t, Y_{et} is the estimated value of Y in year t, $\bar{X}$ is the mean value of X, N is the number of observations and $e_t = Y_t - Y_{et}$ is the error term, and k is the number of estimated coefficients (2 in the case of a bivariate regression equation, a and b). In fact, $N - k$ is the *degree of freedom*. In our example, the degree of freedom equals $N - k = 10 - 2 = 8$.

We may now calculate the value of s_b as follows:

$$s_b = \sqrt{\frac{\Sigma e_t^2}{(N-k)\,\Sigma (X_t - \bar{X})^2}}$$

The process of computing s_b is illustrated in Table 5.3.

Table 5.3 *Calculation of Standard Error of Coefficient*

Year	*X*

2

t

Yt

Yte

Yt – Yte = et

et

(Xt – X)2

2001

5

45

45.29

– 0.29

0.09

88.36

2002

8

50

51.74

– 1.74

3.03

40.96

2003

10

55

56.04

– 1.04

1.09

19.36

2004

12

58

60.34

– 2.34

5.48

5.76

2005

10

58

56.04

1.96

3.85

19.36

2006

15

72

66.79

5.21

27.15

0.36

2007

18

70

73.24

– 3.24

10.50

12.96

2008

20

85

77.54

7.46

55.56

31.36

2009

21

78

79.69

– 1.69

2.68

43.56

2010

25

85

88.29

– 3.29

10.83

112.36

N =10

S *X*

2

t = 144

S *Yt* = 656

S *et* =

S(*Xt* – *X*)2 =

X = 14.4

Y = 65.6

120.26

374.40

Given the numerical values of these variables as estimated in Table 5.3, the numerical value of *sb* can be calculated as follows.

sb =

120.26

= 0.04 = 0.20

(10 − 2) (374.40)

Having estimated the value of *sb*, we need then to estimate the *t-ratio*. The *t*-ratio (*t*) is defined as *b/sb*.

b

2.15

Thus,

t =

=

= 10.75

s

b

0.20

Now that we have obtained the values of two tests—standard error and *t*-ratio—we use them finally to test the null hypothesis, that is, there is no relationship between *Y* (sales) and *X* (ad-expenditure). To test the hypothesis, we need to perform statistical *t* test, i.e., compare the computed *t*-ratio (10.75) with the critical *t* value with different degrees of freedom. Recall that *degrees of freedom* in our example equals *N* – *k* = 10 – 2 = 8. The critical *t* values for different degrees of freedom are given in the Appendix at the end of the chapter. The *t* test is usually performed at 5 per cent level of confidence. Referring to *t* distribution table in the Appendix, we locate the number 8 under the degrees of freedom. When we link 8 with 5 per cent level of confidence, under the column 0.05, we get critical *t* value as 2.306 for the so called ‘two-tailed test’. The value of *t* that we have calculated in our example is 10.75.

This value of *t* (i.e., 10.75) far exceeds the critical *t* value (i.e., 2.306) at the 5 per cent level of significance. Therefore, the null hypothesis that 'there is no relationship between *Y* (sales) and *X* (ad-expenditure)' is rejected. The rejection of null hypothesis at 5 per cent level of significance means that *there is a statistically significant relationship between Y (sales) and X (ad-expenditure)*. More precisely, we arrive at the conclusion that *we are 95 per cent confident that there is a statistically significant relationship between Y (sales) and X (ad-expenditure).* An implication of this conclusion is that the probability of it going wrong is less than 5 per cent.

5.4.4 Test of Goodness of Fit: The Coefficient of Determination

Apart from testing statistical significance of the relationship between *X* (ad-expenditure) and *Y* (sales), another test is performed to test the overall explanatory power of the estimated regression equation. This test is performed by calculating the *coefficient of determination*. The coefficient of determination, denoted usually by the symbol R^2, gives the measure of the overall strength of the association between the dependent variable (*Y*) and the independent variable (*X*). The coefficient of determination (R^2) is defined as the proportion of the total variation in the dependent variable *Y* (about its mean), explained by the variations in the independent variable or what is also called the **explanatory variable**, *X*. Given the definition, the coefficient of determination (R^2) is measured as follows:

$$R^2 = \frac{\text{Explained variation in } Y}{\text{Total variation in } Y} \qquad \ldots(5.21)$$

The *explained variation* is the sum of the squares of the deviation of the measured value of *Y* in each year from the mean of *Y*, (i.e., $\bar{Y}$). That is,

$$\text{Explained variation in } Y = \sum_{t=1}^{n} (Y_t^e - \bar{Y})^2 \qquad \ldots(5.22)$$

The *total variation* in *Y* equals the sum of the squares of the deviation of each observed value of *Y* from the mean of observed *Y*, i.e., $\bar{Y}$. Thus,

$$\text{Total variation in } Y = \sum_{t=1}^{n} (Y_t - \bar{Y})^2 \qquad \ldots(5.23)$$

Once the 'explained variation' and 'total variation' in *Y* are obtained, the coefficient of determination (R^2), as given in Eq. (5.21), can be redefined in terms of Eqs. (5.22) and (5.23) as

$$R^2 = \frac{\sum (Y_t^e - \bar{Y})^2}{\sum (Y_t - \bar{Y})^2} \qquad \ldots(5.24)$$

From explanation point of view, total variation in *Y* is constituted of two parts: (*i*) *explained variation*, and (*ii*) *unexplained variation. Explained variation* has already been defined in Eq. (5.22). The *unexplained variation* in *Y* equals the sum of square of the difference between the observed value of *Y* in each year and the estimated value of *Y* for each year. That is,

$$\text{Unexplained variaton} = \sum_{t=1}^{n} (Y_t - Y_{te})^2 \quad \ldots(5.25)$$

Thus, the total variation in Y can be redefined in terms of Eqs. (5.22) and (5.25) as

Total variation = Explained variation + Unexplained variation

$$= \sum (Y_{te} - \bar{Y})^2 + \Sigma(Y_t - Y_{te})^2 \quad \ldots(5.26)$$

Although computer programmes provide R^2, we will illustrate here the process of its calculation. The process of calculating coefficient of determination (R^2), as defined in Eq. (5.24), is illustrated in Table 5.4.

Since we have already computed the elements of the coefficient of determination, we can calculate R^2. Column 4 of Table 5.4 shows the total variation and column 7 shows the explained variation. Given the values, we get

$$R^2 = \frac{\sum (Y_{te} - \bar{Y})^2}{\sum (Y_t - \bar{Y})^2} = \frac{1730.53}{1862.40} = 0.93 \quad \ldots(5.27)$$

Table 5.4 *Calculation of Coefficient of Determination*

Year	Y_t	$Y_t - \bar{Y}$	$(Y_t - \bar{Y})^2$	Y_{te}	$Y_{te} - \bar{Y}$	$(Y_{te} - \bar{Y})^2$	$Y_t - Y_{te}$	$(Y_t - Y_{te})^2$
2001	45	– 20.60	424.36	45.29	– 20.31	412.50	– 0.29	0.09
2002	50	– 15.60	243.36	51.74	– 13.86	191.83	– 1.74	3.03
2003	55	– 10.60	112.36	56.04	– 9.56	91.40	– 1.04	

1.09
(Contd...)

2004
58
– 7.60
57.76
60.34
– 5.26
27.67 – 2.34
5.48
2005
58
– 7.60
57.76
56.04
– 9.56
91.40
1.96
3.85
2006
72
6.40
40.96
66.79
1.19
1.42
5.21
27.15
2007
70
4.40
19.36
73.24
7.64
58.37 – 3.24
10.50
2008
85
19.40
376.36
77.54
11.94
142.57
7.76
60.22
2009
78
12.40
153.76
79.69
14.09
198.53 – 1.69
2.86
2010
85
19.40
376.36

88.29

22.69

514.84 – 3.29

10.83

$N = 10$ $\Sigma Yt = 656$

$\Sigma(Yt - \bar{Y})^2$

$\Sigma(Yte - \bar{Y})^2$

$\Sigma(Yt - Ye)^2$

$\bar{Y} = 65.6$

= 1862.40

= 1730.53

= 125.10

As shown in Eq. (5.27), $R^2 = 0.93$. It means that 93 per cent of the total variation in the dependent variable *Y* (sales) is explained by the independent variable *X*, i.e., the advertisement expenditure. It means that the regression Eq. (5.19) has a very high explanatory power and that the regression line is a 'good fit'.

An important aspect of the coefficient of determination (R^2) is that its square root gives the ***coefficient of correlation,*** denoted by *r*. That is,

Coefficient of correlatio = r =

2

R

...(5.28)

The coefficient of correlation measures the degree of association between the dependent and the independent variables. It is also important to note here that *while regression equation assumes that the variation in the dependent variable (Y) is caused by the variation in the independent variable (X), the coefficient of correlation gives the measure of only the degree of association or covariance between the dependent and the independent variables.*

We may apply this formula to our example to find the correlation coefficient between sales and ad-expenditure. The formula is given as:

$r = 0.93 = 0.96$

The value of $r = 0.96$ (which is close to 1) indicates that there is a very strong association or correlation between sales and ad-expenditure.

Limitations Although the regression technique is a versatile tool widely used in business and economic analyses, it has certain *limitations*. The analyst must bear in mind that the regression analysis provides only a method of measuring the regression coefficients and test of their reliability or the goodness of fit. It does not provide a theoretical underpinning of the relationship between the dependent and the independent variables. Nor does it prove the existence of a theoretical relationship between the variables. The cause-and-effect relationship must be provided by an acknowledged theory or the analyst must provide a theoretical underpinning thereto. Otherwise, regression coefficients just remain numbers without any policy implications. The users of regression analysis must, therefore, ensure that there exists a theoretical cause-and-effect relationship between the dependent and the independent variables which they are using.

5.5 MULTIVARIATE REGRESSION

We have so far explained the technique of estimating a bivariate regression equation. In this section, we extend the bivariate regression analysis to *multivariate regression analysis* or what is also called *multiple regression*. When a dependent variable (*Y*) is the

function of more than one independent or explanatory variables, it is called *multivariate regression*. In fact, the need for estimating a multivariate equation is much more common in economic analysis and managerial decision-making. For instance, recall our bivariate regression analysis of sales (*Y*) and ad-expenditure (*X*). Here it is assumed that sales are a function of *only* ad-expenditure. It is implicitly assumed that no other factors influence sales or that all *other factors* influencing sales remain *constant*. In reality, however, total sales or the total demand for the product of a firm depends on many other factors, viz., competitive advertisement by the rival firms, price of the substitutes, income of the relevant segment of the consumers, quality control, etc. In that case, a bivariate regression analysis of sales vs. ad-expenditure may lead to misleading conclusions. Therefore, where total sales are found to be a function of other independent variables also, the business analyst is required to include all the relevant and quantifiable independent variables in the sales (*Y*) function. Suppose sales depend on (*a*) the firm's own ad-expenditure, (*b*) ad-expenditure by rival firms, and (*c*) relative prices of the substitute products. In that case, the sales-function (*Y*) can be written as

$$Y = a + b_1 X_1 + b_2 X_2 + b_3 X_3 \qquad \text{...(5.29)}$$

where X_1 = firm's own ad-expenditure, X_2 = ad-expenditure by rival firms, X_3 = relative prices of the substitutes, and a, b_1, b_2 and b_3 are the coefficients.

The general form of a multivariate linear regression with *n* independent variables is written as

$$Y = a + b_1 X_1 + b_2 X_2 + \ldots + b_n X_n + e \qquad \text{...(5.30)}$$

where *e* is the error term.

The method of estimating a multiple regression is the same as that of estimating a bivariate or simple regression. In the case of multivariate regression, it is important that *the number of observations must be greater than the number of independent or explanatory variables to allow a reasonable degree of freedom*. However, estimating a multivariate regression manually is much more complex and time consuming. Therefore, the parameters of multivariate regression are invariably estimated with the aid of computer programmes.

5.6 SOME ADDITIONAL TESTS IN MULTIVARIATE REGRESSION ESTIMATES

Although the process of estimating a multiple regression is in principle the same as that of estimating a simple regression, some problems arise due to multiplicity of independent variables, which make the adjusted R^2 questionable. Therefore, the analyst is required to make some additional adjustments and tests in respect of multivariate regression. The major problems and their adjustment methods are discussed here briefly.

1. ***The Coefficient of Multiple Determination and Adjusted R2*** In a simple regression, the coefficient of determination (R^2) gives the measure of the proportion of the total variation in the dependent variable explained by the variation in the independent variables. It is, therefore, quite logical that the value of R^2 increases even if additional independent variables have no theoretical relationship with the dependent variable. For example, in our example of sales and ad-expenditure, $R^2 = 0.93$. This means that the variation in the independent variable (ad-expenditure) explains 93 per cent of the variation in the sales. If we add one more independent variable to the regression equation, the explanatory power

of the two variables will increase and the value of R^2 will increase, even if the additional independent variable is theoretically irrelevant. For example, if we add 'increasing crime rate in the city' into the simple regression equation, R^2 will increase. But that should not mean that increasing crime rate increases the sale of a normal consumer good.

In order to avoid this kind of problem in multiple regression, an adjustment is made in the coefficient of determination (R^2). The adjusted R^2 is called the *coefficient of multiple determination* and it is refrred to as $\bar{R}^2$. The $\bar{R}^2$ is obtained by the formula given below.

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-m} \qquad \text{...(5.31)}$$

where *n* = number of observations, and *m* = number of variables in the regression equation.

The adjusted coefficient of determination ($\bar{R}^2$) corrects the unadjusted R^2 for the decrease in the *degrees of freedom* due to an increase in the number of independent variables. Therefore, $\bar{R}^2$ gives a better measure of the proportion of the total variation in the dependent variable explained by the variation in the independent variables.

2. ***Problem of Multicollinearity*** One of the basic assumptions of multivariate regression analysis is that the *independent or explanatory variables are not interdependent* or that *the explanatory variables are independent of one another* or that *no independent variable is determined by any other independent variable.* However, it is quite likely that there does exist interdependence between the explanatory variables. For example, in our example of multiple regression, expenditure on advertisement by the rival firms may be interdependent. That is, when one firm increases its advertisement expenditure, the rival firms also increase their advertisement expenditure. This may lead to a competitive increase in advertisement expenditure. As a result, advertisement by rival firms becomes interdependent. This kind of interrelationship between the independent variables is called ***multicollinearity***.

When there is multicollinearity, it is very difficult to separate the effect of change in interrelated independent variables on the dependent variables. And, if effects are not separated, then the regression analysis fails to reveal the true relationship between the dependent and the independent variables. The major problem with collinearity is that it makes the *standard errors* of the estimated coefficients (b_1, b_2 ... b_n) of the correlated variables unduly large and this makes them statistically insignificant. Statistically insignificant coefficients have no theoretical value, nor are they of any use in policy formulation. It is, therefore, essential that the analyst checks whether multicollinearity exists or not.

One simple rule of checking the existence of multicollinearity is to check whether there exists a significant correlation between any pair of independent variables. The general method is to prepare a *correlation coefficient matrix* and check whether there exists a high correlation between any pair of independent variables. If a high correlation—the correlation coefficient being 0.9 or so—is detected, *one simple remedy to the problem is to drop one or more of the correlated variables.* Another way is to change the form of the regression equation.

3. ***Problem of Heteroscedasticity*** Heteroscedasticity exists when there is a *systematic relationship* between the error terms and the magnitude of one or more of the independent variables. The regression analysis assumes *homoscedasticity*, i.e., it assumes that error terms or the deviation from the line of best fit occur randomly with respect to the

magnitude of the independent variables. But, when error terms are systematically related to the magnitude of one or more of the independent variables, there exists *heteroscedasticity*. The existence of heteroscedasticity causes the coefficient of multiple determination to overstate the explanatory power of the regression equation and give misleading results.7

A simple way to test whether heteroscedasticity exists is to plot the error terms the values of residuals against the values of the independent variables. A visual inspection reveals whether there is heteroscedasticity. This problem can be solved by (*a*) changing the form of the regression equation, (*b*) re-specifying the independent variables having a systematic relationship with the error terms, and (*c*) using a weighted least square regression technique.

4. ***Problem of Autocorrelation or Serial Correlation*** The problem of autocorrelation arises in time series data. The existence of autocorrelation is indicated by a sequential pattern in the error terms. In simple words, the existence of autocorrelation can be tested by plotting the error terms against the chronologically arranged time-series data. If the size of the error terms shows a definite pattern—it increases or decreases progressively or shows a cyclical pattern—then autocorrelation exists. A definite pattern of error terms implies that some other independent variable is also simultaneously influencing the dependent variable. In addition, there is a more profound test available to test the existence of autocorrelation, called the **Durbin-Watson statistics**8 or in its abbreviated form, D-W statistics. The formula for estimating **D-W statistics** is given below.

n

2

$\sum(tR - tR\ 1-)$

t 1

$D\text{-}W$

=

=

…(5.32)

n

2

∑ tR

t 1

=

where R_t = residual (error term) for period t and R_{t-1} = residual for the preceding period.

The Durbin-Watson statistic provides the test of existence of autocorrelation. A Durbin-Watson statistic around the value of 2 indicates the absence of autocorrelation. But if it is significantly greater or less than the value of 2, it shows the existence of autocorrelation. Incidentally, computerized results produce the measure of the Durbin-Watson statistic and also test whether there is autocorrelation.

The existence of autocorrelation leads to misleading results or unacceptable estimates of coefficients. The problem of autocorrelation can be removed by any of the following methods.

(*a*) By re-specifying the regression equation,

(*b*) By including in the regression equation the variable that can explain the autocorrelation,

(*c*) If error terms follow a cyclical or seasonal pattern, a "dummy" variable can be added, and

(*d*) Where the magnitude of cyclical variations in the residuals is very high, GNP is often added as an explanatory variable.

7. Interested students may refer to a standard text on econometrics.

8. The test of autocorrelation was developed by J. Durbin and G.S. Watson, in their paper "Testing for Serial Correlation in Least Squares Regression", Biometrika, 38(1951).

5. *Analysis of Variance* The analysis of variance is a technique to test the overall explanatory power of the regression equation. For this purpose, the analysis of variance uses *F statistic* or *F ratio.* The formula for computing the value of *F* statistic is give below.

$$F = \frac{\text{Explained variation}/(k-1)}{\text{Unexplained variation}/(n-k)} \quad ...(5.33)$$

where k = number of estimated parameters, n = number of observations.

The *F* statistic can also be calculated by the following ormula:

$$F = \frac{R^2 (k-1)}{(1-R^2)(n-k)} \quad ...(5.34)$$

The *F* statistic is used to test the hypothesis that the variations in the independent variables (*Xs*) explain a significant proportion of the variation in the dependent variable (*Y*). The *F* statistic so calculated is checked in *F* distribution table with respect to the degrees of freed on and the critical values. The computerized results also provide the analysis of variation.

REVIEW QUESTIONS

1. What is meant by hypothesis? How is a hypothesis formulated?
2. What is regression analysis? Describe the nature of business problems which can be solved by using the regression technique?
3. What is meant by cause-and-effect relationship between any two variables? Can the regression technique be used to measure the relationship between the level of education and crime rate? Give reasons for your answer.
4. What is meant by a bivariate linear function? What is the process of formulating an estimable bivariate function?
5. What is meant by an error term? How are the error terms estimated?
6. What is meant by Ordinary Least Square? How is it used in estimating a regression equation?
7. What is meant by test of significance? Why is it necessary to perform the test of significance

before accepting the regression coefficients?

8. What is meant by the coefficient of determination? What is the test of explanatory power of an estimated regression equation?

9. A sociologist conducts a survey of rural families to find out whether there is any relationship between the level of education (given in years of education) and monthly income. His survey produces the following data.

Years of Education and Levels of Income

Years of Education	Income (’000 `)
5	2
6	2
8	3
10	4
12	5
13	8
14	10
16	12

(*i*) Find from a scatter diagram whether there is any relationship between the level of education and monthly earning;

(*ii*) Specify the form of regression equation that be can be fitted to the above data;

(*iii*) Estimate the regression equation and perform the test of significance, and

(*iv*) Find the coefficient of determination.

10. A computer company faced with the problem of stagnation in its sales is planning to reduce its computer price to increase its sales revenue. Past experience of the company regarding price-cut and computer sales does not give a clear indication about the change in sales as a result of price-cut. To find a solution to its problem, the company provides its price and sales data to you as given below.

Year	*Price (` ’000)*	*Sales (’000 units)*
2003	50	60
2004	48	61
2005	45	70
2006	40	75
2007	38	82
2008	35	80
2009		

32

85

2010

30

86

2011

28

90

2012

25

95

Given the price sales data of the computer company, find

(*a*) demand function for computers,

(*b*) coefficient of determination and

(*c*) should the company reduce the price of its computers?

11. What is the use of a multivariate regression analysis? What are the problems that arise when multiple analysis is used?

12. What is meant by (*a*) multicollinearity and (*b*) autocorrelation? How is the regression result affected in the presence of these factors? How are these problems removed?

APPENDIX: Degree of Freedom

Probabilities

Degree

of

Freedom

.80

.60

.40

.20

.10 •

. .05

.02

.01

1

0.325

0.727

1.376

3.078

6.314

12.706

31.821

63.657

2

0.289

0.617

1.061

1.886

2.920

4.303

6.965

9.925

3

0.277

0.584

0.978
1.638
2.353
3.182
4.541
5.841
4
0.271
0.569
0.941
1.533
2.132
2.776
3.747
4.604
5
0.267
0.559
0,920
1.476
2.015
2.571
3.365
4.032
6
0.265
.0.553
0.906
1.440
1.943
2.447
3.143
3.707
7
0.263
0.549
0.896
1.415
1.895 .
2.365
2.998
3.499
8
0.262
0.546
0.889
1.397
1.860
2.306
2.896
3.355
9
0.261
0.543
0.883
1.383
1.833
2.262

2.821
3.250
10
0.260
0.542
0.879
1.372
1.812
2.228
2.764
3.169
11
0.260
0.540
0.876
1.363
1.796
2.201
2.718
3.106
12
0.259
0.539
0.873
1.356
1.782
2.179
2.681
3.055
13
0.259
0.538
0.870
1.350
.1.771
2.160
2.650
3.012
14
0.258
0.537
0.868
1.345
1.761
2.145
2.624
2.977
15
0.258
0.536
0.866
1.341
1.753
2.131
2.602
2.947
16
0.258

0.535
0.865
1.337
1.746
2.120
2.583
2.921
17
0.257
0.534
0.863
1.333
1.740
2.110
2.567
2.898
18
0.257
0.534
0.862
1.330
1.734
2.101
2.552
2.878
19
0.257
0.533
0.861
1.328
1.729
2.093
2.539
2.861
20
0.257
0.533
0.860
1.325
1.725
2.086
2.528
2.845
21
0.257
0.532
0.859
1.323
1.721
2.080
2.518
2.831
22
0.256
0.532
0.858
1.321
1.717

2.074
2.508
2.819
23
0.256
0.532
0.858
1.319
1.714
2.069
2.500
2.807
24
0.256
0.531
0.857
1.318
1.711
2.064
2.492
2.797
25
0.256
0.531
0.856
1.316
1.708
2.060
2.485
2.787
26
0.256
0.531
0.856
1.315
1.706
2.056
2.479
2.779
27
0.256
0.531
0.855
1.314
1.703
2.052
2.473
2.771
28
0.256
0.530
0.855
1.313
1.701
2.048
2.467
2.763
29

0.256
0.530
0.854
1.311
1.699
2.045
2.462
2.756
30
0.256
0.530
0.854
1.310
1.697
2.042
2.457
2.750
40
0.255
0.529
0.851
1.303
1.684
2.021
2.423
2.704
60
0.254
0.527
0.848
1.296
1.671
2.000
2.390
2.660
120
0.254
0.526
0.845
1.289
1.658
1.980
2.358
2.617
00
0.253
0.524
0.842
1.282
1.645
1.960
2.326
2.576

Part II

THEORY OF CONSUMER
DEMAND AND DEMAND
FORECASTING

CHAPTER

6 Analysis of Consumer Demand

CHAPTER OBJECTIVES

The objective of this chapter is to explain the theory of consumer behaviour and the theory of demand. By going through this chapter, you will have the knowledge of the following aspects of the consumer behaviour and theory of demand:

- Meaning of demand in economic sense
- The concept of cardinal and ordinal utility
- Method of analysing consumer behaviour by cardinal utility approach and how a consumer attains his/her equilibrium
- Meaning and properties of the indifference curve – a tool of analysis
- Method of analysing consumer behaviour by ordinal utility approach and conditions for consumer equilibrium
- Derivation of consumer demand through cardinal and ordinal utility approaches
- Revealed preference theory of consumer behaviour

6.1 INTRODUCTION

In Part I of the book, we have introduced Managerial Economics, certain basic economic concepts used in business decision-making, some important tool and techniques of economic analysis applied in business decision making, and optimization technique. This sets the ground for the study of managerial economics. In Part II of the book, we enter the central theme of the managerial economics beginning with the *analysis of consumer behaviour* and the *theory of demand*.

Consumer demand is the basis of all productive activities. Just as 'necessity is the mother of invention', *demand is the mother of production*. Increasing demand for a product offers high business prospects for it in future and decreasing demand for a product diminishes its business prospect. For example, increasing demand for computers, cars, mobile phones etc. in India has enlarged the business prospect for both domestic and foreign companies selling these goods. On the other hand, declining demand for black and white TV sets and manual typewriters is forcing their companies to switch over to modern substitutes or else go out of business. It is, therefore, essential for business managers to have a clear understanding of the following aspects of demand for their products:

(*i*) What is the basis of demand for a commodity?

(*ii*) What are the determinants of demand?

(*iii*) How do the buyers decide the quantity of a product to be purchased?

(*iv*) How do the buyers respond to change in pr oduct prices, their incomes and prices of the related goods?

(*v*) How can the total or market demand for a product be assessed and forecasted?

These questions are answered by the analysis of consumer behaviour. The study of consumer behaviour gives rise to the ***Theory of Consumer Demand***. In this and the following three chapters, we will discuss the theory of individual and market demand. This chapter deals with individual demand for a product. Let us begin with the meaning of demand.

6.2 MEANING OF DEMAND

In general sense of the term 'demand' means desire, need, want or requirement of a commodity. In economic sense, however, the term 'demand' means a ' *desire for a commodity backed by the ability and willingness to pay for it*.' It implies that unless a person has adequate purchasing power or resources and the willingness to spend his resources, his *desire* alone for a commodity would not be considered as his demand. For example, if a man wants to buy a car but he does not have sufficient money to pay for it, in economic sense, his want is not his demand for car. And, if a rich miserly person wants to buy a car but is not willing to pay for the car, his desire too is not his demand for a car. But if a man has sufficient money and is willing to pay the price of the car, his desire to buy a car is an *effective demand*.

The desires without adequate purchasing power and willingness to pay do not affect the market, nor do they generate production activity. A want with *three attributes— desire to buy, ability to pay and willingness to pay*—becomes effective demand. Only an effective demand figures in economic analysis and business decisions.

In regard to market demand, the term 'demand' for a commodity (i.e., quantity demanded) always has a reference to 'a price', 'a period of time' and 'a place'. Any statement regarding the demand for a commodity without reference to its price, time unit and place of demand is meaningless and is of no practical use. For instance, to say 'the demand for TV sets is 50,000' carries no meaning for a business decision, nor does it have any use in any kind of economic analysis. A meaningful statement regarding the demand for TV sets may be expressed as 'the annual demnad for TV sets in Delhi at a price of `25,000 per unit is `50,000 units.

6.3 BASIS OF CONSUMER DEMAND: THE UTILITY

Consumers demand a commodity because they derive or expect to derive *utility* from the consumption of that commodity. The expected utility from a commodity is the basis of demand for it. Though 'utility' is a term of common usage, it has a specific meaning and use in the analysis of consumer demand. We will, therefore, describe in this section the meaning of utility, the related concepts and the law associated with utility.

6.3.1 Meaning of Utility

The concept of utility can be looked upon from two angles—from the product angle and from the consumer's angle. From the product angle, *utility is the want-satisfying property*

of a commodity. From consumer's angle, *utility is the psychological feeling of satisfaction, pleasure, happiness or well-being, which a consumer derives from the consumption, possession or the use of a commodity.*

There is a subtle difference between the two concepts which must be borne in mind. The concept of a want-satisfying property of a commodity is 'absolute' in the sense that this property is ingrained in the commodity irrespective of whether one needs it or not. For example, a pen has its own utility irrespective of whether a person is literate or illiterate. Another important attribute of the 'absolute' concept of utility is that it is '*ethically neutral*' because a commodity may satisfy a frivolous or socially immoral need, e.g., alcohol, drugs, porn-CDs, etc.

On the other hand, from a consumer's point of view, utility is a post-consumption phenomenon as one derives satisfaction from a commodity only when one consumes or uses it. Utility in the sense of satisfaction is a 'subjective' or 'relative' concept because (*i*) a commodity need not be useful for all—cigarettes do not have any utility for non-smokers, and meat has no utility for strict vegetarians; (*ii*) utility of a commodity varies from person to person and from time to time; and (*iii*) a commodity need not have the same utility for the same consumer at different points of times, at different levels of consumption and for different moods of a consumer. In consumer analysis, only the 'subjective' concept of utility is used.

Having explained the concept of utility, we now turn to some quantitative concepts related to utility used in utility analysis, viz. *total utility* and *marginal utility*.

6.3.2 Total Utility

Assuming that utility is measurable and additive, total utility may be defined as the sum of the utility derived by a consumer from the various units of a good or service he consumes at a point or over a period of time. Suppose a consumer consumes four units of a commodity, *X*, at a time and derives utility from the successive units of consumption as u , u , u and u . His total utility (U) from commodity *X* can be then measured as follows.

1

2

3

4

x

$U = u + u + u + u$

x

1

2

3

4

If a consumer consumes a large number of commodities, say, *n* number of commodities, his total utility, TU , is the sum of the total utility derived from each commodity. For

n

instance, if the consumption goods are *X*, *Y* and *Z* and their total respective utilities are U , U and U , then

x

y

z

$TU = U + U + U$

n

x

y

z

6.3.3 Marginal Utility

Marginal utility is another very important concept used in economic analysis. Marginal utility can be defined as the utility derived from the marginal or one additional unit consumed. It may also be defined as the addition to the total utility resulting from the consumption of one additional unit. Marginal Utility (*MU*) thus refers to the change in the Total Utility (i.e., Δ *TU*) obtained from the consumption of an additional unit of a commodity, say *X*. It may be expressed as

$MU = \Delta\ TUx$

x

$\Delta\ Qx$

where *TU* = total utility, and Δ *Q* = change in quantity consumed by one unit.

x

x

Another way of expressing marginal utility (*MU*), when the number of units consumed is *n*, can be as follows.

MU of *n* th unit = *TU* – *TU*

n

n–1

6.4 CARDINAL AND ORDINAL CONCEPTS OF UTILITY

The cardinal and ordinal concepts of utility arise out of question whether 'utility is measureable'. Utility is a psychological phenomenon. It is a feeling of satisfaction, pleasure or happiness. Measurability of utility has, however, been a contentious issue. Classical economists, viz., Jeremy Bentham, Leon Walrus, Carl Menger, etc. and neo-classical economists, notably Alfred Marshall, believed that utility is cardinally or quantitatively measurable like height, weight, length, temperature and air pressure. This belief resulted in the *Cardinal Utility* concept. However, modern economists, most notably J.R. Hicks and R.G.D. Allen, hold the view that utility is not quantitatively measurable—it is not measurable in absolute terms. Utility can be expressed only ordinally or in terms of 'less than' or 'more than'. It is, therefore, possible to list the goods and services in order of their preferability or desirability. For example, suppose a person prefers chocolate to ice cream and ice cream to cold drink. He or she can express his/her preference as chocolate > ice cream > cold drink. This is known as the *ordinal* concept of utility. Let us now look into the origin of the two concepts of utility and their use in the analysis of demand.

6.4.1 Cardinal Utility

Some early psychological experiments on an individual's responses to various stimuli led neo-classical economists to believe that utility is measurable and cardinally quantifiable. According to neo-classical economists, utility can be measured in terms of money. That is, utility of a unit of a commdity for a person is equal to the amount of money he is willing to pay for it. This belief gave rise to the concept of cardinal utility. It implies that utility can be assigned a cardinal number like 1, 2, 3, etc. Neo-classical economists built up the theory of consumption on the assumption that utility is cardinally measurable. They coined and used a term 'util' meaning 'units of utility'. In their measure of utility, they assumed (*i*) that one 'util' equals one unit of money, and (*ii*) that utility of money remains constant. It has, however, been realized over time that *absolute* or cardinal measurement of utility is not possible. Difficulties in measuring utility have proved to be insurmountable. Neither economists nor scientists have succeeded in devising a technique or an instrument for measuring the feeling of satisfaction, i.e., utility. Nor could an appropriate measure of unit be devised. Numerous factors affect the state of consumer's mood, which are impossible to determine and quantify. *Utility is therefore not measurable in cardinal terms*. Nevertheless, cardinal utility concept continues to remain the basis of the analysis of consumer behaviour.

6.4.2 Ordinal Utility

The modern economists have discarded the concept of *cardinal utility* and have instead employed the concept of *ordinal utility* for analyzing consumer behaviour. The concept of *ordinal utility* is based on the fact that it may not be possible for consumers to express the utility of a commodity in absolute or quantitative terms, but it is always possible for a consumer to tell introspectively whether a commodity is more or less or equally useful

compared to another. For example, a consumer may not be able to say that ice cream gives 5 utils and chocolate gives 10 utils. But he or she can always specify whether chocolate gives more or less utility than ice cream. This assumption forms the basis of the ordinal theory of consumer behaviour.

While neo-classical economists maintained that cardinal measurement of utility is practically possible and is meaningful in consumer analysis, modern economists maintain that utility being a psychological phenomenon is inherently immeasurable, theoretically, conceptually as well as quantitatively. They also maintain that the concept of ordinal utility is a feasible concept and it meets the conceptual requirement of analyzing consumer behaviour in the absence of any cardinal measures of utility.

6.4.3 Two Approaches to the Consumer Demand Analysis

Based on cardinal and ordinal concepts of utility, there are two approaches to the analysis of consumer behaviour.

(*i*) ***Cardinal Utility Approach***, attributed to Alfred Marshall and his followers, is also called the Neo-classical Approach.

(*ii*) ***Ordinal Utility Approach***, pioneered by J.R. Hicks, a Nobel laureate, and R.G.D. Allen, is also called the Indifference Curve Analysis.

The two approaches are not in conflict with one another. In fact, they represent two levels of sophistication in the analysis of consumer behaviour. Both the approaches are important for assessing and analyzing consumer demand for a commodity—be it for thoretical purpose or for business decision-making, depending on the level of sophistication required.

It is **important** to note in this regard that in spite of tremendous developments in consumption theory based on ordinal utility, the classical demand theory based on cardinal utility has retained its appeal and applicability to the analysis of consumer behaviour. Besides, the study of classical demand theory serves as a foundation for understanding the advanced theories of consumer behaviour. The study of classical theory of demand is of particular importance and contributes a great deal in managerial decisions.

We commence our analysis of consumer demand based on the *cardinal utility approach*. The *ordinal utility* approach to analyse consumer demand will be discussed in the next section.

6.5 ANALYSIS OF CONSUMER BEHAVIOUR: CARDINAL UTILITY APPROACH

The central theme of the consumption theory – be it based on ordinal utility or cardinal utility approach – is the *utility maximizing behaviour* of the consumer. The fundamental postulate of the consumption theory is that all the consumers—individuals and households—aim at *utility maximization* and all their decisions and actions as consumers are directed towards utility maximization. The specific questions that the consumption theory seeks to answer are:

(*i*) how does a consumer decide the optimum quantity of a commodity that he or she chooses to consume, i.e., how does a consumer attain his/her equilibrium in respect to each commodity?

(*ii*) how does he or she allocate his/her disposable income between various commodities of consumption so that his/her total utility is maximized?

The theory of consumer behaviour based on the cardinal utility approach seeks to answer the above questions on the basis of the following *assumptions*.

Assumptions

(*i*) **Maximization of satisfaction.** Every rational consumer intends to maximize his/her satisfaction from his/her given money income.

(*ii*) **Rationality.** It is assumed that the consumer is a rational being in the sense that he or she satisfies his/her wants in the order of their preference. That is, he or she buys that commodity first which is expected to yield the highest utility and the last which is likely to the least utility.

(*iii*) **Limited money income.** The consumer has a limited money income to spend on the goods and services he or she chooses to consume. Limitedness of income, along with utility maximization objective makes the choice between goods inevitable.

(*iv*) **Utility is cardinally measurable.** The cardinalists have assumed that utility is cardinally measurable and that utility of one unit of a commodity equals the units of money a consumer is prepared to pay for it, i.e., 1 util = 1 unit of money.

(*v*) **Diminishing marginal utility.** The consumption of a commodity is subject to the law of diminishing marginal utility. That is, it is assumed that the utility gained from the successive units of a commodity consumed decreases as a person consumes them. This is an axiom of the theory of consumer behaviour.

(*vi*) **Constant marginal utility of money.** The cardinal utility approach assumes that marginal utility of money remains constant whatever the level of a consumer's income.

(*vii*) **Utility is additive.** Cardinalities utility approach assumes the cardinally measurable utility is additive. It implies that utility derived from each unity of commodity and from different commodities can be added numerically. For example, suppose a consumer consumes three units of a commodity and *MU* derived from each unit is given as U_1, U_2, and U_3. In that case,

Total Utility *(TU)* = $U_1 + U_2 + U_3$.

6.5.1 Law of Diminishing Marginal Utility

The law of diminishing marginal utility is one of the fundamental laws of economics pertaining to the analysis of consumer behaviour. This law states that *as the quantity consumed of a commodity goes on increasing, the utility derived from each successive unit consumed goes on decreasing, consumption of all other commodities remaining constant.* In simple words, when a person continues to consume more and more units of a commodity at a point of time, e.g., chocolates, the utility that he derives from each successive chocolate goes on diminishing. The law of diminishing marginal utility is illustrated below numerically.

To illustrate the law of diminishing marginal utility, let us assume that a consumer consumes only one commodity *X*, and that utility is measurable in quantitative terms. Let us also suppose that total and marginal utility schedules of commondity *X* are given as in Table 6.1. The law of diminishing marginal utility is illustrated numerically in Table 6.1 and graphically in Fig. 6.1.

Table 6.1 *Total and Marginal Utility Schedules for Commodity X.*

Units of commodity X	*Total utility* (TU_x)	*Marginal utility* (MU_x)
1	30	30
2	50	20
3	60	10

4

65

5

5

60

– 5

6

45

– 15

As shown in Table 6.1, with the increase in the number of units of commodity *X*
consumed per unit of time, *TU* , increases but at a diminishing rate. The diminishing
x
MU is shown in the last column of Table 6.1. Figure 6.1 illustrates graphically the law
x
of diminishing *MU* . The rate of increase in *TU* as a result of increase in the number of
x
x
units consumed is shown by the *MU* curve in Figure 6.1. The downward sloping *MU*
x
x
curve shows that marginal utility goes on decreasing as consumption increases. At 4 units
consumed, the *TU* reaches its maximum level, the point of saturation marked by point
x
M. Beyond this point, *MU* becomes negative and *TU* begins to decline. The downward
x
x
sloping *MU* curve illustrates the law of diminishing marginal utility.
x

Why Does MU Decrease? The utility gained from a unit of a commodity depends on the intensity of the desire for it. When a person consumes successive units of a commodity, his need is satisfied by degrees in the process of consumption of the commodity and the intensity of his need goes on decreasing. Therefore, the utility obtained from each successive unit goes on decreasing.

Assumptions The law of diminishing marginal utility holds only under certain conditions. These conditions are referred to as the *assumptions* of the law. The assumptions of the law of diminishing marginal utility are listed below.

First, the unit of the consumer good must be a standard one, e.g., a cup of tea, a bottle of cold drink, a pair of shoes or trousers, etc. If the units are excessively small or large, the law may not hold.

Second, the consumer's taste or preference

Fig. 6.1 *Total and Diminishing*

must remain the same during the period of

Marginal Utility of Commodity X

consumption.

Third, there must be continuity in consumption. In case a break in continuity is necessary, the time interval between the consumption of two units must be appropriately short.

Fourth, the mental condition of the consumer must remain normal during the period of consumption. A person drinking whisky may feel a greater pleasure with successive pegs because of change in his mental status due to intoxication.

Given these conditions, the law of diminishing marginal utility holds universally.

In some cases, e.g., accumulation of money, collection of hobby items like stamps, old coins, rare paintings and books, melodious songs, etc., the marginal utility may initially increase rather than decrease. But eventually it does decrease. As a matter of fact, the law of diminishing marginal utility operates generally under normal conditions.

6.5.2 Consumer's Equilibrium: Cardinal Utility Approach

Having discussed the concept of *marginal utility* and the law of diminishing marginal utility, we take up the most important aspect of the theory of consumer behaviour. The issue is: how does a consumer attain, his equilibrium? In this section, we discuss the consumer's equilibrium by cardinal utility approach.

Conceptually, a consumer reaches his equilibrium position when he has maximized the level of his satisfaction, given his resources and other conditions. Technically, a utility-maximizing consumer reaches his equilibrium position when allocation of his expenditure is such that the last penny spent on each commodity yields the same utility. In order to explain how a consumer reaches his equilibrium, we begin with the simplest case of a consumer consuming only one commodity. The analysis will then be extended to explain equilibrium of a consumer consuming several goods.

(i) Consumer's Equilibrium: One-Commodity Model. Suppose that a consumer with a given money income consumes only one commodity, *X*. Since both his money income and commodity *X* have utility for him, he can either spend his money income on commodity *X* or retain it in the form of asset. If the marginal utility of commodity *X*, (MU_x) is greater than marginal utility of money (MU_m)[1], a utility-maximizing consumer will exchange his money income for the commodity. By assumption, MU_x is subject to diminishing returns (assumption 5), whereas marginal utility of money (MU_m) remains constant (assumption 6). Therefore, the consumer will continue to spend his money income on commodity *X* so long as $MU_x > P_x$. The utility-maximizing consumer reaches his equilibrium, i.e., the level of maximum satisfaction, where

$$MU_x = P_x (MU_m)$$

where $MU_m = 1$.

This is equilibrium condition in one commodity case. Alternatively, the consumer reaches equilibrium where

$$\frac{MU_x}{P_x (MU_m)} = 1$$

Consumer's equilibrium in a single commodity model is graphically illustrated in Fig. 6.2. The horizontal line marked $P_x (MU_m)$ shows the constant utility of money weighted by the price of commodity *X* (i.e., P_x) and MU_x curve represents the diminishing marginal utility of commodity *X*. The $P_x (MU_m)$ line and MU_x curve interest at point *E*. Consumer equilibrium condition, i.e., $MU_x = P_x (MU_m)$ is satisfied at point *E*. Therefore, the consumer is in equilibrium at point *E*.

At any point on the *MU* curve above point *E*, e.g., at point *M*, *MU* > *P* (*MU*). Therefore,
x
x
x
x

1. The utility of money income can be viewed in terms of its purchasing power, liquidity, asset function, social prestige, etc. However, in Marshallian analysis, utility of money is assumed to remain constant and one unit of money equals one util.

if the consumer exchanges his money for
commodity *X*, he will increase his total utility
because his *MU* is greater than his cost in
x
terms of *MU* . So the consumer makes a
m
net gain. For example, at point *M*, consumer
gains *MU* = *MQ* whereas the price that he
x
0
pays equals *NQ* . His marginal net gain equals
0
MQ – *NQ* = *MN*. This condition exists till
0
0
he reaches point *E*. When the consumer
reaches point *E*, his net utility gain is equal
to zero, as *MU* = *P* (*MU*).
x
x
m
Similarly, at any point below *E*, *MU* <
x
P (*MU*). Therefore, if he consumes more
x
m
than *OQ* , he loses more utility in terms of
x
price paid [*P* (*MU*)] than he gains. He
x
m
is therefore a net loser. The consumer can,

Fig. 6.2 *Consumer's Equilibrium*

therefore, increase his satisfaction by reducing his consumption. This means that at any point other than *E*, consumer's total satisfaction is less than maximum. Therefore, point *E* is the point of equilibrium.

(ii) Consumer's Equilibrium: Multiple Commodity Model. We have explained above consumer's equilibrium under an assumption that the consumer consumes a single commodity. In real life, however, a consumer consumes a large number of goods and services. So the question arises: How does a consumer consuming multiple goods reach his equilibrium? In this section, we explain consumer's equilibrium in the multi-commodity case.

We know from assumptions 2 and 5, that the consumer has limited income and that the utility which he derives from various commodities is subject to diminishing returns. We know also that the *MU* schedules of various commodities may not be the same. Some commodities yield a higher marginal utility and some lower for the same number of units consumed. In some cases, *MU* decreases more rapidly than in case of others for the same number of units consumed. A rational and utility-maximising consumer consumes commodities in the order of their utilities. He first picks up the commodity which yields the highest utility followed by the commodity yielding the second highest utility and so on. He switches his expenditure from one commodity to the other in accordance with their marginal utilities. He continues to switch his expenditure from one commodity to another till he reaches a stage where *MU* of each commodity is equal to per unit of money spent on each commodity. Theoretically, the consumer spends on different goods he consumes so that marginal utility derived from each good is equal. This is called the law of equi-marginal utility.

The law of equi-marginal utility explains the consumer's equilibrium in a multi-commodity model. This law states that a consumer consumes various goods in such quantities that the *MU* derived per unit of expenditure from each good is the same. In other words, a rational consumer spends his income on various goods he consumes in such a manner that each rupee spent on each good yields the same *MU*.

Let us now explain consumer's equilibrium in a multi-commodity model. For the sake of simplicity, however, we consider only a two-commodity case. Suppose that a consumer

consumes only two commodities, *X* and *Y*, their prices given as P_x and P_y, respectively.

Following the equilibrium rule of the single commodity case, the consumer will spend his income on commodities *X* and *Y* in such proportions that

$$MU_x = P_x (MU_m)$$

and

$$MU_y = P_y (MU_m)$$

Given these conditions, the consumer's equilibrium condition can be expressed as

$$\frac{MU_x}{P_x (MU_m)} = 1 = \frac{MU_y}{P_y (MU_m)} \quad \text{...(6.1)}$$

Since, according to assumption (6), *MU* of each unit of money is constant at 1, Eq. (6.1) can be rewritten as

$$\frac{MU_x}{P_x} = \frac{MU_y}{P_y} \quad \text{...(6.2)}$$

or

$$\frac{MU_x}{MU_y} = \frac{P_x}{P_y} \quad \text{...(6.3)}$$

y

Equation (6.2) leads to the conclusion that the consumer reaches his equilibrium when the marginal utility derived from each rupee spent on the two commodities *X* and *Y* is the same. In two-goods case, Eq. (6.3) reveals that a consumer is in equilibrium when *MU* ratio of any two goods equals their price ratio.

The two-commodity case can be extended to generalize the rule for consumer's equilibrium for a multiple commodity case, i.e., the case of a consumer consuming a large number of goods and services with a given income and at different prices. Supposing, a consumer consumes *A* to *Z* goods and services, his equilibrium condition may be expressed as

MUA MUB

MUZ

=

= =

= *MUm*

...(6.4)

A

P

B

P

Z

P

Equation (6.4) gives the Law of Equi-marginal Utility.

It is important to note that, in order to achieve his equilibrium, what a utility maximizing consumer intends to equalize is not the marginal utility of each commodity he consumes, but the marginal utility per unit of his money expenditure on various goods and services.

6.5.3 Derivation of Consumer's Demand Curve

The final outcome of the analysis of consumer behaviour is the formulation of the ***law of demand.*** In this section, we illustrate the derivation of ***demand curve*** and the basis of the formulation of the *law of demand*. In fact, the basis of consumer equilibrium itself provides the basis for the derivation of demand curve for a commodity2 demanded by an individual. Incidentally, it was Alfred Marshall who provided the logical basis for derivation of the *demand curve* and formulated the *law of demand* by using the consumer utility function.

The derivation of the demand curve is illustrated in Fig. 6.3 assuming a consumer consuming only one commodity, *X,* with his limited income. Panel (a) shows consumer's

2. Here, the commodity is supposed to be a 'normal good'. A normal good is one which is generally consumed by the people and whose consumption increases with decrease in its price.

equilibrium consumption at different levels of price of commodity *X*. Suppose marginal utility curve of the consumer is given by the curve *MU* and price of commodity *X*

X

is given at *P* . Given the *MU* curve and

3

price, the consumer is in equilibrium at point *E* and consumes *OQ* quantity of

1

1

commodity *X*. It implies that at price *P* ,

3

the consumer demands *OQ* quantity of

1

commodity. When price decreases from *P* to *P* , consumer equilibrium shifts to

3

2

point *E* and consumer's demand increases

2

from OQ to OQ. Similarly, when price

1

2

falls down further to P, consumer demand

1

increases to OQ. It may thus be concluded

3

that as price deceases, demand increases.

This inverse relationship between demand

and price gives rise to demand curve.

The derivation of demand curve is

illustrated in panel (b) of Fig. 6.3. The

demand curve D is drawn on the basis of

x

price-quantity relationship shown by panel

(a) of Fig. 6.3. Price-quantity combination

corresponding to equilibrium point E at

1

price P in panel (a) is shown at point

3

J in panel (b). Similarly, price-quantity

combinations at equilibrium points E and

2

Fig. 6.3 *Derivation of Demand Curve*

E are shown at points K and L, respectively

3

in panel (b). By joining points J, K and L, we derive the individual's demand curve for

commodity X. The demand curve D in panel (b) of Fig. 3.6 is the usual downward sloping

x

Marshallian demand curve.

6.5.4 Law of Demand under Variable MU of Money

The law of demand, as explained above, is based on the assumption that *utility of money remains constant.* It implies that if utility of money is variable, the law of demand based on constant utility of money will not hold. However, some economists have argued that the law of demand holds even under the condition of variable MU of money. This proposition is explained below.

Suppose a consumer with a limited income consumes only one commodity, X, and he is in equilibrium at a given price, P. At equilibrium, consumer's marginal utility of

x

commodity X, (MU) is equal to P times marginal utility of money, P (MU). Now

X

x

x

m

suppose P falls but the consumer buys only as much as he did before the fall in P.

x

x

Under this condition, the consumer saves money and hence his idle cash balance, i.e., money held idle, increases. With increase in money held idle, the MU deciles under

m

the condition of variable MU, whereas MU remains constant. Under this condition,

m

x

equilibrium condition is disturbed as MU is not equal to P (MU). In fact, P (MU) <

x

x

m

x

m

MU . To attain equilibrium, the consumer will have to buy more of commodity *X* to the

x

extent that *MU* = *P* (*MU*). It implies that with decrease in price, the consumer will have

x

x

m

to buy more of a good to maximize his utility. Thus, the law of demand holds under the condition of variable marginal utility of money also.

6.5.5 Why Demand Curve Slopes Downward to Right

As has been shown in Fig. 6.3(b), a normal *demand curve slopes downward to the right***.** A question arises: Why demand curve slopes downward to the right? The primary answer to this question lies in *utility maximizing behaviour of the consumer*. However, utility maximizing behaviour does not disclose certain inherrent factors that lie behind the consumer behaviour. In fact, there are some other factors also that cause increase in demand with increase in price. These factors are discussed in this section.

1. Income Effect of Price Change. When price of a commodity decreases, all other factors affecting demand remaining constant, then the real income of the consumer in terms of that product increases. In other words, purchasing power of the consumer increases. As is the general practice, increase in real income motivates consumers to consume more of normal goods. Therefore, demand for goods and services increases due to increase in real income caused by decrease in their price. The increase in demand due to increase in income is known as ***income effect.***

For example, suppose petrol price is fixed at `70 per litre and weekly demand for petrol by a car-owner is 50 litres. Under these conditions, the car-owner spends `750 per week. Now let the petrol price decrease tò50 per litre and the car-owner spends the same amount on petrol per week. Then he would be able to buy 70 litres (= `750/`50) per week. Note that decrease in petrol price (expenditure on petrol remaining the same) causes increase in petrol demand from 50 litres to 70 litres. This is ***income effect*** of decrease in price. The ***income effect*** of a decrease in price is illustrated graphically in Fig. 6.4.

As Fig. 6.4 shows, at petrol pricè70 per litre, the car-owner buys 50 litres of petrol. The car-owner is in equilibrium at point *A*. When petrol price decreases, he shifts to equilibrium point *B* where he consumes 70 litres of petrol. Note that petrol consumption increases but his total expenditure of petrol remaining the same, i.e., `350. This increase in petrol consumption arises due to increase in real income caused by the fall in price. This is *income effect*.

Similarly, when price of a product increases, consumer's real income (purchasing power) decreases. As a result demand for the product decreases. This kind of effect of change in price on real income makes demand curve slope downward to right.

It may be added that increase in demand for petrol due to decrease in petrol price may be more than the price effect, i.e., 20 more than litres in our example. This kind of change in demand takes place because people tend to spend more on a normal good when its price falls. In that case, the *income effect* is higher than the price effect.

Fig. 6.4 *Income Effect of Price Change*

2. Substitution Effect of Price Change. Substitution effect arises due to change in *relative price* 3 caused by change in the price of substitute goods. When price of a commodity falls, price of its substitute remaining constant, then the substitute goods becomes relatively costlier. Or, in other words, the commodity whose price falls becomes relatively cheaper. As it is a general practice, utility maximizing consumers tend to substitute cheaper goods for the costlier ones. As a result, demand for the relatively cheaper product increases. The increase in demand caused by this factor is called *substitution effect.* For example, suppose prices of petrol and diesel are given and both the goods are demanded in certain quantities per unit of time. As such, if the price of petrol decreases, diesel price remaining the same, then petrol becomes relatively cheaper. As result, consumers shift their demand, at least partially, from diesel to petrol. This kind of increase in demand is known as *substitution effect.*

3. Utility Maximizing Behaviour. As has already been explained in Section 6.5.2, the utility maximizing behaviour of the consumers is the basic factor behind the downward slope in demand curve. As mentioned above, a utility maximizing consumer of a commodity, say *X*, reaches equilibrium under the following condition.

$P_x = MU_X$

Given the equilibrium condition, when the price of commodity *X* falls, $P_x < MU_X$.

Under this condition, the utility maximizing consumer will have to reduce MU_X. To this end, the consumer will have to buy more of commodity *X*. This leads to increase in demand making demand curve to slope downward.

3. Relative price is the price of a commodity in relation to the price of its substitute.

6.5.6 Exceptions to the Law of Demand

The law of demand is a fundamental economic law. This law applies to deamnd for almost all consumer goods. However, there are some ***exceptions*** to the law of demand. The exceptions of the law of demand are described here briefly.

(*a*) ***Expectations of further price rise in future.*** When consumers expect a continuous increase in the price of a durable commodity, they buy more of it despite the increase in its price with a view to avoiding the pinch of a much higher price in future. For instane, in pre-budget months, prices generally tend to rise. Yet, people buy more storable goods in anticipation of further rise in prices due to raising of or imposition of indirect taxes. Similarly, when consumers anticipate a further fall in future in the falling prices, they postpone their purchases rather than buying more when there is a fall in the price just to take the advantge of any further fall in price.

(*b*) ***Status Goods.*** The law of demand does not apply to the commodities which are used as a 'status symbol' or for enhancing social prestige or for displaying wealth and richness, e.g., gold,4 precious stones, rare paintings, antiques, etc. Rich people buy such goods mainly because their prices are high and buy more of them when their prices move up rather than buying less. For example, gold price was shorting up in August and September 2013. Yet demand for gold continued of increase so much so gold that import increased to such an extent that current account deficit increased heavily.

(*c*) ***Giffen Goods.*** Another exception to the law of demand is the classic case of Giffen goods.5 A Giffen good is defined as an inferior good whose demand increases when its price increases. There are several inferior commodities (much cheaper than its superior substitutes) consumed by the poor households as an essential commodity. If the price of such goods increases (price of its substitute remaining constant), its demand increases instead of decreasing because, in case of a Giffen good, income effect of a price rise is greater than its substitution effect. The reason is, when price of an inferior good increases, income remaining the same, poor people cut the consumption of the superior substitutes so that they are able to buy sufficient quantity of the inferior good to meet their basic need. For instance, let us suppose that the monthly minimum consumption of foodgrains by a poor household is 20 kg of *bajra* (an inferior good) and 10 kg of wheat (a superior good). Suppose also that *bajra* sells at `5 per kg and wheat at `10 per kg and that the household spends its total income of `200 on these items. Now, if price of *bajra* increases tò6 per kg, the household will be forced to reduce the consumption of wheat by 5 kg6 and increase that of *bajra* by the same quantity in order to meet its minimum monthly consumption requirement of 30 kg of food-grains, its expenditure on foodgrains remaining the same. The consumer substitutes *bajra* for wheat because he can in no other way meet his basic minimum needs. Obviously, the household's demand for *bajra* increases from 20 kg to 25 kg per month despite the increase in its price.

4. Goods of this category are also purchased and stocked to store the value.

5. Named after Robert Giffen (1837–1970), a British Economist.

6. This can be worked out as follows. Suppose that the household substitutes *x* kgs of *bajra* for the same quantity

of wheat, so that (20 + x) + (10 – x) = 30 kg. Since the consumer spends only `2000, his pattern of consumption expenditure, at the new price of *bajra* (i.e., `6 per kg) will be 6 (20 + x) + 10 (10 – x) = `200.

By solving this equation for x, we get x = 5 kg.

6.5.7 Shift in Demand Curve

When a demand curve changes its position (retaining its slope though not necessarily), the change is known as a *shift in the demand curve*. For example, suppose that the original demand curve for commodity *X* is given as D_1 in Fig. 6.5. As shown in the figure, at price OP_2, the consumer would buy OQ_1 units of *X*, other factors remaining constant. But, if any of the other factors (e.g., consumer's income or price of the substitutes) changes, it will change the consumer's ability and willingness to buy commodity *X*. For example, if consumer's disposable income increases due to decrease in income tax, he would be able to buy OQ_2 units of *X* instead of OQ_1 at the same price, OP_2. His equilibrium point shifts from *A* to *B*. Suppose, this applies to the whole range of prices of *X*. Similarly, if price is given at OP_1 and consumer's income increases, he would demand more of *X*. His demand

Fig. 6.5 *Shift in Demand Curve*

increases form OQ_2 to OQ_3 and his equilibrium shifts from point *C* to *D*. By drawing a demand curve through points *B* and *D*, we get a new demand curve, D_2. This shows an *upward* shift in demand curve form D_1 to D_2. Similarly, decrease in disposable income of the consumer due to, say, rise in taxes may cause a *downward shift* in the demand curve from D_2 to D_1.

Factors Behind Shifts in the Demand Curve

Shifts in a price-demand curve may take place owing to the change in one or more non-price determinants of the demand for a commodity. Consider, for example, the increase in demand for commodity *X* by Q_1Q_2 in Fig. 6.5. Given the price OP_2, the demand for *X* might have increased by Q_1Q_2 for any of the following reasons.

(*i*) Increase in consumer's income so that he can buy OQ_2 of *X* at price OP_2: this is *income effect*;

(*ii*) Price of commodity *Y*, the substitute of *X*, rises so that the consumers find it gainful to prefer Q_1Q_2 of *X* for its substitute: this is *substitution effect*;

(*iii*) Advertisement by the producer of the commodity *X* changes consumer's taste or preference in favour of commodity *X* so much that the consumer buys more of *X* or he prefers *Q Q* to its substitute. This is again a *substitution effect* of advertisement.

1 2

(*iv*) Price of a complement of *X* falls so much that the consumer can afford *OQ* of *X*; and

2

(*v*) Price remaining the same, demand for *X* might increase also for such reasons as *X* gaining fashion status, improvement in its quality, change in production technology and seasonality of the product.

It is **important** for the business decision-makers to bear in mind the distinction between changes in demand due to (*i*) shift in price-demand curve; and (*ii*) movement along the demand curve. For instance, in Fig. 6.5, the increase in quantity demanded from

OQ to *OQ* can be explained in two different ways: ***one***, by moving down from point

1

2

A to *C* along the demand curve *D* which results from a fall in price from *P* to *P* ; and

1

2

1

two, through upward shift in demand curve from *D* to *D* . In the former case, additional

1

2

demand is obtained at the cost of some revenue. In the latter case, demand increases due to a shift in the demand curve on account of some other factors, such as increase in consumer's income, increase in the price of substitutes, increase in population, etc. This kind of increase in demand results in increase in revenue without decrease in price. However, in case the demand curve is made to shift through advertisement or other sales promotion devices, the additional demand is not free of cost. Moreover, it is the increase in demand attained by advertisement which is hoped for and attempted by business firms.

6.6 ANALYSIS OF CONSUMER BEHAVIOUR: ORDINAL UTILITY APPROACH

In the preceding section, we have discussed the consumer behaviour based on *cardinal utility approach*. In this section, we proceed to discuss consumer behaviour on the basis of ***cardinal utility approach***. The concept of *ordinal utility* has already been explained in Section 6.4. As mentioned there, the cardinal utility approach is based on the assumption that utility is cardinally or numerically measurable while ordinal utility approach is based on the assumption that utility is measurable only ordinally, not cardinally or numerically. In other words, utility can be expressed only comparatively. For example, a consumer consuming two googs, *X* and *Y*, can express his feeling of utility as utility of *X* is 'higher than' or 'lower than' that of *Y*. Although the concept of 'ordinal utility' was evolved much earlier, the analysis of consumer behaviour based on ordinal utility concept was developed by J.R. Hicks and R.G.D. Allen in 1934. The *ordinal utility approach* is also known as Hicks-Allen Approach. However, the basic theory of consumer behaviour was developed by J. R. Hicks. In order to analyse consumer behaviour, Hicks used a new tool of analysis called ***Indifference curve***. In this section, we will discuss consumer behaviour following the ordinal utility approach. We will first explain the 'indifference curve', the tool of analysis and then proceed to analyse consumer behaviour through the indifference curve.

6.6.1 Assumptions of Ordinal Utility Theory

1. ***Rationality.*** The consumer is assumed to be a rational being. Rationality means that a consumer aims at maximizing his total satisfaction given his income and prices of the goods and services he consumes and his decisions are consistent with this objective.

2. ***Ordinal utility.*** Indifference curve analysis is based on the assumption that utility is only ordinally expressible. That is, the consumer is only able to express the order of his preference for different baskets of goods.

3. ***Transitivity and consistency of choice.*** Consumer's choices are assumed to be transitive. *Transitivity* of choice means that if a consumer prefers *A* to *B* and *B* to *C*, he must prefer *A* to *C*. Or, if he treats *A* = *B* and *B* = *C*, he must treat *A* = *C*.

Consistency of choice means transitivity of choice over a period of time. For example, it means that if a consumer prefers *A* to *B* in one period, he does not prefer *B* to *A* in another period or even treat them as equal.

4. ***Nonsatiety.*** It is also assumed that the consumer has not reached the point of saturation in case of any commodity. Therefore, a consumer always prefers a larger quantity of all the goods.

5. ***Diminishing marginal rate of substitution.*** The consumer consumes substitutable goods and can substitute one good for another. The marginal rate of substitution is the rate at which a consumer is willing to substitute one commodity (*X*) for another (*Y*) so that his total satisfaction remains the same. The marginal rate of substitution is given as $\Delta Y/\Delta X$. The ordinal utility approach assumes that $\Delta Y/\Delta X$ goes on decreasing when a consumer continues to substitute *X* for *Y*.

6.6.2 Meaning and Nature of Indifference Curve

An indifference curve may be defined as the *locus of points each representing a different combination of two substitute goods, which yield the same utility or level of satisfaction to the consumer. Therefore, he is indifferent between any two combinations of two goods when it comes to making a choice between them*. Such a situation arises because he consumes a large number of goods and services and often finds that one commodity can be substituted for another. The consumer can, therefore, substitute one commodity for another, and can make various combinations of two substitute goods which give him the same level of satisfaction. Since each combination yields the same level of satisfaction, he would be indifferent between the combinations when he has to make a choice. When such combinations are plotted graphically, it produces a curve. This curve is called the *indifference curve*. An indifference curve is also called *iso-utility curve* or *equal utility curve*.

For example, let us suppose that a consumer consumes two goods, *X* and *Y*, and he makes five combinations *a*, *b*, *c*, *d* and *e* of the two substitute commodities, *X* and *Y*, as presented in Table 6.2. All these combinations yield the same level of satisfaction.

Table 6.2 *Indifference Schedule of Commodities X and Y*

Combination		*Units of Commodity Y*	+	*Units of Commodity X*	=	*Total Utility*
a	=	25	+	3	=	*U*
b	=	15	+	5	=	*U*
c	=	8	+	9	=	*U*
d						

=

4

+

17

=

U

e

=

2

+

30

=

U

Table 6.2 is an indifference schedule—a schedule of various combinations of two goods— *X* and *Y*–yielding the same level of utility. Therefore, the consumer is indifferent between the combinations. The last column of the table shows an undefined utility (*U*) derived from each combination of *X* and *Y*. The combinations *a*, *b*, *c*, *d* and *e* given in Table 6.2 are plotted and joined by a smooth curve (as shown in Fig. 6.6). The resulting curve is known as ***indifference curve***. On this curve, one can locate many other points between any two points showing different combinations of *X* and *Y* which yield the same level of satisfaction. Therefore, the consumer is indifferent between the combinations which may be located on the indifferent curve.

Indifference Map We have drawn a single indifference curve in Fig. 6.6 on the basis of the indifference schedule given in Table 6.2. The combinations of the two commodities, *X* and *Y*, given in the indifference schedule or those indicated by the indifference curve are by no means the only combinations of the two commodities. The consumer may make many other combinations with less of one or both of the goods—each combination yielding the same level of satisfaction but less than the level of satisfaction indicated by the indifference curve *IC* in Fig. 6.6. As such, an indifference curve below the

Fig. 6.6 *Indifference Curve*

one given in Fig. 6.6 can be drawn, say, through points *f*, *g* and *h*. Similarly, the consumer may make many other combinations with more of one or both the goods—each combination yielding the same satisfaction but greater than the satisfaction indicated by *IC*. Thus, another indifference curve can be drawn above *IC*, say, through points *j*, *k* and *l*. This exercise may be repeated as many times as one wants, each time generating a new indifference curve.

In fact, the space between *X* and *Y* axes is known as the ***indifference plane*** or ***commodity space***. This plane is full of finite points and each point on the plane indicates a different combination of goods *X* and *Y*. Intuitively, it is always possible to locate any two or more points on the indifference plane indicating different combinations of goods *X* and *Y* yielding the same level of satisfaction. It is thus possible to draw a number of indifference curves without intersecting or being

Fig. 6.7 *The Indifference Map*

tangent to one another as shown in Fig. 6.7. The set of indifference curves IC_1, IC_2, IC_3 and IC_4 drawn in this manner make the ***indifference map***. In fact, an indifference map may contain any number of indifference curves, ranked in the order of consumer's preferences.

6.6.3 Marginal Rate of Substitution (*MRS*)

An indifference curve is formed by substituting one good for another. The rate at which one good is substituted for another is called Marginal Rate of Substitution (*MRS*). In

other words, the *MRS* is the rate at which one commodity can be substituted for another, the level of satisfaction remaining the same. The *MRS* between two commodities *X* and *Y*, may be defined as the quantity of *X* which is required to replace one unit of *Y* (or quantity of *Y* required to replace one unit of *X*) in the combination of the two goods so that the total utility remains the same. This implies that the utility of *X* (or *Y*) given up is equal to the utility of additional units of *Y* (or *X*) added to the combination. The *MRS* is expressed as $\Delta Y/\Delta X$, moving down the curve.

MRS goes on Diminishing The basic postulate of ordinal utility theory is that if a consumer goes on substituting one good for another, the $MRS_{y,x}$ (or $MRS_{x,y}$) goes on decreasing. It means that the quantity of a commodity that a consumer is willing to sacrifice for an additional unit of another goes on decreasing when he goes on substituting one commodity for another. The diminishing $MRS_{x,y}$ obtained from combinations of *X* and *Y* given in Table 6.2 is presented in Table 6.3.

Table 6.3 *The Diminishing MRS between Commodities X and Y*

Indifference Points	*Combinations* $Y + X$	*Change in Y* $(-\Delta Y)$	*Change in X* (ΔX)	$MRS_{y,x}$ $(\Delta Y/\Delta X)$
a	25 + 3	–	–	–
b	15 + 5	– 10	2	– 5.00
c	8 + 9	– 7	4	– 1.75
d	4 + 17			

– 4

8

– 0.50

e

2 + 30

– 2

13

– 0.15

As Table 6.3 shows, when the consumer moves from point *a* to *b* on his indifference curve (Fig. 6.6) he gives up 10 units of commodity *Y* and gets only 3 units of commodity *X*. Thus,

−Δ *Y*

10

MRS

−

=

=

= −

y x

5.00

,

Δ *X*

2

As he moves down from point *b* to *c*, he loses 7 units of *Y* and gains 4 units of *X*, giving

−Δ *Y*

7

MRS

−

=

=

= −

y x

1.75

,

Δ *X*

4

Note that as the consumer moves from point *a* to *b* and from point *b* to *c*, the *MRS* decreases from –5.00 to –1.75. The *MRS* goes on decreasing as the consumer moves

y, *x*

further down along the indifference curve, from point *c* through *d* and *e*. *The diminishing marginal rate of substitution causes the indifference curves to be convex to the origin.*

Why Does MRS Diminish? The *MRS* diminishes along the *IC* curve because, in most cases, no two goods are perfect substitutes for one another. In case any two goods are perfect substitutes, the indifference curve will be a straight line having a negative slope showing constant *MRS*. Since goods are not perfect substitutes *, the subjective value attached to the additional quantity (i.e., subjective MU) of a commodity decreases fast in relation to the other commodity whose total quantity is decreasing*. Therefore, when the quantity of one commodity (*X*) increases and that of the other (*Y*) decreases, the subjective *MU* of *Y* increases and that of *X* decreases. Therefore, the consumer becomes increasingly *unwilling* to sacrifice more units of *Y* for one unit of *X*. But, if he is required to sacrifice

additional units of *Y*, he will demand increasing units of *X* to maintain the level of his satisfaction. As a result, the *MRS* decreases.

Furthermore, when combination of two goods at a point on the indifference curve is such that it includes a large quantity of one commodity (*Y*) and a small quantity of the other commodity (*X*), then consumer's *capacity* to sacrifice *Y* is greater than to sacrifice *X*. Therefore, he can sacrifice a larger quantity of *Y* in favour of a smaller quantity of *X*. For example, at combination *a* (see the indifference schedule, Table 6.2), the total stock of *Y* is 25 units and that of *X* is 5 units. That is why the consumer is willing to sacrifice

10 units of *Y* for 3 units of *X* (Table 6.3). This is an observed behavioural rule that the consumer's willingness and capacity to sacrifice a commodity is greater when its stock is greater and it is lower when the stock of a commodity is smaller.

These are the reasons why *MRS* between the two substitute goods decreases all along the indifference curve.

6.6.4 Properties of Indifference Curve

Indifference curves have the following four basic properties:

1. Indifference curves slope downward to right;
2. Indifference curves of imperfect substitutes are convex to the origin;
3. Indifference curves do not intersect nor are they tangent to one another;
4. Upper indifference curves indicate a higher level of satisfaction.

These properties of indifference curves, in fact, reveal the consumer's behaviour, his choices and preferences. They are, therefore, very important in the modern theory of consumer behaviour. Let us now look at their implications.

1. *Indifference Curves Slope Downward to Right.* In the words of Hicks,7 "... so long as each commodity has a positive marginal utility, the indifference curve must slope downward to the right", as shown in Fig. 6.6. The negative slope of an indifference curve implies two requisits: (*a*) that the two commodities can be substituted for each other; and (*b*) that if the quantity of one commodity decreases, qunatity of the other commodity must so increase that the consumer stays at the same level of satisfaction. The conditions make the indifference curve slope downward to the right. In case two goods are *perfect substitutes*, then the change in the combination of two goods produces an *indifference line,* not a curve.

2. *Indifference Curves are Convex to Origin.* Indifference curves are not only negatively sloped, but are also *convex to the origin* as shown in Fig 6.6. The *convexity of the indifference curves* is caused by two factors:

(*i*) the two commodities are *imperfect substitutes* for one another, and

(*ii*) the marginal rate of substitution (*MRS*) between the two goods decreases as a consumer moves along an indifference curve.

The postulate of diminishing *MRS*, is based on the fact that if a consumer substitutes one commodity (*X*) for another (*Y*), his willingness to sacrifice more units of *Y* for one additional unit of *X* decreases, as quantity of *Y* decreases. There are two reasons for

7. *Value and Capital*, 1956, p. 13.

diminishing MRS: (*i*) two commodities are not, in general, perfect substitutes for one another, and (*ii*) *MU* of a commodity increases as its quantity decreases and *vice versa*, and, therefore, more and more units of the other commodity are needed to keep the total utility constant.

3. *Indifference Curves Neither Intersect Nor are Tangent with One Another.* If two indifference curves intersect or are tangent with one another, it will yield two impossible conclusions: (*i*) that two equal combinations of two goods yield two different levels of satisfaction, and (*ii*) that two different combinations—one being larger than the other–yield the same level of satisfaction. Such conditions are impossible if the consumer's subjective valuation of a commodity is greater than zero.

Let us now test what happens when two indifference curves, intersect. As shown in Fig. 6.8, indifference curves, *IC* and *IC*¢, intersect each other at point *A*. Note that point *A* falls on both the indifference curves, *IC* and *IC*¢. It means that the same basket of goods (*OM* of *X* + *AM* of *Y*) yields different levels of utility below and above point *A* on the same indifference curve. This implies inconsistency in consumer's choice. The inconsistency that two different baskets of *X* and *Y* yield the same level of utility can be proved as follows. Consider two other points—point *B* on indifference curve *IC*¢ and point *C* on indifference curve *IC*, both points being on a vertical line.

Points *A*, *B* and *C* represent three different combinations of commodities *X* and *Y*. Let us call these combinations as *A*, *B* and *C*, respectively. Note that points *A* and *B* fall on the same indifference curve, *IC′*. It means that, in terms of utility, *A* = *B*

Note also that points *A* and *C* are placed

on the same indifference curve, *IC*. It means

that, in terms of utility,

A = *C*

Now look at its, implication. Since

A = *B*

and

A = *C*

Fig. 6.8 *Intersecting Indifference Curves*

∴

B = *C*

But if *B* = *C*, it would mean that in terms of utility,

ON of *X* + *BN* of *Y* = *ON* of *X* + *CN* of *Y*

Since ' *ON* of *X*' is common to both the sides, it means that

BN of *Y* = *CN* of *Y*.

But as Fig. 6.8 shows, *BN* > *CN*. Therefore, combinations *B* and *C* cannot be equal in terms of satisfaction. The intersection, therefore, violates the *transitivity rule* which is a logical necessity in indifference curve analysis. The same reasoning is applicable when two indifference curves are tangent with each other.

4. *Upper Indifference Curves Represent a Higher Level of Satisfaction than the Lower Ones.* An indifference curve placed above and to the right of another represents a higher level of satisfaction than the lower one. In Fig. 6.9, indifference curve *IC* is placed above

2

the curve *IC* . Therefore, *IC* represents a higher level of satisfaction. The reason is that an

1

2

upper indifference curve contains all along its length a larger quantity of one or both the goods than the lower indifference curve. And a larger basket of commodities is supposed to yield a greater satisfaction than the smaller one, provided *MU* of goods is greater than zero.

For example, consider the indifference

curves *IC* and *IC* in Fig. 6.9. The vertical

1

2

movement from point *a* on the lower indifference

curve *IC* to point *b* on the upper indifference

1

curve *IC* means an increase in the quantity of

2

Y by *ab*, the quantity of *X* remaining the same

(*OX*). Similarly, a horizontal movement from

point *a* to *d* means a greater quantity (*ad*) of

commodity *X*, quantity of *Y* remaining the same

(*OY*). The diagonal movement, i.e., from *a* to *c*,

means a larger quantity of both *X* and *Y*. Unless

the utility of additional quantities of *X* and *Y* are

equal to zero, these additional quantities will

yield additional utility. Therefore, the level of

satisfaction indicated by the upper indifference **Fig. 6.9** *Comparison between Lower and*

curve (*IC*) would always be greater than that

2

Upper Indifference Curves

indicated by the lower indifference curve (*IC*).

1

6.7 BUDGETARY CONSTRAINTS ON CONSUMER'S CHOICE:

LIMITED INCOME AND PRICES

Given the indifference map, a utility maximizing consumer would like to reach the highest possible indifference curve on his indifference map. But the consumer has two strong constraints: (*i*) he has a limited income, and (*ii*) he has to pay a price for the goods. Given the prices, the limitedness of income acts as a *constraint on how high a consumer can ride on his indifference map*. This is known as *budgetary constraint*. In a two-commodity model, the *budgetary constraint* may be expressed through a ***budget equation*** as

$$P_x \cdot Q_x + P_y \cdot Q_y = M$$

where P_x and P_y are prices of goods *X* and *Y* respectively; Q_x and Q_y are their respective quantities; and *M* is the consumer's money income.

The budget equation states that the total expenditure of the consumer on goods *X* and *Y* cannot exceed his total income, *M*. The quantities of *X* and *Y* that a consumer can buy, given his income (*M*) and prices, P_x and P_y, can be easily obtained from the budget equation, as shown below.

$$Q_y = \frac{M}{P_y} - \frac{P_x}{P_y} Q_x$$

and

$$Q_x = \frac{M}{P_x} - \frac{P_y}{P_x} Q_y$$

Now, Q_x or Q_y may be alternatively assigned any positive numerical value and the corresponding values of Q_y and Q_x may be obtained. When the values of Q_x and Q_y are plotted on the *X* and *Y* axes, we get a line with a negative slope, which is called the ***budget line*** or ***price line*,** as shown in Fig. 6.10.

An easier method of drawing the *budget line* is to mark point M/P_y on the *Y* axis (assuming Q = 0) and point M/ P on X-axis (assuming Q = 0) and to join these points by

x

x

y

a line. This gives the same budget line as given by the equation in Fig. 6.10. *The budget line shows the alternative options of commodity combinations available to the consumer given his income and the prices of X and Y.*

As can be seen in Fig. 6.10, budget line divides the *commodity space* into two parts: (*i*) feasible area, and (*ii*) non-feasible area. The area under the budget line (including the budget line) is *feasible* area. Any combination of goods *X* and *Y* represented by a point within this area (e.g., point *A*) or point *P* on the boundary line (i.e., on the budget line) is a feasible combination, given *M*, *P* and *P* . The area beyond the budget

x

y

line is *non-feasibile* area because any point falling in this area, e.g., point *B*, is unattainable (given *M*, *P* and *P*).

x

y

Shifts in the Budget Line

Fig. 6.10 *Budget Line and Budget Space*

The budget line is drawn on the basis of the budget equation, given as $M = P \,.\, Q + P \,.\, Q$. Any change in the parameters of the

x

x

y

y

budget equation, viz., *M*, *P* and *P* , make the budget line shift upward or downward or

x

y

swivel left or right and up or down. If consumer's income (*M*) increases, prices remaining constant, the budget line shifts upwards remaining parallel to the original budget line. Suppose the original budget line is given by line *AB* in Fig. 6.11. If *M* increases (prices remaining the same), the budget line *AB* will shift to *CD*. And, if *M* decreases by the same amount, the budget line will shift backward to its original position *AB*. Income remaining the same, if prices change, the budget line changes its position and slope. For example, if *M* and *P* remain constant and *P* decreases to half then the budget line will be

y

x

AF. Similarly, *M* and *P* remaining constant, if *P* increases, the budget line shifts to *EB*.

x

y

Slope of the Budget Line

Another important aspect of the budget line that matters in determining a consumer's equilibrium is its *slope*. The slope of the budget line (*AB*) in Fig. 6.11, is given as

Δ *Qy OA*

=

Δ *Q*

OB

x

Since $OA = M/P_y$ (at $X = 0$) and $OB = M/P_x$ (at $Y = 0$), the *slope of the budget line AB* in Fig. 6.11 may be rewritten as

$$\frac{OA}{OB} = \frac{M/P_y}{M/P_x} = \frac{P_x}{P_y}$$

Thus, the slope of the budget line is the same as the *price ratio* of the two commodities.

6.8 CONSUMER'S EQUILIBRIUM: ORDINAL UTILITY APPROACH

Now that we have discussed the indifference map and the budget line, we turn to analyze consumer's equilibrium. As noted earlier, a consumer attains his equilibrium when he maximizes his total utility, given his income and market prices of the goods and services that he consumes.

Fig. 6.11 *Shift in the Budget Space*

The ordinal utility approach specifies two conditions for the consumer's equilibrium:

(*i*) Necessary condition, i.e., the first order condition, and

(*ii*) Supplementary condition, i.e., the second order condition.

In a two-commodity model, the necessary or the ***first order condition*** under ordinal utility approach is the same as equilibrium condition under cardinal utility approach. It is given (see section 6.5.2, Eq. (6.3) as

$$\frac{MU_x}{P_x} = \frac{MU_y}{P_y}$$

Since, by implication (as shown below), $MU_x / MU_y = MRS_{x, y}$, the necessary condition of equilibrium under ordinal utility approach can be expressed as

$$MRS_{x, y} = \frac{MU_x}{MU_y} = \frac{P_x}{P_y}$$

This is a necessary but not a *sufficient condition* of consumer's equilibrium. The ***second order*** **or** ***supplementary condition*** requires that the necessary condition be fulfilled at the *highest possible indifference curve*.

Consumer's equilibrium is illustrated in Fig. 6.12. The indifference curves IC_1, IC_2 and IC present a hypothetical indifference map of the consumer. The line *AB* is the

3

hypothetical budget line. The budget line *AB* and the indifference curve *IC* are tangent

2

at point *E*. The slope of the indifference curve gives the *MRS* and slope of budget line

gives *P* / *P* . At point *E* of *IC* , *MRS* = *P* / *P* . Therefore, consumer is in equilibrium at

x

y

2

x, *y*

x

y

point *E*. This point can be proved as follows.

We know that between any two points on an indifferent curve,

Δ *Y* · *MU* = Δ *X* · *MU* .

y

x

Given this condition, the slope of an indifference curve (Δ *Y*/Δ *X*) can be worked out as

Δ *Y*

MUx

=

= *MRS*

Δ *X*

MU

y, *x*

y

We know also that the slope of the budget line is given by

OA

x

P

=

OB

y

P

As shown in Fig. 6.12, at point *E*, *MRS* , the slope of the IC is equal to *P* / *P* , the slope

y,*x*

2

x

y

of the budget line *AB*. Therefore, the consumer is in equilibrium at point *E*. The tangency

of *IC* with the budget line *AB*, indicates

2

that *IC* is the highest possible indifference

2

curve which the consumer can reach, given

his income and the prices. Since *IC* is the

2

highest possible attainable indifference curve,

both necessary and supplementary conditions

of equilibrium are satisfied at point *E*. The

consumer is, therefore, in equilibrium at

point *E*.

Why Equilibrium Not at Points

J and K?

Note that the necessary condition is satisfied

also on two other points, *J* and *K* (i.e., the

points of intersection between the budget

line *AB* and indifference curve *IC*). But

1

Fig. 6.12 *Equilibrium of the Consumer*

these points do not satisfy the *second order*
condition as indifference curve *IC* is not the highest possible curve on which the necessary
1
condition is fulfilled. Since indifference curve *IC* lies below the curve *IC* , the level of
1
2
satisfaction at any point on *IC* is lower than the level of satisfaction indicated by *IC* . So
1
2
long as the utility maximizing consumer has an opportunity to reach the curve *IC* , the
2
rationality condition demands that he would not like to settle on a lower indifference curve.

From the information contained in Fig. 6.12, it can be proved that the level of
satisfaction at point *E* is greater than that on any other point on *IC* . Suppose the consumer
1
is at point *J*. If he moves to point *M*, he will be equally well-off because points *J* and
M are on the same indifference curve. If he moves from point *J* to *M*, he will have to
sacrifice *JP* of *Y* and take *PM* of *X*. But in the market, he can exchange *JP* of *Y* for *PE*
of *X*. That is, he gets extra *ME* (= *PE* – *PM*) of *X*. Since *ME* gives him extra utility, the
consumer moves to point *E* which means a utility higher than the point *M*. Therefore,
point *E* is preferable to point *M*. The consumer will, therefore, have a tendency to move
to point *E* from any other point on the curve *IC* in order to reach the highest possible
1
indifference curve, all other things (taste, preference and prices of goods) remaining the
same.

Another fact which is obvious from Fig. 6.12 is that, due to budget constraint, the
consumer cannot move to an indifference curve placed above and to the right of *IC* . The
2
reason is that his income would be insufficient to buy any combination of two goods at
the curve *IC* . Note that the indifference curve *IC* falls in the infeasibility area.
3
3

6.8.1 Effects of Change in Income on Consumption

We have been concerned so far with the consumer's behaviour under the assumption
that consumer's income and market prices of goods and services remain constant. Let
us now drop these assumptions one by one and examine the consumer's response to the
changes in his income and prices of goods. In this section, we examine the effects of
changes in consumer's income on his consumption behaviour, assuming that prices of all
goods and services, and consumer's tastes and preferences remain constant. The change
in consumption basket due to change in income is called **income effect**.

Income Effect on Normal Goods

The effect of change in income on the consumption of normal goods is illustrated in
Fig. 6.13. To analyze the effect of change in income on consumption, let us suppose that
a consumer is initially in equilibrium
at point *E* on indifference *IC* . Now
1
1
let the consumer's income increase
so that his budget line shifts upward
from *AJ* to *BK* and the consumer
reaches a new equilibrium point, *E* on
2
IC . Similarly, if his income increases
2
further, he moves from equilibrium

E_2 to E_3 and then to E_4. Thus, with each successive upward shift in the budget line, the equilibrium position of the consumer moves upward. The successive equilibrium combinations of goods (*X* and *Y*) at four different levels of income are indicated by points E_1, E_2, E_3 and E_4 in Fig. 6.13. If these points of equilibrium are joined by a curve, we get the path of increase in consumption

Fig. 6.13 *Income Consumption Curve of Normal Goods*

resulting from the increase in income.

This curve is called the ***income consumption curve (ICC)***. The income-consumption curve may be defined as *the locus of points representing various equilibrium quantities of two commodities consumed by a consumer at different levels of income, all other things remaining constant.* The movement from point E_1 towards point E_4 indicates increase in the consumption of both the normal goods *X* and *Y* with increase in income. This is called ***income effect***.

Income-Effect on Inferior Goods

The effect of change in income on the consumption of different kinds of goods is not uniform. It may be *positive* or *negative* or *even neutral* depending on the nature of a commodity. In case of normal goods, income-effect is positive and in case of inferior goods, it is negative. By definition, an ***inferior good*** is one whose consumption decreases when income increases. Fig. 6.14 (*a*) and (*b*) present the case of inferior goods and *negative income effect* on their consumption. In Fig. 6.14 (*a*), *X* is assumed to be an

inferior good—its consumption decreases when consumer's income increases indicated by the leftward trend in the *ICC* curve. Similarly, in Fig. 6.14 (*b*), *Y* is considered to be an inferior commodity. Therefore, consumption of *Y* decreases with increase in income. It means that income effect on the consumption of inferior goods is negative.

In fact, whether a commodity is a 'normal good' or an 'inferior good' in economic sense, depends on whether income-effect on its consumption is positive or negative. If income effect is positive, the commodity is considered to be a 'normal good' and if it is negative, the commodity is said to be an 'inferior good'. Thus, the income-consumption-curve may take various shapes depending on whether a commodity is a 'normal good' or an 'inferior good'.

Fig. 6.14 *Income-Consumption Curve of Inferior Goods*

6.8.2 Effects of Change in Prices on Consumption

Let us now examine the effects of change in price on consumer behaviour, income remaining constant. As noted earlier, when price of a commodity changes, the slope of the budget line changes, which changes the condition for consumer's equilibrium. A rational consumer adjusts his consumption basket with a view to maximizing his satisfaction under the new price conditions. This change in consumption basket is called ***price-effect***. ***Price-effect*** *may be defined as the change in the quantity consumed of a commodity due to change in its price.* To examine price-effect, let us recall our two-commodity model, and assume change in price of commodity *X*, holding constant the consumer's income, his taste and preference and the price of commodity *Y*. The consumer's response to a change in the price of *X* and the resulting change in the combination of the two goods are illustrated in Fig. 6.15.

Suppose that the consumer is initially in equilibrium at point E_1. Now let the price of X fall, *ceteris paribus*, so that the consumer's budget line shifts from its initial position LR to the position LS. As a result, the consumer reaches a higher indifference curve IC_2 and his new equilibrium point is E_2. Here, his consumption of X increases by $FE_2 = UR$ and consumption of Y decreases by E_1F. This is the *price-effect* on the consumption of commodity X.

The Price-Consumption Curve

As shown in Fig. 6.15, with a successive fall in the price of X, consumer's equilibrium shifts from E_2 to E_3 and from E_3 to E_4. By joining the points of equilibrium E_1, E_2, E_3 and E_4, we get a curve called ***price-consumption-curve (PCC)***. *Price-consumption-curve is a locus of points of equilibrium on indifference curves, resulting from the change in the price of a commodity.* The price-

Fig. 6.15 *Price-Consumption Curve*

consumption-curve (PCC) shows the change in consumption basket due to a change in the price of commodity X. It can be seen from Fig. 6.15 that the quantity of X consumed goes on increasing whereas consumption of Y first decreases and then increases. The decrease in the consumption of Y due to fall in the price of X is the substitution effect and increase in consumption of Y is the ***real-income effect***.

6.8.3 Income and Substitution Effects of Price Change

As noted above, the change in consumption basket due to change in the price of consumer goods is called 'price effect'. Price-effect consists of two effects: (*i*) income-effect and (*ii*) substitution-effect. ***Income-effect*** results from the increase in *real income* due to a decrease in the price of a commodity. ***Substitution-effect*** arises due to the consumer's inherent tendency to substitute cheaper goods for the relatively expensive ones.

In this section, we will discuss the methods of measuring the total *price-effect* (PE), and its *income effect* (IE) and *substitution effect* (SE). There are two approaches of decomposing the total price-effect into income and substitution-effects, viz., (*i*) *Hicksian approach* and (*ii*) *Slutsky's approach*.

Measuring Income and Substitution Effects by Hicksian Approach

Hicksian method of measuring income and
substitution effects of a price change is
illustrated in Fig. 6.16. Let the consumer
be in equilibrium initially at point *P* on
indifference curve *IC* and budget line

Fig. 6.16 *Income and Substitution Effects:*

1

Hicksian Approach

M N . At point *P*, the consumer consumes *PX* of *Y* and *OX* of *X*. Now let the price of

1 1

1

1

X fall, price of *Y* remaining the same, so that the new budget line is *M N* . The new budget

1 3

line (*M N*) is tangent to *IC* at point *Q*. At this point, the consumer buys an additional

1 2

2

quantity (*X X*) of *X*. That is, total price effect = *X X* .

1 3

1 3

Now the problem is how to split the price-effect (*X X*) into *income* and *substitution*

1 3

effects. We know that *X X* = *IE* + *SE*. Given this equation, if any of the two effects is

1

3

known, the other can be easily measured. The general practice is to first measure income-effect component of the price-effect and then deduct it from the price effect to find the substitution-effect.

The Hicksian method of finding income-effect is to reduce the consumer's income (by way of taxation) so that he returns to his original indifference curve *IC* , to an equilibrium

1

point conforming to the new price ratio. This has been done by drawing an imaginary budget line (*M N*) parallel to *M N* and tangent to indifference curve *IC* at point *R*. Point

2 2

1 1

1

R is thus the income-adjusted equilibrium of the consumer at the new price ratio of *X* and *Y*, after the elimination of the real income-effect caused by the fall in the price of *X*. The shift in equilibrium from *Q* to *R* means that the consumer cuts his comsumption of *X* by *X X* due to fall in his income. This gives the measure of **income-effect** (*X X*)

2 3

2 3

caused by the increase in real income of the consumer due to fall in price of *X*. With income effect measured at *X X* , the substitution effect (*SE*) can be easily obtained as

2 3

SE = *PE* – *IE*. By using this method, we get substitution effect as *X X* – *X X* = *X X* .

1 3

2 3

1 2

In Fig. 6.16, the movement of the consumer from *P* to *R* shows his response to the change in relative price ratio, his real income being held constant at its original level. The consumer's movement from point *P* to *R* means an increase in quantity demanded of *X* by *X X* . This change in quantity demanded is called ***substitution-effect***.

1 2

The outcome of the above exercise may be summarized as follows:

Price effect

= *X X*

1 3

Income effect

= *X X* – *X X* = *X X*

1 3

1 2

2 3

Substitution effect

= *X X* – *X X* = *X X*

1 3

2 3

1 2

Income and Substitution Effects by Slutsky's Approach

An alternative method of measuring income and substitution effects of price change was suggested by a Russian mathematician, Eugene Slutsky. Slutsky's method of measuring income and substitution effects is similar to the Hicksian method. There is however an important difference between Hicks' and Slutsky's methods of measuring the real income-effect of a fall in the price of a commodity. Under Hicksian method, consumer's income has to be so reduced that he moves back to his original *IC* curve though on a different equilibrium point whereas, under Slutsky's method, consumer's income has to be so reduced that he moves back not only to the original indifference curve but also to his original equilibrium point.

The Slutskian method of splitting the total price-effect into income and substitution effects is demonstrated in Fig. 6.17. Let us suppose that the consumer is initially in equilibrium at point *P* on indifference curve *IC* . When price of *X* falls, other things

1

remaining the same, the consumer moves to another equilibrium point *Q* on indifference curve *IC* . The movement from point *P* to *Q* increases the consumer's purchase of *X* by

3

X X . This is the ***total price-effect*** caused by the fall in the price of *X*.

1

3

Now the problem is to measure the substitution and income effects by Hicksian approach. According to the Slutskian approach, a consumer's real income is so reduced that he returns to his original equilibrium point *P* and he is able to purchase his original combination of the two goods (i.e., *OX* of *X* and *PX* of *Y*) at the new

1

1

price ratio. This is accomplished by drawing an imaginary budget line, *M N* through the point *P*. Since

2 2

the whole commodity space is full of indifference curves, one of the indifference curves (*IC*) is tangent

2

to the imaginary budget line *M N* at

2 2

point *R*. The movement from point *Q*

Fig. 6.17 *Income Substitution Effects: Slutsky's*

to *R* shows a fall in the consumption of

Approach

X by *X X* . This is **Slutskian income effect**. We may now easily find out the substitution effect

2

3

(*SE*) by subtracting the income-effect (*IE*) from the total price-effect (*PE*), as shown below.

Substitution Effect = *PE* – *IE* = *SE*

= *X X* – *X X* = *X X*

1

3

2

3

1

2

Thus, according to Hicksian approach, the movement from *P* to *R* and the consequent increase in the quantity purchased of *X* (i.e., *X X*) is the ***substitution effect***. Similarly,

1

2

the consumer's movement from *R* to *Q* and the increase in the quantity purchased of *X* is the ***income-effect.***

Hicksian Approach Versus Slutskian Approach Figure 6.18 presents a comparison of the Hicksian and Slutskian approaches of splitting price-effect into substitution and income effects and also their results. The Slutskian approach holds the *apparent real income* constant by adjusting consumer's real income by the amount of 'cost-difference' so that the consumer is left with an income just sufficient to buy the *original combination* of the goods. The Hicksian approach, on the other hand, holds constant the *real income* so that the consumer is able to stay on the *original indifference* curve though consumer basket may change. To express the difference graphically, Hicksian method puts the consumer on the original indifference curve whereas Slutskian method makes the consumer move to an *upper indifference curve*.

Fig. 6.18 *Hicksian Approach vs. Slutskian Approach*

The difference between the two methods illustrated graphically in Fig. 6.18. Let the consumer be initially in equilibrium at point *P* on indifference curve *IC* . When the price of *X*

1

falls the consumer moves to a new equilibrium point *Q*. The movement from *P* to *Q* gives the total price-effect which equals *X X* of commodity *X*. Up to this point, there is no difference

1

4

between Slutsky and Hicks. Beyond this point, they differ. According to the Slutskian approach, the income adjusted budget line is drawn through the original equilibrium point, *P*, as shown by the line *M N* . As a result the consumer moves to an uppor *IC* curve *IC* at equilibrium

2 3

2

point *I*. The movement from *P* to *T* gives the substitution effect and the movement from *T* to *Q* gives the income effect. According to the Hicksian approach, the income-adjusted budget line is so drawn that it is tangent to the original indifference curve *IC* . So the consumer reaches a

1

new equilibrium point *T*. The movement from *P* to *R* gives substitution effect and movement from *R* to *Q* gives the income effect. The substitution and income effects of Slutskian and Hicksian approaches are summed up in quantitative terms in the following table.

Method

Price-effect

Substitution-effect

Income-effect

Hicksian

X X

X X

X X

1 4

1 2

2 4

Slutskian

X X

X X

X X

1 4

1 3

3 4

Other Differences

Apart from the above difference between the two methods, there are some other differences also. While the Hicksian approach is considered as a 'highly persuasive solution' to the problem of splitting price-effect into substitution and income effects, the Slutskian approach is intuitively 'perhaps less satisfying'.8 But the merit of the Slutskian approach is that substitution and income effects can be directly computed from the observed facts, whereas the Hicksian measure of these effects cannot be obtained without the knowledge of a consumer's indifference map.

Both the methods have, however, their own merits. The merit of the Slutskian method, which Hicks calls the 'cost-difference' method, lies in its property that it makes income effect easy to handle. Hicks has himself recognized this merit of the Slutskian method.

The merit of Hicksian method or 'compensating variation method' is that it is a more convenient method of measuring the substitution effect. In Hicks' own words, "The merit of the cost-difference method is confined to [its] property... that its income effect is peculiarly easy to handle. The compensating variation method [i.e., his own method] does not share in this particular advantage; but it makes up for its clumsiness in relation to income-effect by its convenience with relation to the substitution effect."9

Applicability of Income and Substitution Effects

Let us now look into the purpose of separating the income and substitution effects. As Hicks10 has pointed out, "substitution effect is absolutely certain; it must always work in favour of an increase in demand for a commodity when the price of that commodity falls." Thus, the behaviour of substitution effect is predictable; it follows directly from the principle of diminishing marginal rate of substitution. On the contrary, "income effect is not so reliable'11 and its behaviour is

8. Baumol, W.J., *Economic Theory and Operations Analysis*, Fourth Edition, New Delhi.

9. Hicks, J.R. *A Revision of Demand Theory*, (Oxford, 1969), p. 69.

10. Hicks, J.R. *Value and Capital* p. 32.

11. Hicks, J.R. *Value and Capital*, p. 32.

unpredictable in general. In fact, whether income-effect is positive or negative depends on whether a commodity is treated by the consumer as a 'superior' or an 'inferior' good. Since the subjective valuation of a commodity may vary from person to person, the response of the consumer in general to a change in real income becomes uncertain and unpredictable. It is quite likely that in some cases, substitution effect works in a positive direction, while income-effect works in a negative direction. In such cases a systematic analysis of price-demand relationship becomes an extremely difficult task. It becomes necessary, therefore, to eliminate the unpredictable income effect so that the behaviour of the predictable substitution effect can be known. From application point of view, apart from its analytical importance, the knowledge of 'how powerful is the substitution effect' is essential for formulating an appropriate pricing strategy.

6.9 DERIVATION OF INDIVIDUAL DEMAND CURVE

The basic purpose of the entire analysis of indifference curve technique is to construct the individual demand curve for a commodity. As stated earlier, the individual demand curve shows the relationship between the quantity demanded of a commodity (say *X*) by an individual and its price (*P*) under the *ceteris paribus* assumption. Thus, to draw an individual

x

demand curve, we need different levels of *P* and the corresponding *Q* . This information can

x

x

be obtained from the price-consumption curve. The price-consumption curve (*PCC*), in Fig.

6.19 (*a*), contains the information required for constructing the individual demand curve for *X*.

In Fig. 6.19 (*a*), *X*-axis measures the quantity of commodity *X* and *Y*-axis measures

commodity *Y*. It is assumed that consumer's money income is given as *M*. Given the

money income, *P* can be measured as *M/Q* . It is also assumed that *P* remains constant.

x

x

y

As Figs. 6.19 (*a*) and (*b*) show, with *P* decreasing from say, *P* to *P* and then to *P* , the

x

3

2

1

budget line rotates anti-clockwise, from *MN* to *MN* and then to *MN* . As a result, the

1

2

3

consumer moves from equilibrium point *E* to *E* and finally to point *E* on the *PCC*. The

1

2

3

shift in equilibrium indicates rise in consumption of *X* following the fall in *P* . *x*

From the information contained in Fig. 6.19 (*a*), we may construct a demand schedule

as given in the following table. The demand curve may be constructed by plotting the

demand schedule given below.

Price

Equilibrium

Quantity demanded of X

P = M/ON

E

OX

3

1

1

1

P = M/ON

E

OX

2

2

2

2

P = M/ON

E

OX

1

3

3

3

Note that *P* > *P* > *P* and *OX* < *OX* < *OX* .

3

2

1

1

2

3

The demand curve may be constructed directly from Fig. 6.19 (*a*). This has been shown in Fig. 6.19 (*b*). In this figure, vertical axis of panel (*b*) measures the price of commodity *X* and horizontal axis measures quantity of commodity *X*. As the figures shows, *P* , *P* and *P* represent

3

2

1

the three price levels we have considered in part (*a*). If we draw horizontal lines from *P* , *P* ,

3

2

and *P* , we get the three price levels in part (*b*). We know from panel (*a*) that at price *M*/ *ON*

1

1

= *P* , the quantity consumed is *OX* . If we extend the ordinate *E X* to the *X*-axis of part (*b*), it

3

1

1 1

intersects the line *P* = *M*/ *ON* at point *e* . By repeating this process for equilibrium points *E* and

3

1

1

2

E in Fig. 6.19 (*a*), we get points *e* and *e* in Fig. 6.19 (*b*). By joining the points *e* , *e* , and *e* 3

2

3

1

2

3

we get the demand curve *D* . This demand curve is the same as the Marshallian demand curve.

x

Fig. 6.19 *Derivation of Individual Demand Curve*

Possible Shapes of the Demand Curve The precise shape and slope of a demand curve derived from indifference curves depends on the direction in which income and substitution effects work as a result of fall in the price of a commodity. In fact, the *substitution effect is always negative but income-effect is uncertain*. Therefore, given the negative substitution effect, the shape and nature of demand curve depends on the direction and magnitude of the income-effect. There are four possible combinations of substitution and income effects for a fall in the commodity price and the corresponding nature of the demand curve. These may be summarized as follows:

1. When substitution-effect is negative and income-effect is positive, quantity demanded of *X* increases as *P* decreases. The demand curve, therefore, slopes downward to the

x

right. This is the case with ‘normal goods’.

2. If income-effect is negative but less than the (negative) substitution effect (as happens in case of inferior goods) the demand curve slopes downward to the right more steeply than usual.

3. If income-effect is zero, the demand curve follows the substitution effect, i.e., as price decreases, demand increases. The demand curve has a negative slope, but is relatively flatter.

4. If income-effect is negative and more powerful than the substitution-effect (as happens in the case of Giffen goods) the demand curve becomes backward

bending, as shown in Fig. 6.20. But it is most unlikely that any demand curve will slope downward to the left throughout its whole length.[12] It will be so only over that range of price changes over which the negative income-effect is stronger than the substitution-effect. Therefore, the most likely shape of the demand curve[13] for a Giffen good is one shown in Fig. 6.20. The demand curve for a Giffen good slopes downward till price falls to P_2. But if price falls further, the income-effect may become negative and so powerful that it outweighs the substitution effect. Then the demand curve for a Giffen good becomes a backward sloping one. If price continues to fall, say below P_1, the demand may once again increase for the Giffen good. This seems to be the most likely shape of the demand curve for a Giffen good.

Fig. 6.20 *Demand Curve for a Giffen Good*

6.10 THE REVEALED PREFERENCE THEORY

The cardinal and ordinal utility approaches to demand analysis, discussed in the preceding sections, are based on two different concepts of utility–cardinal or ordinal. The cardinal approach assumes absolute or cardinal measurability of utility and ordinal approach assumes relative or ordinal or introspective measurability of utility. While cardinal measurement of utility is not practicable, the introspective ordinal utility is non-observable. Thus, both these approaches involve problems of measurability of utility. In an attempt to overcome this problem, Samuelson[14] proposed in 1947 another theory of demand called 'Revealed Preference Theory' of consumer behaviour.

The main merit of the revealed preference theory is that the 'law of demand' can be directly derived from the revealed preference axioms without using indifference curves and most of its restrictive assumptions. What is needed is simply to record the observed behaviour of the consumer in the market. The consumer reveals his behaviour by the basket of goods he buys at different prices.

12. Stonier A.W. and D.C. Hague, *A Textbook of Economic Theory*, (London, Longman, 1972), p. 71.

13. Stonier and Hague, *op. cit.* , p. 71.

14. Samuelson, *P.A., Foundation of Economic Analysis,* (Cambridge, Mass., Harvard University Press, 1947), Chs. 5 and 6; and "Consumption Theory in Terms of Revealed Preference," *Economica*, November 1948, pp. 243–53. Samuelson had however conceived the idea much earlier in his paper "A Note on the Pure theory of Consumer's Behaviour," *Economica*, February and August 1938.

Besides, the revealed preference theory is also capable of establishing the existence of indifference curves and their convexity. For its merits, revealed preference theory is treated as the 'third root of the logical theory of demand'.

Assumptions

Samuelson's revealed preference theory is based on the following straight forward assumptions:

1. ***Rationality.*** The consumer is assumed to be a rational being. In his order of preferences, he prefers a larger basket of goods to the smaller ones.
2. ***Transitivity.*** Consumer's preferences are assumed to be *transitive*. That is, given alternative baskets of goods, *A*, *B* and *C*, if he treats $A > B$ and $B > C$, then he treats $A > C$.
3. ***Consistency.*** It is also assumed that during the analysis, consumer's taste remains constant and consistent. *Consistency* implies that if a consumer, given his circumstances, prefers *A* to *B* he will not prefer *B* to *A* under the same conditions.
4. ***Effective price inducement.*** Given the collection of goods, the consumer can be induced to buy a particular collection of goods by providing him sufficient price incentives. That is, for each collection, there exists a price line which makes it attractive for the consumer.

Revealed Preference Axiom

The revealed preference axiom can be stated as follows. Given the budgetary constraint and alternative baskets of goods having the same price, if a consumer chooses a particular basket, he reveals his preference for the basket. For example, suppose there are two alternative baskets *A* and *B* of two goods *X* and *Y*. Both the baskets being equally expensive, if a consumer chooses basket *A* rather than basket *B*, he reveals his preference for basket *A*.

If a consumer chooses a particular
basket, he does so either because he likes
it more or it is less expensive than the
other. In our example, if the consumer
chooses basket *A* rather than *B* because *A*
is cheaper, then the preference for *A* is not
revealed because the consumer might regret
not having been able to buy basket *B*. But,
if both the baskets are equally expensive,
then there is only one plausible explanation,
i.e., the consumer likes basket *A* more than
basket *B*. In this case, the consumer reveals
his preference for basket *A*.

The revealed preference axiom has been

Fig. 6.21 *Revealed Preference*

illustrated in Fig. 6.21. The consumer's
budgetary constraint has been shown by his budget line *MN*. If he chooses a particular
bundle of *X* and *Y* represented by point *A* on the budget line, it implies that he prefers
point *A* to any other point on the budget line, say point *B*. Since point *B* is on the budget
line, it is as much expensive as *A*. If the consumer chooses point *A*, it means that *A* is
revealed to be preferable to *B* and *B* is revealed inferior to *A*. Any point below the budget
line, like point *C*, represents a smaller and cheaper basket of *X* and *Y* and hence it is
revealed inferior to *A*. Therefore, any point above the budget line, like point *D*, represents

a larger and more expensive basket of goods than indicated by point *A*. Hence point *D*
is preferable to point *A*.

Derivation of Demand Curve

We have explained above Samuelson's *revealed preference axiom* and noted its assumptions.
By using his logic, we can now derive some of the results about consumer's behaviour.
We shall first examine the income and substitution effects and then derive the law of
demand, by using revealed preference theory.

Let us assume that the initial budget line of the consumer is given by *M N* in Fig.

1 1

6.22 and the consumer chooses a bundle of goods *X* and *Y* indicated by point *A*, (i.e., a
bundle of *AX* of *Y* and *OX* of *X*). Since all the bundles represented by the various points

1

1

on the line *M N* are equally expensive, by choosing bundle *A*, the consumer reveals his

1 1

preference for it. Let us now suppose that the price of *X* falls, price of *Y* remaining the
same, so that the budget line *M N* shifts to *M N* and the consumer shifts to the point

1 1

1 3

C. This shift results from the two effects of price change, viz., income and substitution
effects. Let us now decompose the income and substitution effects of the price effect by
using the Slutskian method. This has been done by drawing a budget line *M N* through

2 2

point *A*. Since the income adjusted budget line passes through point *A*, it implies that the
basket of *X* and *Y* indicated by point *A* is still available to the consumer. The consumer
will, therefore, not choose any point between *A* and *M* as they are inferior to the bundle

2

of goods indicated by point *A*. Hence he will buy either the bundle represented by point
A or by any other point, say *B*, on the *AN* segment of the budget line. Note that if he

2

continues to buy the basket at point *A*, substitution effect will be zero. And, if he chooses
point *B*, substitution effect is *X X* . Thus, demand for *X* increases as a result of decrease

1 2

in its price. A series of similar points can be traced by assuming subsequent decrease in the price of *X*, each point showing increase in demand for *X* due to fall in its price. This establishes the law of demand, i.e., as price of *X* falls, the quantity demanded of *X* increases. Having derived the law of demand, the Marshallian demand curve with a negative slope can easily be drawn.

Fig. 6.22 *Substitution and Income Effects: Revealed Preference Approach*

Appraisal of Revealed Preference Theory

Samuelson's revealed preference theory is a major contribution to the theory of demand. It has certain advantages over the Marshallian and Hicks-Allen approaches.

One, unlike the Marshallian demand theory and Hicks-Allen indifference curve analysis, it can be used to derive demand curves directly from consumer's revealed preferences without using the utility concept.

Two, in his approach to consumer analysis, Samuelson uses *behaviourist method* which is empirically observable in the market, whereas Marshallian *psychological method* or Hicksian *introspective method* are not empirically verifiable.

Three, revealed preference theory abandons most of the restrictive assumptions of the indifference curve analysis, e.g., the assumption of rationality or utility maximization and continuity, etc. It can be used to construct indifference curves under weaker assumptions.

Four, revealed preference theory provides also a basis for constructing the index number of cost of living.

The discussion on the revealed preference theory takes us to the end of individual demand analysis. In the next chapter, we will take up market demand analysis.

SUMMARY

• Chapter 6 presents a detailed analysis of consumer demand, including cardinal and ordinal approaches to analyze consumer behaviour.

• Conceptually, demand is the desire for a commodity for which the consumer has ability and willingness to pay.

• There are two concepts of demand for a commodity: (*i*) individual demand, and (*ii*) market demand. Individual demand means the desire of an individual consumer for a commodity for which he/she has ability to pay and willingness to pay for the commodity. Market demand for a commodity refers to the sum of individual demands for the commodity.

• Demand for a commodity arises because the use of the commodity gives *utility* to the consumer. Conceptually, *utility* means pleasure and satisfaction derived from consumption from a commodity.

• At theoretical level, there are two approaches to the measurability of utility: (*i*) cardinal utility approach, and (*ii*) ordinal utility approach. Accordingly, there are two approaches to the analysis of consumer behaviour.

• Cardinal utility approach was adopted by the classical and the neo-classical economists, especially Alfred Marshall, who assumed that utility is measurable cardinally, i.e., in cardinal numbers. It can be measured in terms of money one is willing to pay for a unit of a commodity under the condition that MU of money remains constant.

• Cardinal utility approach develops the law of diminishing marginal utility, i.e., the utility derived from marginal unit consumed goes on diminishing.

• According to cardinal utility approach, a consumer given his/her income and prices of goods he/she consumes tries to maximize utility. A consumer maximizes utility from a commodity at the level of consumption at which price of the good is equal to its marginal utility. The same rule applies to all the goods. This rule gives the principle of equi-marginal utility.

• Ordinal utility approach, known also as Hicksian approach, refutes the cardinal measurability of utility. This approach is based on the assumption that utility is measurable ordinally or comparatively only, i.e., in order of the importance of different goods, as first, second, third, etc.

• Ordinal utility approach uses 'indifference curve' (*IC* for short) as a tool of analysis. An *IC* shows different combinations of two substitute goods yielding the same level of utility. The *IC* has certain properties: (*i*) ICs are convex to origin, (*ii*) upper IC indicates a higher level of utility than the lower one, and (*iii*) *ICs* do not intersect. The slope of the *IC* gives the marginal rate of substitution (*MRS*), the rate at which one good can be substituted for the other.

• Consumer equilibrium is traced by introducing the budget line (*BL*) – a line that indicates various combinations of two goods that can be purchased given the consumer's income and prices of goods. The slope of the *BL* gives the price ratio. A consumer finds his/her equilibrium *MRS = price ratio*. This condition is satisfied at point of tangency between the *IC* and *BL*.

• *IC* technique is used to derive the consumer demand curve, the usual Marshallian demand curve.

• Given the law of demand, change in price causes change in demand, called price effect (*PE*). IC technique has the merit of being used to measure the income effect (*IE*) and substitution effect (*SE*) on change in demand due to change in price.

• Ordinal utility approach is treated to be a better tool of analysis than the ordinal utility approach. The economists have, however, pointed out the deficiencies of the ordinal utility approach to analyzing consumer behaviour.

REVIEW QUESTIONS

1. What is the relevance of consumer demand to business decision making? How does analysis of demand contribute to business decision-making?
2. What is meant by utility? Distinguish between cardinal and ordinal concepts of utility. How is utility measured under cardinal utility approach?
3. What is law of diminishing marginal utility? What are the conditions for the application of the law of diminishing utility? Explain the law with the help of *MU* schedule and *MU* curve.
4. What is meant by consumer equilibrium? Suppose total utility schedule of a consumer from his consumption of *X* is given as follows.

Units of *X* consumed

1

2

3

4

5

6

7

8

Total utility

20

35

45

52

57

60

60

50

Suppose also that price commodity *X* is `5 per unit. Find consumer's equilibrium.

5. What is meant by consumer equilibrium? What are the conditions of consumer's equilibrium under cardinal utility approach? Illustrate consumer's equilibrium diagrammatically.
6. Suppose marginal utility schedule of consumption of a commodity is given as follows.

Units consumes

1

2

3

4

5

6

MU

100

80

60

40

20

0

(*a*) Find the find Consumer's equilibrium if price of commodity is `60.

(*b*) Derive the demand curve for the commodity.

7. Prove that a consumer consuming two goods *X* and *Y* given their prices as P_x and P_y, respectively, is in equilibrium where

$$\frac{MU_x}{P_x} = \frac{MU_y}{P_y}$$

8. What is the law of equi-marginal utility? What is the condition for equi-marginal utility?

9. What is the law of demand? What is the basis of law of demand? Prove your point of view by applying cardinal utility approach to consumer analysis.

10. Why does demand curve slope down to the right? Under what conditions can a demand curve slope upward to right? Explain with examples?

11. What is meant by shift in demand curve? What are the factors that cause upward or downward shift in the demand curve?

12. Distinguish between cardinal and utility concepts of utility? What is the difference between cardinal and ordinal utility approaches to the analysis of consumer behaviour?

13. What is indifference curve? What are the properties of indifference curve? What is the difference between linear and curvy-linear indifference curves?

14. Why is an indifference curve for two normal goods convex to origin? Why it cannot be concave to origin or a straight line?

15. What is meant by marginal rate of substitution? Why does marginal rate of substitution decrease along the indifference curve?

16. Define budget line. How is basis of deriving the budget line? What purpose does budget line serve in consumer analysis?

17. What is meant the budget line? How is the budget line derived? Derive budget assuming a budget equation. How is the budget line affected by change in price of one good, all other factors remaining the same?

18. What are the conditions for consumer's equilibrium under ordinal utility approach to consumer analysis? Explain and illustrate graphically.

19. What is meant by consumer's equilibrium? Explain and illustrate consumer's equilibrium under the condition of measurability of utility and immeasurability of utility.

20. The equilibrium condition is satisfied also on the point of intersection between the budget line and indifference curve. Why does a consumer not stay at this point of equilibrium?

21. How is consumer equilibrium affected by the change in price given the consumers income and by the change in consumer's income, prices remaining constant?

22. What is meant by price effect, income effect and substitution effects of a change in price? What are the two methods of measuring income and substitution effects of a change in price?

23. The ultimate purpose of analysis of consumer behaviour is to derive consumer demand curve. How can indifference curve be used to derive consumer demand curve?

24. What is the revealed preference theory of consumer behaviour? How is the revealed preference theory treated to be superior to indifference curve approach?

25. Explain and illustrate the revealed preference theory of consumer behaviour. Can demand curve be derived by applying the revealed preference theory?

Exercises

26. Which of the following sentences are incorrect?

(*a*) Utility is not measurable.

(*b*) Consumer is in equilibrium where *MU* > price.

(*c*) Consumer surplus is maximum when *MU* = *P*.

(*d*) *MU* of money does not remain constant.

(*e*) Budget line can be curvy-linear.

(*f*) Consumer can be equilibrium on a lower *IC* curve if *MRS = price ratio.*

27. Suppose demand function for a product is given as *D* = 500 – 10 *P.* Find:

(*a*) Quantity demanded at pricè10.

(*b*) Price to sell 200 units.

(*c*) Price at which demand equals zero.

(*d*) Quantity demanded at zero price.

28. Distinguish between the following concepts.

(*a*) Cardinal utility and ordinal utility

(*b*) Income consumption curve and price consumption curve

(*c*) *MU* curve and demand curve

(*d*) Substitute goods and complementary goods

(*e*) Income effect and substitution effect

29. State whether following statements are 'right' and 'wrong'.

(*a*) Demand curve slopes'down to the right.

(*b*) Marginal utility increases in case of prestige goods.

(*c*) Income and substitution effects are greater than price effect.

(*d*) Consumer equilibrium takes place where *MRS* = price ratio.

(*e*) Revealed preference theory of consumer demand was developed by R.G.D. Allen.

(*f*) An indifference curve can be concave to origin.

30. Give reasons for the following statements.

(*a*) *MU* of consumer goods decreases with their consumption.

(*b*) *MRS* decreases because of diminishing marginal utility.

(*c*) Higher indifference curve yields a higher level of utility than the lower one.

(*d*) The law of equi-marginal utility implies that *MU* derived per rupee is the same.

FURTHER READING

Baumol, W. J., *Economic Theory and Operations Analysis* (Prentice Hall of India, New Delhi, 1985), Ch. 9.

Browning, E. K. and Browning, J. M., *Macroeconomic Theory and Application* (Kalyani Publisher, New Delhi, 1986), Chs. 2 & 3.

DwiveDi, D. N., *Microeconomics: Theory and Applications, 3rd Ed.* (Vikas Publishing House, Delhi, 2012), Chs. 7–9. 77.

Furguson, C.E., *Microeconomic Theory,* 3rd Ed. (Richard D. Irwin, Inc., Illinois, 1972), Chs. 1–4.

HirscHey, Mark, *Economics for Managers* (Thomson South-Western, 2006), Ch 4.

HirsHleiFer, J., *Price Theory and Applications* (Prentice Hall of India, New Delhi, 1985), Part II.

Koutsoyiannis, A., *Modern Microeconomics*, 2nd Ed. (Macmillan, London, 1979), Ch. 2.

marsHall, A., *Principles of Economics* (Macmillan & Co London, 1920). Book III, Ch IV and Book V, Chs. I & II.

PerloFF, J.M., *Microeconomics,* (Addison Wesley, New York, 2001), Ch. 4&5.

salvatore, D., *Microeconomics: Theory and Applications,* 4th Ed. (Oxford University Press, 2003). Chs. 3&4.

salvatore, D. (Adopted by Ravikesh Srivatsva, *Managerial Economics,* 7th Ed. (Oxford University Press, New Delhi, 2012), Ch. 3.

tHomas, C. R. And Maurice, S. C., *Managerial Economics: Concepts and Applications,* 8th Ed. (Tata McGraw-Hill, New Delhi, 2005), Ch. 5.

CHAPTER

7 Analysis of

Market Demand

CHAPTER OBJECTIVES

The objective of this chapter is to provide an analysis of market demand. By going through this chapter, you will know:

• The meaning and derivation of market demand

• Determinants of market demand for different kinds of consumer goods

• Meaning and formation of consumer demand function

• Kinds of consumer demand functions

• Application of demand functions

7.1 INTRODUCTION

From the analysis of individual demand, we move on in this chapter to analyze the market demand for a product. The analysis of market demand for a firm's product plays a crucial role in business decision-making. The market demand or the size of the market at a point in time at different prices gives the overall scope of business; it gives prospects for expanding business; and it plays a crucial role in planning for future production, inventories of raw materials, advertisement, and setting up sales outlets. Therefore, the information regarding the magnitude of the current and future demand for the product is indispensable. Theory of demand provides an insight helps in analysing these problems.

In this chapter, we discuss the meaning of market demand, types of market demand and their distinctive features, and determinants of market demand for a product.

7.2 DEFINITION OF MARKET DEMAND

Meaning of Market Demand

Market demand can be defined as the sum of individual demands for a product at a price per unit of time. We may recall that the quantity demanded of a commodity by an individual per unit of time, at a given price, is known as 'individual demand' for that commodity. *The aggregate of individual demands for a product is called market demand for the product*. In other words, the sum of quantity demand by all the consumers/users of a commodity per unit of time at a given price, all other things remaining the same, is called 'market demand' for that product.

For example, suppose there are only three consumers— *A, B* and *C*—of a commodity *X* which has a fixed price. Consumers *A, B* and *C* consume 100 units, 200 units and 300 units respectively, of commodity *X* monthly. Therefore, the monthly *market demand* for commodity *X* equals 100 + 200 + 300 = 600 units.

If individual demand schedules or individual demand functions are known, the market demand schedule and curve can be obtained by (*i*) adding up individual demand at different prices, and (*ii*) summing up individual demand functions. The derivation of market demand curves by the two methods are illustrated below.

7.3 DERIVATION OF MARKET DEMAND CURVE

If individual demand schedules or individual demand functions are known, the market de mand schedule and market demand curve can easily be derived. The market demand curve can be derived by adding up (*i*) the individual demand schedules, and (*ii*) the individual demand functions. In this section, we illustrate the derivation of market demand curve by using these two methods.

Derivation of Market Demand Curve from Individual Demand Schedules

Suppose again that there are three consumers *(A, B* and *C*) of a commodity *X* and their monthly demand schedules for the commodity are given in Table 7.1. The table shows the quantity demanded of commodity *X* individually by the three consumers at different prices of commodity *X*. The last column shows the market demand, i.e., the sum of individual demands for commodity *X*. The market demand shows the total quantity of commodity *X* demanded per month by the three consumers at different prices.

Given the individual and market demand schedules, the market demand curve can be obtained by plotting the market demand against the respective prices. This is illustrated in Fig. 7.1 by the curve *D* .

M

Alternatively, market demand curve can be drawn by horizontal summation of the individual demand curves. In Fig. 7.l, curves marked *D* , *D* and *D* show the individual demand curves

A

B

C

based on demand schedule given in Table 7.1. The derivation of market demand curve by horizontal summation of individual demand curves is illustrated by the demand curve *D* .

M

Table 7.1 *Price of Commodity X and*

Quantity Demanded

Price of X

Quantity of X Market Demand

(`)

Demanded by

=(A+B+C)

A

B

C

25

0

0

0

0

20

5

0

0

5

15

10

5

0

15

10

15 10 5

30

5

20 15 10

45

0

25 20 15

60

Fig. 7.1 *Derivation of Market Demand Curve*

Derivation of Market Demand Curve from Individual Demand Functions

Market demand curve can also be drawn by summing up the individual demand functions. Suppose individual demand functions of consumers *A, B* and *C* for commodity *X* are given as follows.

A's demand function: $D_x^A = 100 – 10 P$

B's demand function: $D_x^B = 75 – 7.5 P$

C's demand function: $D_x^C = 50 – 5 P$

Given the individual demand functions, the market demand function can be obtained by adding them up. Thus,

Market Demand function: $D_x^M = (100 – 10 P_x) + (75 – 7.5 P_x) + (50 – 5 P_x)$

$D_x^M = 225 – 22.5 P$

This market demand function can be converted into a market demand schedule by assigning numerical values to P_x. The market demand curve can then be drawn by plotting the demand schedule.

7.4 KINDS OF DEMAND

The managers are supposed to be clear about the kind of demand they are dealing with. The demand for various goods and services are generally classified on the basis of the consumers of the product, suppliers of the product, nature of the product, seasonal nature of the demand, interdependence of demand for two products, etc. Here we discuss some major kinds of demands that figure in business decisions.

(i) Individual Demand and Market Demand. As noted above, ***individual demand*** refers to the quantity of product demanded by an individual at a point in time or over a period of time given the price of the product, given his income, price of the related goods (substitutes and complements), consumer’s taste and preferences, price expectations, and external influences (e.g., bandwagon and demonstration effects).

As explained above, ***market demand*** refers to the quantity that 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). 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 details, see also the previous section).

(ii) Demand for Firm’s Product and Industry’s Product. The quantity that a firm disposes 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 demands is not of much use in a *highly competitive market*, e.g., the fruit and vegetable markets. In these markets, each seller has an insignificant share in the market. Therefore, demand for firm’s product is not of significance.

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 decision point

of view. For, in such markets, product of each firm is so differentiated from the product 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. For examples, there was price competition1 between MUL and other car companies in 2005.

In 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 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 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.

(iii) 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 of the people 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 arises directly from the biological or physical needs of human beings, e.g., demand for food, clothes, shelter, etc. Demand for these goods and the like is *autonomous demand*. Autonomous demand may also arise as a result of ‘demonstration effect’, 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 of demand for food. Similarly, demand for steel, bricks, cement etc. is a derived demand – derived from the demand for housing and commercial 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 businessman's point of view to the extent that the former can serve as an indicator of the latter.

(iv) 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

1. MUL cut down the price of its M 800 model (both standard and air-conditioned) by `16,286 in June 2005 because it had lost its 30% market to other common brands. Maruti sales had fallen by 38% in the first two months of fiscal year 2005-06.

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, refrigerators, scooters, cars, cell phones, etc. The durable producer goods include mainly items under 'fixed assets', such as buildings, plants, machinery, office furniture and fixtures etc. The durable goods, both consumer and producer durable goods, may be further classified as 'semi-durables' (e.g., clothes and furniture) and 'durables' (e.g., residential and factory buildings, 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. The goods of 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 category, fall such goods as raw materials, fuel and power, finishing materials and packing items, etc.

The demand for non-durable goods depends largely on their current prices, consumers' income and fashion and is subject to frequent changes whereas the demand for durable goods is influenced also 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 non-durable goods do not. Also, the demand for non-durable 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.*

(v) Short-term and Long-term Demand. From demand-analysis point of view, demand is classified also under short-term demand and long-term demand. Short-term demand refers to the demand for goods that are required over a short period of time. In this category are found mostly 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 a short-term demand though the demand for generic goods (trousers, shoes, ties, etc.) continues to remain a long-term demand. Similarly demand for umbrellas, raincoats, gum-boots, cold-drinks, ice creams etc., is short-term demand of 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, etc. 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.

The ***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 non-durable 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 consumer, their ability to adjust their consumption pattern and their susceptibility to advertisement of a new product. The long-term demand depends, by and large, 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 new entrepreneurs, in pricing policy, and in determining and phasing the advertisement expenditure.

7.5 DETERMINANTS OF MARKET DEMAND

Market demand for a product depends on a number of factors, called **determinats of demand**. The knowledge of the determinants of market demand for a product and the nature of relationship between the demand and its determinants proves very helpful in analyzing and estimating demand for the product. It may be noted at the very outset that a host of factors determine the market demand for a product.

In general, however, following are the factors that determine the market demand for a product:

1. Price of the product,
2. Price of the related goods—substitutes, complements and supplements,
3. Level of consumers' income,
4. Consumers' taste and preferences,
5. Advertisement of the product,
6. Consumers' expectations about future price and supply position,
7. Demonstration effect and 'bandwagon effect',
8. Consumer-credit facility,
9. Population of the country (for the goods of mass consumption),
10. Distribution pattern of national income, etc.

To this list, one may add such factors as off-season discounts and gifts, number of uses of a commodity, level of taxation and the general social and political environment of the country (especially with respect to demand for capital goods).

All these factors are, however, not equally important. Besides, some of them are not even quantifiable. For example, consumer's preferences, utility, demonstration effect, expectations, etc., are difficult to measure. Nevertheless, we will discuss here both quantifiable and non-quantifiable determinants of the demand for a product.

1. *Price of the Product* The price of a product is one of the most important determinants of its demand in the long-run and the only determinant in the short-run. The price of a product and its quantity demanded are inversely related. The law of demand states that the quantity demanded of a product which its consumers/users would like to buy per unit of time, increases when its price falls and decreases when its price increases, *other factors remaining constant*. (For details, review chapter 6.). The assumption 'other factors remaining constant' implies that factors other than price remain constant, particularly income of the consumers, prices of the substitutes and complementary goods, consumer's

taste and preferences, and number of consumers, remain unchanged. Other factors remaining constnat, price is the main determinant of market demand especially in short run.

2. *Price of the Related Goods* The demand for a commodity is also affected by the changes in the prices of its related goods. Related goods are classified under two categories: (i) ***substitutes,*** and (ii) ***complementary goods***.

(i) Substitutes. Two commodities are deemed to be *substitutes* for one another if they satisfy the same want and change in the price of one changes the demand for its substitute in the same direction. That is, increase in the price of a good increases demand for its substitutes. For instance, tea and coffee, hamburgers and hot-dogs, petrol and CNG, alcohol and drugs are some common examples of substitutes in the case of consumer goods. What is important from managerial point of view is that increase in the price of a good (say, *X*) causes increase in demand for its substitute (say *Y*).

The demand function for *X* and *Y* with respect to the price of their substitutes can be written as follows.

$D = f(P)$, $\Delta D / \Delta P > 0$ and $D = f(P)$, $\Delta D / \Delta P > 0$

x

y

x

y

y

x

y

x

Fig. 7.2 *Demand for Substitutes and Complements*

When price of a substitute good (say, coffee) of a product (tea) falls (or increases), the demand for the product falls (or increases). The demand-price relationship of this nature is given in Fig. 7.2(*a*).

(ii) Complements. A commodity is considered to be a *complement* for another when it complements the use of the other. In case of complements, the use of the two goods goes together. For example, petrol is a complement to cars, butter and jam to bread, milk and sugar to tea and coffee, electricity to computer, chair to table, etc. In economic sense, two goods are termed as complementary to one another if an increase in the price of one causes a decrease in demand for the other. By definition, there is an inverse relation 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 cars and other petrol-run vehicles, other things remaining the same. The demand function for car (*D*) in relation to petrol price (*P*) can be written as

c

p

$D = f(P)$, D D /D $P < 0$

c

p

c

p

The relationship between the demand for a product (car) and the price of its complement (petrol) is given in Fig. 7.2(*b*).

3. *Consumer's Income* Income is the basic determinant of quantity of a product demanded as it determines the purchasing power of the consumer. That is why people with higher current disposable incomes spend a larger amount on consumer goods and services than those with lower income. Income-demand relationship is of a more varied nature than that between demand and its other determinants. While other determinants of demand, e.g., product's own price and the price of its substitutes are more significant in the short-run, income as a determinant of demand is equally important in both short-run and long-run.

The relationship between the demand for a commodity, say *X*, and the household income (*Y*), assuming all other factors to remain constant, is expressed by a demand function such as

$D = f(Y)$, D D /D $Y > 0$

x

x

Before we proceed to discuss income-demand relationships, it will be useful to note that consumer goods of different nature have different relationships with incomes of different categories of consumers. The managers need, therefore, to be fully aware of the goods they are dealing with and their relationship with the income of consumers, particularly in regard to the assessment of both existing and prospective demand for a product.

For the purpose of income-demand analysis, consumer goods and services may be grouped under four broad categories, viz. (*a*) essential consumer goods, (*b*) inferior goods, (*c*) normal goods, and (*d*) luxury or prestige goods. Let us now look into the relationship between income and the different goods. This relationship is presented through Engel curves.[2]

(*a*) *Essential consumer goods* *(ECG)*. The goods and services in this category are called 'basic needs' and are consumed by all persons of a society, e.g., foodgrains, salt, vegetable oils, matches, cooking fuel, minimum clothing and housing. Quantity demanded of this category of goods increases with increase in consumer's income but only upto a certain limit, even though the total expenditure may increase in accordance with the quality of goods consumed, other factors remaining the same. The relationship between goods of this category and consumer's income is shown by the curve *ECG* in Fig. 7.3. As the curve *EGC* shows, a consumer's demand for essential goods increases only until his income rises to OY_2. It tends to saturate beyond this level of income.

(*b*) *Inferior goods* (*IG*). Inferior and superior goods are widely known to both

Fig. 7.3 *Income Demand Curves*

consumers and sellers. For instance, every

2. Named after 19th century German Statistician, Christian Lorenz Ernst Engel.

consumer knows that millet is inferior to wheat and rice; *bidi* (indigenous cigarette) is inferior to cigarette, cotton clothes are inferior to silk clothes, kerosene is inferior to cooking gas; bike is inferior to car; non-AC car is inferior to AC-car, and so on and so forth. In economic sense, however, *a commodity is deemed to be inferior if its demand decreases with the increase in consumer's income beyond a certain level of income*. The nature of relation between income and demand for an inferior good is shown by the curve *IG* in Fig. 7.3 under the assumption that other determinants of demand remain the same. Demand for such goods rises only up to a certain level of income (say, OY_1) and declines as income increases beyond this level.

(*c*) *Normal goods* (*NG*). Technically, normal goods are those that are demanded in increasing quantities as consumers' income rises. Clothing, house, furniture, and automobiles are some of the important examples of this category of goods. The nature of relation between income and demand for the goods of this category is shown by the curve *NG* in Fig. 7.3. 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 increases rapidly with the increase in the consumer's income but slows down with further increases in income. It may be noted from Fig. 7.3 that up to a certain level of income (Y_1) the relation between income and demand for all types of goods is similar. The difference is only of degree. The relation becomes distinctly different beyond the Y_1 level of income. From a managerial point of view, therefore, it is important to view the income-demand relations in the light of the nature of product and the level of consumers' income.

(*d*) *Luxury and prestige goods* (*LG*). What is and what is not a luxury good is a matter of consumer's perception of the need for a commodity. Conceptually, however, all such goods that add to the pleasure and prestige of the consumer without enhancing his earning capacity or efficiency fall in the category of luxury goods. For example, stone-studded jewellery, costly brands of cosmetics, luxury cars, accommodation in 5-star hotels, travel by first-class railway *AC* cars, business class air travel, etc., can be treated as luxury goods. A special category of luxury goods is that of prestige goods, e.g., precious stones, ostentatious decoration of buildings, rare paintings and antiques, diamond-studded jewellery and watches, prestigious schools, etc. Demand for such goods arises beyond a certain level of consumer's income, i.e., consumption of luxury goods at a certain level high level of income. Producers of such items, while assessing the demand for their product, should consider the income change in the richer section of the society, and not merely the per capita income (see curve *LG* in Fig. 7.3).

4. ***Consumer's Taste and Preference*** Consumer's taste and preference play an important role in determining the demand for a product. Taste and preference generally depend on life-style, social customs, religious values attached to a commodity, habit of the people, the general levels of living of the society, and age and sex of the consumers. Change in these factors changes consumers' taste and preferences. As a result, consumers reduce or give up the consumption of some goods and add new ones to their consumption pattern. For example, following the change in fashion, people switch their consumption pattern from cheaper, old-fashioned goods over to costlier, modern goods, so long as price differentials are commensurate with their preferences. Consumers are prepared to pay higher prices for 'modern goods' even if their utility is virtually the same as that of old-fashioned goods, e.g., new fashion suits to old design suits, flat TV sets to box TV sets, and modern shoes to old design shoes, etc.

This piece of information is useful for the manufacturers of goods and services subject to frequent changes in fashion and style, at least in two ways: (*i*) they can make quick profits by designing new models of their product and popularising them through advertisement, and (*ii*) they can plan production better and can even avoid over-production if they keep an eye on the changing fashions.

5. ***Advertisement Expenditure*** Products are advertised with the objective of promoting sales of the product. Advertisement helps in increasing demand for the product in at least four ways: (*a*) by informing the potential consumers about the availability of the product; (*b*) by showing its superiority over the rival product; (*c*) by influencing consumers' choice against the rival products; and (*d*) by setting new fashions and changing tastes. The impact of such effects shifts the demand upward to the right.

In other words, other factors remaining the same, as expenditure on advertisement increases, volume of sales increases to an extent. The relationship between sales (*S*) and advertisement outlays (*AD*) is expressed by the function $S = f(AD)$. The relation between advertisement outlays and sales is shown in Fig. 7.4.

Assumptions. The relationship between demand and advertisement cost shown in Fig. 7.4 is based

Fig. 7.4 *Advertisement and Sale*

on the following assumptions:

(*a*) Consumers are fairly sensitive and responsive to various modes of advertisement,

(*b*) The rival firms do not react to the advertisements made by a firm,

(*c*) The level of demand has not already reached the saturation point. Once demand reaches the saturation point, advertisement makes only a marginal impact on demand,

(*d*) Advertisement cost added to the price does not make the price prohibitive for consumers, compared to the price of substitutes,

(*e*) Other determinants of demand, e.g., income and tastes, etc. are not operating in the reverse direction.

In the absence of these conditions, the effect of advertisement on sales may be unpredictable.

6. ***Consumers' Expectations*** Consumers' expectations regarding the future prices, income, and supply position of goods, etc. play an important role in determining the demand for goods and services in the short-run. If consumers expect a high rise in the price of a storable commodity, they would buy more of it at its high current price with a view to avoiding the pinch of a high price rise in future. 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 goods whose prices are expected to decrease in the future.

Similarly, an expected increase in income increases demand. For example, announcement of 'dearness allowance', bonus, revision of pay-scale, etc., induces increase in current purchases. Besides, if scarcity of certain goods is expected by the consumers/users on account of a reported fall in future production, strikes on a large scale, diversion of civil supplies towards military use, etc., the current demand for such goods tends to increase,

more so if their prices show an upward trend. Consumers demand more for future consumption and profiteers demand more to make money out of an expected scarcity.

7. *Demonstration and Snob Effect* When new commodities or new models of existing ones appear in the market, rich people buy them first. For instance, when a new model of a car appears in the market, rich people would mostly be the first buyers. Colour TV sets and VCRs were first seen in affluent households. Some people buy goods or new models of goods because they have a genuine need for them or have excess purchasing power. Some others do so because they want to exhibit their affluence. According to a social philospher, Will Smith, “Too many people spend money they haven’t earned, to buy things they don’t need, to impress people they don’t like.”[3]. But once new commodities are in vogue, many households buy them not because they have a genuine need for them but because their neighbours have bought these goods. The purchases made by the latter category of the buyers arise out of such feelings as jealousy, competition and 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 what economists call ‘Demonstration effect’ or the ‘Bandwagon effect’. These factors have a positive effect on demand. On the contrary, when a commodity becomes the thing of common use, some people, mostly 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.[4]

8. *Consumer-Credit Facility* Availability of credit to the consumers from the sellers,[5] banks, relations and friends, or from other sources, enduces the consumers to buy more than what they would buy in the absence of credit facility. That is why consumers who can borrow more can consume more than those who cannot borrow. Credit facility mostly affects the demand for durable goods, particularly those which require bulk payment at the time of purchase. The car-loan facility may be one reason why Delhi has more cars than Calcutta, Chennai and Mumbai all put together. Realty business boomed in Delhi, NOIDA and Gurgaon mainly because of housing loans made available by the banks. The managers who are assessing the prospective demand for their products should, therefore, take into account the availability of credit to the consumers.

9. *Population of the Country* The total domestic demand for a product of mass consumption depends also on the size of the population. Given the price, per capita income, tastes and preferences etc., the larger the population, the larger the demand for a product. With an increase (or decrease) in the size of population and with the employment percentage remaining the same, demand for the product tends to increase (or decrease). The global perception that India offers the largest market in the world is based on the fact that she has the second largest population – albeit with a low purchasing power – in the world.

10. *Distribution of National Income* The level of national income is the basic determinant of the market demand for a product—the higher the national income, the higher the demand for all normal goods and services. Apart from its level, the distribution pattern of national income is also an important determinant of the overall demand for a product. If national income is unevenly distributed, i.e., if a majority of the population belongs to

3. Quoted in *Times of India*, on 21/3/2010.

4. For details see, Harvey Leibenstien, “Bandwangon, Snob, and Veblen Effect in the Theory of Consumers’ Demand”, *Qly. Jl. of Eco.*, 65 (May 1950).

5. Sellers provide ‘buy now pay later’ facility.

the lower income groups, market demand for essential goods, including inferior ones, will be the largest whereas the demand for other goods will be relatively lower.

7.6 DEMAND FUNCTION

The demand function and different kinds of demand functions have already been introduced in Chapter 4, of couse, in a different context. Here, we repeat the exercise in the context of market demand analysis. Here the term ‘demand function’ has been used in the sense of *market demand function*.

As noted earlier (Ch. 4), a function is a symbolic statement of a relationship between the dependent and the independent variables. Demand function states the relationship between the demand for a product (the dependent variable) and its determinants (the independent variables). Let us consider a very simple case of market demand function. Suppose all the determinants of the aggregate demand for commodity *X*, other than its price, remain constant. This is a case of a *short-run demand function*. In the case of a short-run demand function, quantity demanded of *X*, (*D*) depends on its price (*P*). The

x

x

market demand function can then be symbolically written as

D = *f* (*P*)

...(7.1)

x

x

The function (7.1) reads 'demand for commodity *X* (i.e., *D*) is the function of its price

x

(*P*)'. In this function, *D* is a dependent and *P* is an independent variable. It implies that

x

x

x

a change in *P* (the independent variable) causes a change in *D* (the dependent variable).

x

x

The function (7.1) however does not reveal the change in *D* for a given percentage

x

change in *P* , i.e., it does not give the quantitative relationship between *D* and *P* . When

x

x

x

the quantitative relationship between *D* and *P* is known, the demand function may be

x

x

expressed in the form of an equation. For example, a linear demand function is written as

D = *a* – *bP*

...(7.2)

x

x

where ' *a*' is a constant, denoting total demand at zero price, and *b* = Δ *D*/Δ *P*,6 is also a constant—it specifies the change in *D* in response to a change in *P* .

x

x

Table 7.2 *Demand Schedule*

P

D = 100 – 5 *P*

D

x

x

x

x

0

D = 100 – 5 × 0

100

x

5

D = 100 – 5 × 5

75

x

10

D = 100 – 5 × 10

50
x
15
$D = 100 – 5 \times 15$
25
x
20
$D = 100 – 5 \times 20$
0
x

Fig. 7.5 *Linear Demand Function*

6. Δ *D*/Δ *P* is, in fact, reciprocal of the slope of a linear demand curve.

The form of a demand function depends on the nature of demand-price relationship. The two most common forms of demand functions are *linear* and *non-linear* demand function. Here we briefly discuss the linear and non-linear forms of demand functions.

7.6.1 Linear Demand Function

A demand function is said to be linear when Δ *D*/Δ *P* is constant and the function it results in is a linear demand curve7. Eq. (7.2) represents a linear form of the demand function. Assuming that in an estimated demand function $a = 100$ and $b = 5$, demand function Eq. (7.2) can be written as

$D = 100 – 5 P$

...(7.3)

x
x

By substituting numerical values for *P* , a demand schedule may be prepared as given in
x
Table 7.2.

This demand schedule when plotted, gives a linear demand curve as shown in Fig. 7.5. As can be seen in Table 7.2, each change in price, i.e., Δ *P* = 5 and each
x
corresponding change in quantity demanded, i.e., Δ *D* = 25. Therefore, Δ *D* /Δ *P* = *b* = 25/5
x
x
x
= 5 throughout. That is why demand function Eq. (7.3) produces a linear demand curve.

Price Function

From the demand function, one can easily derive the price function. For example, given the demand function Eq. (7.2), the price function may be written as follows.

$P = a - Dx$
x
b
$P = a$ 1
$- D$
x
b b x

Assuming *a*/ *b* = *a* and 1/ *b* = *b* , the price
1
1
function may be written as

$P = a – b D$

...(7.4)

x
1
1
x

7.6.2 Non-linear Demand Function

A demand function is said to be non-linear or curvilinear when the slope of the demand curve, ($\Delta P/\Delta D$) changes all along the curve. A non-linear demand function yields a demand curve instead of a demand line, as shown in Fig. 7.6. A non-linear demand function takes the form of a power function of the form given below.

$D = aP – b$ …(7.5)

x

x

Fig. 7.6 *Non-linear Demand Function*

7. For example, Phillip L. Paalberg and Robert Thompson have estimated the US demand function for wheat as follows. *Qw* = 87.9041—0.0924 *Pw* where *Qw* = demand for wheat and *Pw* is wheat price (See their paper "Interrelated Products and the Effects of an Import Tariffs" in *Agri*. *Eco*. *Res*., 32, (October 1980). Quoted in J.R. Davis and Seemon Chang. *Principles Managerial Economics*, Prentice-Hall, 1968, p. 13.

a

and

$D =$

$- b$

…(7.6)

x

x

$P + c$

where $a > 0$, $b > 0$ and $c > 0$.

7.6.3 Multi-variate or Dynamic Demand Function: Long-Term Demand Function

We have discussed above a single variable demand function, i.e., price as a single independent variable. This may be termed as a short-term demand function. In the long run, however, neither the individual nor the market demand for a product is determined by any one of its determinants because other determinants do not remain constant. The long-run demand for a product depends on the composite impact of all its determinants operating simultaneously. Therefore, for the purpose of estimating long-term demand for a product, all its relevant determinants are included in the demand function. They are then expressed in a functional form. The function describes the relationship between the demand (a dependent variable) and its determinants (the independent or explanatory variables). A demand function of this kind is called a *multi-variate* or *dynamic demand function*. For instance, consider this statement: the demand (*D*) for a commodity *X*,

x

depends on its price (*P*), consumer's money income *M*, price of its substitute *Y*, (*P*),

x

y

price of complementary goods (*P*) and consumer's taste (*T*) and advertisement expenditure

c

(*A*). This statement can be expressed in a functional form as,

$D = f (P , M, P , P , T, A)$

...(7.7)

x

x

y

c

The demand function (7.7) describes the demand for commodity *X* which depends on such determinants as *P* , *M*, *P* , *P* , *T* and *A*. If the relationship between *D* and the

x

y

c

x

quantifiable independent variables, *P* , *M*, *P* , *P* and *A* is of linear form, the estimable

x

y

c

form of the demand function is expressed as

$D = a - bP + cM + dP - gP + jA$

...(7.8)

x

x

y

c

where ‘ *a*’ is a constant term and constants *b*, *c*, *d*, *e*, *g* and *j* are the coefficients of relation between *D* and the respective independent variables8.

x

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.

7.7 CONCLUSION

While Chapter presents the analysis of consumer behaviour and individual consumer demand, this chapter presents the analysis of market demand. The discussion here includes explanation of the concept and derivation of market demand, determinants of market demand and the nature of relationship between the demand determinants and market demand, and different forms of demand function. These aspects are important for business managers to understand how market size is determined. Another and a rather more important aspect from pricing-decision point of view is how market demand reacts to change in price of a product. This takes us to

8. Note that all coefficients in emperically estimated market demand functions, *bP* and *gP* bear a minus sign

x

c

(–) because of inverse relationship between these variables and *D* . *x*

another aspect of market demand, i.e., the elasticity of demand. The elasticity of demand and its application in business decision-making has been discussed in the next Chapter.

SUMMARY

• Chapter 7 deals with various aspects of market demand, including meaning of market demand, derivation of market demand curve, kinds of market demand, determinants of market demand, and market demand function.

• Market demand refers to the sum of individual demands for a product at a given price at a point of time, all other factors given.

• Market demand curve for a product is derived by adding up individual demands at different levels of prices. Or, given the individual demand curves, market demand curve can be derived by summing up individual demand curves for the product.

• From analytical point of view, demand for a product is classified as (*i*) individual demand and market demand, (*ii*) demand for firm’s product and industry product, (*iii*) Autonomous demand and derived demand, (iv) demand for durable and non-durable goods, and (*v*) short-term demand and long-term demand.

• Determinants of market demand for product are (*i*) price of the product, (*ii*) price of the substitute goods, (*iii*) price of the complementary goods, (*iv*) consumers’ income, (*v*) taste and fashion, (*vi*) advertisement, (*vii*) expectation about future price, (*viii*) demonstration effect and bandwagon effects, (*ix*) credit facility, (*x*) population of the relevant market segment.

• Demand function is mathematical statement showing the nature and extent of relationship between demand for a product and demand determinants. Demand function may be short-run function or a long-run function.

• A short-run demand function is price related demand function, expressed as $Q = f(P)$. The factual or empirical demand function is formulated by estimating the relationship between different prices of a product and corresponding demand. The general form of a linear short-run demand function takes the form as $Q = a - bP$.

• A long-term demand function includes all the determinants of the demand for a product. The general form of long-run demand function for a consumer good, say *X,* is expressed as $Q = f(P, P, P, Y, T/F, \ldots)$.

x

x

s

c

REVIEW QUESTIONS AND EXERCISES

1. Define market demand. How is the analysis of market demand for a product important for the manager of the firm producing the product?

2. How is market demand for a commodity derived? Derive market demand curve by assuming three individual demand curves for a product.

3. How is market demand for a product classified with the purpose of business decision making? Distinguish between (*a*) autonomous demand and derived demand, (*b*) individual demand and market demand, and (*c*) short-term demand and long-term demand for consumer goods.

4. What are the determinants of demand for a product? How do the changes in the following factors affect the demand for a product?

(*a*) Price of the product

(*b*) Income of consumers

(*c*) Price of the substitute

(*d*) Advertisement.

5. Distinguish between demand schedule and demand function. Suppose demand function for commodity *X* is given as *D* = *a* – *bP* . Derive the demand schedule and demand curve.

x

x

6. Distinguish between demand schedule and demand function. Suppose demand function for commodity *X* is given as *D* = *a* – *b P* . Derive the demand schedule and demand curve.

x

x

7. What is difference between linear demand function and non-linear demand function? Illustrate and explain the properties of linear and non-linear demand functions.

8. Define market demand. Suppose there are three consumers of commodity *X* and their respective demand schedules are given as follows.

Price of *X* (`)	Quantity of *X* demanded by		
	A	B	C
10	0	0	0
8	4	2	0
6	8	4	0

Find the market demand schedule and market demand curve for commodity *X.*

9. Suppose monthly demand function of Ruchi, Neha and Sanchi for Pepsi are given as follows.

Ruchi : *D* = 50 – 2 *P*

R

Neha : *D* = 60 – 3 *P*

N

Sanchi *: D = 20 – P*

S

Find the aggregate demand function and total weekly consumption of Pepsi if price of Pepsi is fixed at `10 per bottle.

10. Specify the nature of the following demand functions.

(*i*) Q = 50 – 5 P (*ii*) Q = 100 – $P2$

d

d

(*iii*) $Qd = aP – b$

(*iv*) Q = ($a/p + c$) – b

d

11. A publishing company plans to publish a new book. It collects sales data from other publishers of similar book. By using the data, it finds a demand function given as Q = 5000 – 5 P. Find out:

(*a*) Demand schedule and demand curve,

(*b*) Number of books saleable at a price of `20 per book

(*c*) Price for selling 2500 copies of the book.

FURTHER READING

Baumol, W. J., *Economic Theory and Operations Analysis* (Prentice Hall of India, New Delhi, 1985), Ch. 9.

Browning, E. K. and Browning, J. M., *Macroeconomic Theory and Application* (Kalyani Publisher, New Delhi, 1986), Chs. 2 & 3.

DwiveDi, D. N., *Microeconomics: Theory and Applications, 2nd Ed*. (Vikas Publishing House, Delhi, 2012), Chs. 7 – 9. 77.

Furguson, C. E., *Microeconomic Theory,* 3rd Ed. (Richard D. Irwin, Inc., Illinois, 1972), Chs. 1–4.

HirscHey, Mark, *Economics for Managers* (Thomson South-Western, 2006), Ch 4.

HirsHleiFer, J., *Price Theory and Applications* (Prentice Hall of India, New Delhi, 1985), Part II.

Koutsoyiannis, A., *Modern Microeconomics*, 2nd Ed. (Macmillan, London, 1979), Ch. 2.

marsHall, A., *Principles of Economics* (Macmillan & Co London, 1920), Book III, Ch. IV and Book V, Chs. I & II.

PerloFF, J. M., *Microeconomics,* (Addison Wesley, New York, 2001), Ch. 4&5.

salvatore, D., *Microeconomics: Theory and Applications,* 4th Ed. (Oxford University Press, 2003). Chs. 3 & 4.

salvatore, D. (Adopted by Ravikesh Srivatsva, *Managerial Economics*, 7th Ed. (Oxford University Press, New Delhi, 2012), Ch. 3.

tHomas, C. R. and Maurice, S. C., *Managerial Economics. Concepts and Applications,* 8th Ed. (Tata McGraw-Hill, New Delhi, 2005), Ch. 5.

CHAPTER

8 Elasticity of Demand

CHAPTER OBJECTIVES

The objective of this chapter is to introduce the concept of demand elasticity and the methods of its measurement. This chapter gives you a clear understanding of

• Meaning and kinds of demand elasticity

• The concept and measurement of price elasticity of demand

• How price elasticity affects total revenue and marginal revenue

• Concept and measurement of income elasticity, cross elasticity, ad-elasticity and price expectation elasticity

• Application of demand elasticity in pricing decisions

8.1 INTRODUCTION

The theory of demand, discussed in two preceding chapters, states the *direction of change in demand* due to change in its determinants. For example, according to the theory of demand, all other factors remaining constant, when price of a product increases, demand for the product decreases and *vice versa*. The theory of demand does not tell ‘demand decreases by how much due to a certain increase in price’. However, the knowledge of direction or kind of change in demand due to change in its determinant is not sufficient from the angle of price management. What is more important is the *extent of relationship* between the demand for a product and its determinants. For example, suppose a firm reduces the price of its product from `10 tò8 with the objective of increasing demand for its product and its total revenue. In effect, however, total revenue decreases from `1000 tò880. The reason is that demand increased by a lower percentage (10%) than the percentage of decrease in price (20%).

Look at another case of pricing decision issue. When price of the substitute of a firm’s product increases, the question arises: Should the firm change its own price or keep the price unchanged? Similarly, if a firm plans to advertise its product, a question arises: Will the sales promotion yield sales revenue in excess of advertisement cost? The theory of demand, as such, does not provide answer to these questions. Answer to "these questions lies in the *degree of responsiveness* of demand to a change in its determinants. The degree of responsiveness of demand for product to change in its price is called the ***elasticity of demand.*** This chapter presents a detailed discussion on the concept of the elasticity of demand, the various methods of measuring demand, and its application to business decisions. The

concepts of elasticity of demand generally used in business decisions are:

(*i*) Price elasticity of demand,

(*ii*) Cross elasticity of demand,

(*iii*) Income elasticity of demand,

(*iv*) Advertisement elasticity of demand, and

(*v*) Elasticity of price expectations.

Before we discuss these elasticities of demand, let us understand the concept of demand elasticity and the method of its measurement.

8.2 DEFINITION AND MEASUREMENT OF ELASTICITY CONCEPT

In general terms, the *elasticity of demand* is defined as *the degree of responsiveness* of demand for a product to change in its determinants. The measure of the degree of responsiveness of demand to change in its determinants gives the measure of the *extent of relationship* between the demand for a product and any of its determinants. In technical terms, the measure of elasticity of demand is called *elasticity coefficient* measured by the following formula:

E

X

d = Percentage Change in Quantity Demanded of Product

Percentage Change in Demand Determinant Factor *Y*

For instance, suppose a determinant of demand for a product changes by 10 percent and, as a result, demand changes by 15 percent. In that case, the elasticity coefficient equals 15/10 = 1.5.

The general formula for measuring the elasticity of demand can be expressed as follows.

Δ /

Q Q Δ *Q Y*

E

=

×

d = Δ *Y*/ *Y*

Δ *F Q*

…(8.1)

where *Q* = quantity demanded initially; Δ *Q* = change in demand, *Y* = the original value of demand determinant factor; and Δ *Y* = change in the determinant factor.

Let us now discuss the elasticity of demand with reference to its different determinants.

8.3 PRICE ELASTICITY OF DEMAND

Price elasticity of demand is generally defined as the responsiveness or sensitiveness of demand for a commodity to the changes in its price. More precisely, *elasticity of demand is the percentage change in demand due to one per cent change in the price of the commodity*. A formal definition of price elasticity of demand (*ep*) is given as

ep = Percentage change in quantity demanded

Percentage change in price

A general formula1 for calculating coefficient of price elasticity, as given in Eq. (8.1), is given as follows:

1. The elasticity formula is derived as follows:

1

Q − 2

Q ×100 Δ *Q*

1

Q

1

Q

Δ *Q* 1

P

ep =

=

=

·

1

P − 2

P

Δ P

Δ P

1

×100

Q

1

P

1

P

where P 1 is original price, P 2 is new price Q 1 is quantity demanded at P 1 and Q 2 is quantity demanded at P 2.

Δ Q Δ P Δ Q

P

e

÷

=

×

p = Q

P

Q

Δ P

Δ Q P

=

×

Δ P Q

…(8.2)

where Q = original quantity demanded, P = original price, Δ Q = change in quantity demanded and Δ P = change in price.

It is important to note here that a minus sign (–) is generally inserted in the formula before the fraction in order to make the elasticity coefficient a non-negative value.2

The price elasticity can be measured between any two points on a demand curve (called *arc elasticity*) or at a point (called *point elasticity*). The measurement of the two kinds of price elasticity are discussed below in detail.

8.3.1 Arc Elasticity

The concept of arc elasticity of demand refers to the measurement of demand elasticity for a significant change in price and consequent change in demand. When there is a substantial or a big change in price, then the demand – price point shifts from one point to another on the demand curve. Thus, the measure of elasticity of demand between any two finite points on a demand curve is known as ***arc elasticity***. For example, measure of elasticity between points J and K (Fig. 8.1) is the measure of arc elasticity. The movement from point J to K on the demand curve (Dx) shows a big (50%) fall in the price from `20 tò10 so that Δ P = 20 – 10 = 10. The big fall in price causes a large increase in demand from 43 units to 75 units so that Δ Q = 43 – 75 = – 32. The arc elasticity between points J and K (moving from J to K) can be calculated by substituting these values into the elasticity formula as follows:

e

−Δ

p = –

Q P

· (with minus sign)

Δ *P Q*

32

−

20

= −

.

= 1.49

10 43

...(8.3)

This means that a one per cent

decrease in price of commodity *X* results

in a 1.49 per cent increase in demand

for it.

Problem in Using Arc Elasticity The

arc elasticity should be measured and

Fig. 8.1 *Linear Demand Curve*

used carefully, otherwise it may lead to

2. Price-elasticity of demand calculated without a minus sign will always be a negative value because either Δ *P* or Δ *Q* will carry a negative sign due to inverse relationship between price and quantity demanded. This gives a negative value of elasticity whereas in the concept of elasticity, a negative value has no meaningful interpretation expect that it indicates inverse relationship between *P* and *Q*. The negative elasticity coefficient is rather misleading. The 'minus' sign is, therefore, inserted as a matter of 'linguistic' convenience, to make the coefficient of elasticity non-negative. Sometimes, 'it is also suggested to ignore the negative sign in the numerator and denominator of the elasticity formula'. The elasticity in Eq. (8.7) ignores the negative sign.

wrong decisions. Arc elasticity co-efficients differ between the same two finite points on a demand curve if *direction* of change in price is reversed. For instance, as estimated in Eq. (8.3), the elasticity between points *J* and *K*—moving from *J* to *K* equals 1.49. It may be wrongly interpreted that the elasticity of demand for commodity *X* between points *J* and *K* equals 1.49 irrespective of the direction of price change. But it is not true. A reverse movement in the price, i.e., the movement from point *K* to *J* implies a different elasticity co-efficient (0.43). Movement from point *K* to *J* gives *P* = 10, Δ *P* = 10 – 20 = –10, *Q* = 75 and Δ *Q* = 75 – 43 = 32. By substituting these values into the elasticity formula, we get

ep = – 32 10

.

= 0.43

...(8.4)

10

−

75

The measure of elasticity co-efficient in Eq. (8.4) for the reverse movement in price is obviously different from one given by Eq. (8.3). It means that *the elasticity depends also on the direction of change* in price. Therefore, while measuring price elasticity, the direction of price change should be carefully noted. Otherwise, it will lead to a wrong decision regarding the change in price. For instance, if price elasticity between points *J* and *K* is taken to be the same whether price increases or decreases, it leads to the conclusion that total sales revenue will remain the same whether price increases or decreases. But, this is a wrong conclusion. The movement from point *J* to *K* yields a sales revenue tò10 × 75 = `750. But movement from point *K* to yields a sales revenue of `20 × 43 = `860. It means increasing price is beneficial and decreasing price is harmful.

Some Modifications Some modifications have been suggested in economic literature to resolve the problems associated with arc elasticity.

First, the problem arising due to the change in the direction of price change may be avoided by using the lower values of *P* and *Q* in the elasticity formula. In that case,

e

Δ *Q P*

p =

. *l*

Δ *P Ql*

where *Pl* = 10 (the lower of the two prices) and *Ql* = 43 (the lower of the two quantities). Thus,

ep = – 32 10

.

= 0.74

...(8.5)

10

−

43

This method is however devoid of the logic of calculating percentage change in price and demand because the choice of lower values of *P* and *Q* is arbitrary—it is not in accordance with the rule of calculating percentage change.

Second, another method suggested to resolve this problem is to use the average of upper and lower values of *P* and *Q* in fraction *P*/ *Q*. In that case the formula is

Δ *Q* (*P* + *P*) / 2

e

.

p =

1

2

Δ *P* (*Q* + *Q*) / 2

1

2

Q − *Q* (*P* + *P*) 2

or

e

.

p = 2

1

1

2

...(8.6)

P − *P* (*Q* + *Q*) 2

2

1

1

2

where subscripts 1 and 2 denote lower and upper values of prices and quantitites.

By substituting the values from our example, we get,

e

−

+

p = – 75 43 (20

10) 2

.

= 0.81

10 − 20 (43 + 75) 2

This method too has its own drawbacks as the elasticity co-efficient calculated through this formula refers to the elasticity mid-way between *P* 1 *P* 2 and *Q* 1 *Q* 2. The elasticity co-efficient (0.81) is not applicable for the whole range of price-quantity combinations at different points between *J* and *K* on the demand curve (Fig. 8.1)—it only gives a mean of the elasticities between the two points.

8.3.2 Point Elasticity

Point elasticity on a linear demand curve. Point elasticity is also a way to resolve the problem in measuring the elasticity. The concept of point elasticity is used for measuring price elasticity where change in price is infinitesimally small3.

Point elasticity is the elasticity of demand

at a finite point on a demand curve, e.g., at
point *P* or at point *B* on the linear demand curve
MN in Fig. 8.2. This is in contrast to the arc
elasticity between points *P* and *B*. A movement
from point *B* towards point *P* implies change in
price (D *P*) becoming smaller and smaller, such
that point *P* is almost reached. Here the change
in price is infinitesimally small. Measuring
elasticity for an infinitesimally small change
in price is the same as measuring elasticity
at a point. The formula for measuring point
elasticity is given below.

Fig. 8.2 *Point Elasticity*

∂ *Q P*

Point elasticity

(*e*

.

p)

...(8.7)

∂ *P Q*

Note that ∂ *Q* has been substituted for Δ *Q* in the formula for arc elasticity. The

∂ *P*

Δ *P*

derivative ∂ *Q* is reciprocal of the slope of the demand curve *MN*. Point elasticity is

∂ *P*

thus the product of price-quantity ratio at a particular point on the demand curve and the
reciprocal of the slope of the demand line.4 The reciprocal of the slope of the straight
line *MN* at point *P* is geometrically given by *QN/PQ*. Therefore,

Δ *Q QN*

=

Δ *P PQ*

3. The concept of point elasticity is a theoretical concept. Practically, this concept refers to a very small or insignificent change in price. For instance, if an automobile company reduces the price of its car price at `5 lakh by `5,000, it is an insignificant (1%) change in price.

4. Note that the slope of the demand surve, *MN*, at point *P* is given by *PQ/ON*.

Note that at point *P*, price *P* = *PQ* and quantity demand (*Q*) = *OQ*. By substituting
these values in Eq. (8.7), we get

PQ QN QN

e

.

=

p = *OQ PQ OQ*

Given the numerical values for *QN* and *OQ*, elasticity at point *P* can be easily obtained. We
may compare here the arc elasticity between points *J* and *K* and point elasticity at point *J* in
Fig. 8.1. At point *J*,

QN 108 − 43

e

=

p =

= 1.51

OQ

43

Note that point elasticity, *ep* = 1.51, is different from various measures of arc elasticities
(i.e., *ep* = 1.49, *ep* = 0.43, *ep* = 0.7, and *ep* = 0.81).

As has been proved below, geometrically, *QN*/ *OQ* = *PN*/ *PM*. Therefore, elasticity of
demand at point *P* (Fig. 8.2) may be expressed as

PN

ep = *PM*

Proof. The fact that *ep* = *QN*/ *PQ* = *PN*/ *PM* can be proved as follows. Note that in Fig. 8.2, there are three triangles—D *MON*, D *MRP* and D *PQN*—and ∠ *MON*, ∠ *MRP* and ∠ *PQN* are right angles. Therefore, the other corresponding angles of the three triangles will always be equal and hence, D *MON*, D *MRP* and D *PQN* are similar.

According to geometrical properties of similar triangles, the ratio of any two sides of a triangles are always equal to the ratio of the corresponding sides of the other triangles. By this rule, between D *PQN* and D *MRP*,

$$\frac{QN}{RP} = \frac{PN}{PM} \quad \text{...(8.8)}$$

Since *RP* = *OQ*, by substituting *OQ* for *RP*, Eq. (8.8) can be expressed as

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

It follows that

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

It means that price elasticity of demand at point *P* in Fig. 8.2 is given by

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

It may thus be concluded that the price elasticity of demand at any point on a linear demand curve is equal to the ratio of lower segment to the upper segments of the line, i.e.,

$$ep = \frac{\text{Lower segment of Demand Curve}}{\text{Upper segment of Demand Curve}}$$

Fig. 8.3 *Non-linear Demand Curve*

Point elasticity on a non-linear demand curve. The ratio D *Q*/D *P* in respect of a non-linear demand curve is different at each point. Therefore, the method used to measure point elasticity on a linear demand curve cannot be applied straightaway to measure elasticity on a curvilinear demand curve. A simple modification in technique is required. In order to measure point elasticity on a non-linear demand curve, the chosen point is first brought on a linear demand curve. This is done by drawing a tangent through the chosen point. For example, suppose we want to measure elasticity on a non-linear demand curve, *DD′* (Fig. 8.3) at point *P*. For this purpose, a tangent *MN* is drawn through point *P*. Since demand curve *DD′* and the line *MN* pass through the same point (*P*), the slope of the demand curve and that of the line at this point is the same. Therefore, the elasticity of demand curve at point *P* will be equal to that of the line at this point. Elasticity of the line at point *P* can be measured as

P ∂ *P* *PQ* *QN*

QN

e

·

=

.

=

p= Q ∂ P OQ PQ

OQ

As proved above, geometrically = QN

PN

=

Fig. 8.4 *Point Elasticities of Demand*

OQ PM

To conclude, at midpoint of a linear demand

curve, ep = 1. Note that in Fig. 8.4, point P falls

on the mid point of demand curve MN. At point, P, therefore, e = 1. It follows that at any

point above the point P, $ep > 1$, and at any point below the point P, $ep < 1$. According to this

formula, at the extreme point N, ep = 0, and at extreme point M, ep is undefined because division

by zero is undefined. It must be noted here that these results are relevant between points M

and N.

8.3.3 Measuring Price Elasticity from a Demand Function

The price elasticity of demand for a product can be measured directly from the demand function. In this section, we describe the method of measuring price elasticity of demand for a product from the demand function—both linear and non-linear. It may be noted here that if a demand function is given, arc elasticity can be measured simply by assuming two prices and working out ΔP and ΔQ. We will, therefore, confine ourselves here to point elasticity of demand with respect to price.

Measuring Price Elasticity from a Linear Demand Function

Suppose that a linear demand function is given as

$Q = 100 – 5P$

Given the demand function, point elasticity can be measured for any price. For example, suppose one has to measure elasticity at P = 10. Point elasticity is measured as

δ Q P

e

.

p = δ P Q

The term d Q/d P in the elasticity formula gives the slope of the demand curve. The slope of the demand curve can be found by differentiating the demand function. Thus

δ Q δ(100 − 5 P)

=

= 5

−

δ P

δ P

Having obtained the slope of the demand curve as δ Q/δ P = – 5, ep at P = 10 can be calculated as follows. At price P = 10, Q = 100 – 5(10) = 50. By substituting these values into the elasticity formula,

δ Q P

e

.

p = δ P Q

we get,

ep = (– 5) 10 = –1

50

Similarly, at

P = 8, Q = 100 – 5(8) = 60 and

ep = – 5 (8/60) = – 40/60 = – 0.67

And at

P = 15, Q = 100 – 5(15) = – 25, and

ep = – 5(15/25) = – 75/25 = – 3

Measuring Price Elasticity from a Non-linear Demand Function

Suppose a non-linear demand function of multiplicative form is given as follows

$$Q = aP^{-b}$$

and we want to compute the price elasticity of demand. The formula for computing the price elasticity is the same, i.e.,

$$e_p = \frac{\delta Q}{\delta P} \cdot \frac{P}{Q} \qquad \text{...(8.9)}$$

What one needs to compute the price-elasticity coefficient is to find first the value of the first term, d Q/d P, i.e., the slope of the demand curve. The slope can be obtained by differentiating the demand function, Thus,

$$\text{slope of demand curve} = \frac{\delta Q}{\delta P} = -baP^{-b-1} \qquad \text{...(8.10)}$$

By substituting Eq. (8.10) in Eq. (8.9), e_p can be expressed as

$$e_p = -baP^{-b-1}\left(\frac{P}{Q}\right) = \frac{-baP^{-b}}{Q} \qquad \text{...(8.11)}$$

Since $Q = aP^{-b}$, by substitution, we get

$$e_p = \frac{-baP^{-b}}{aP^{-b}} = -b \qquad \text{...(8.12)}$$

Equation (8.12) shows that when a demand function is of a multiplicative or power form, price elasticity coefficient equals the power of the variable P. This means that price *elasticity in the case of a multiplicative demand function remains constant all along the demand curve regardless of a change in price.*

8.3.4 Price Elasticity and Total Revenue

A firm aiming at enhancing its total revenue would like to know whether increasing or decreasing the price would achieve its goal. The price elasticity coefficient of demand for its product at different levels of its price provides the answer to this question. The simple answer is that if $e_p > 1$, then decreasing price will increase total revenue and if $e_q < 1$, then increasing price will increase total revenue. To prove this point, we need to know the total revenue (TR) and the marginal revenue (MR) functions and measures of price-elasticity are required. Since $TR = Q.\ P$, we need to know P and Q. This information can be obtained through the demand function. Let us recall our earlier demand function given as

$$Q = 100 - 5P$$

Price function (P) can be derived from the demand function as

$$P = 20 - 0.2Q \qquad \text{...(8.13)}$$

Given the price function, TR can be obtained as

$$TR = P \cdot Q = (20 - 0.2Q)\,Q$$

$$= 20Q - 0.2Q^2$$

…(8.14)

From this *TR*-function, the *MR*-function can be derived as

∂ *TR*

MR =

= 20 – 0.4 *Q* …(8.15)

∂ *Q*

The demand function and *MR*-function (8.15) are presented graphically in panel (a) and *TR*-function (8.14) in panel (b) of Fig. 8.5. As the figure shows, at point *P* on the demand curve, $e = 1$ where output, $Q = 50$. Below point *P*, $e < 1$ and above point *P*, $e > 1$. It can be seen in panel (*a*) of Fig. 8.5 that *TR* increases so long as $e > 1$; *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 8.1. As the table shows, when demand is *perfectly inelastic* (i.e., $ep = 0$ as is 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. 8.5 *Price Elasticity and Total Revenue*

As shown in panel (a), over the range of demand curve $e > 1$, quantity demanded increases by more than the proportionate decrease in price and hence the total revenue increase when price falls. The total revenue increases till price decreases till $e = 1$.

If demand for a product is *unit elastic* ($ep = 1$) quantity demanded increases (or decreases) in the proportion of decrease (or increase) in the price. It implies that a small change in price leaves total revenue remains unchanged. Therefore, total revenue remains unaffected.

If demand for a commodity has $ep < 1$, change in quantity demanded is greater than the proportionate change in price. Therefore, total revenue decreases 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 8.1 *Elasticity, Price-change and Change in TR*

Elasticity

Change in

Change in

Co-efficient

Price

TR

$e = 0$

Increase

Increase

Decrease

Decrease

$e > 1$

Increase

Decrease
Decrease
Increase
$e = 1$
Increase
No change
Decrease
No change
$e < 1$
Increase
Increase
Decrease
Decrease
$e = \infty$
Increase
Decrease to zero
Decrease
Infinite increase*
*Subject to the size of the market.

8.3.5 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.

Since *TR* = *P*. *Q*,

∂ ·
⎛ ∂ ⎞
⎛ ∂ ⎞
MR
P Q
Q
P
∂ *P*
= (
) = *P*
+ *Q*
= *P* + *Q*
Q
⎜⎝ *Q* ⎟
⎜
⎠
⎝ *Q* ⎟
∂
∂
∂ ⎠
∂ *Q*
⎛
∂ ⎞
=
1 *Q P*
P + ·
⎜⎝ *P Q* ⎟
∂ ⎠
...(8.16)

Q ∂ *P*
Note that ·
in Eq. (8.16) gives the reciprocal of elasticity. That is,

$$\frac{Q}{P}\cdot\frac{\partial P}{\partial Q} = \frac{1}{e_p}$$

Remember that e_p carrier a 'minus' sign.

By substitutng $1/e_p$ for $\frac{Q}{P}\cdot\frac{\partial P}{\partial Q}$ in Eq. (8.16), we get

$$MR = AR\left(1 + \frac{1}{e_p}\right)$$

or

$$MR = P\,[1 + (1/e_p)] \qquad \ldots(8.17)$$

Given this relationship between *MR* and price-elasticity of demand, the decision-makers can easily know whether it is beneficial to change the price. If $e = 1$, $MR = 0$. Therefore, change in price will not cause any change in *TR*. In case $e < 1$, $MR < 0$, *TR* decreases when price decreases and *TR* increases when price increases. And, if $e > 1$, $MR > 0$, *TR* increases if price decreases and *vice versa*.

Price Elasticity, AR and MR Given the Eq. (8.17), the formula for price elasticity (e_p) can be expressed in terms of *AR* and *MR*. We know that $P = AR$. So Eq. (8.17) can be written as

$$MR = AR\left(1 + \frac{1}{e_p}\right)$$

$$MR = AR + \frac{AR}{e_p}$$

By rearranging the terms, we get

$$MR - AR = \frac{AR}{e_p}$$

or

$$\frac{MR - AR}{AR} = \frac{1}{e_p}$$

The reciprocal of this equation gives the measure of the price elasticity (e_p) of demand which can be expressed as

$$\frac{1}{\dfrac{MR - AR}{AR}} = \frac{AR}{MR - AR} = e$$

e

p or $ep = MR - AR$

p

8.3.6 Determinants of Price Elasticity of Demand

We have noted above that price-elasticity of demand for a product may vary between zero and infinity. However, price-elasticity of demand, at a given price, varies from product to product depending on the following factors.

1. ***Availability of Substitutes.*** One of the most important determinants of elasticity of demand for a commodity is the availability of its close substitutes. The higher the degree

of closeness of the substitutes, the greater the elasticity of demand for the commodity. For instance, coffee and tea, rice and wheat, and petrol and diesel may be considered as close substitutes for one another. If price of one of these goods increases, the other commodity becomes relatively cheaper. Therefore, consumers buy more of the relatively cheaper good and less of the costlier one, all other things remaining the same. The elasticity of demand for the substitute goods will be higher. 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 is much greater than that for 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. Demand for luxury goods (e.g., high-price refrigerators, TV sets, cars, decoration items, etc.) is more elastic than the demand for necessities and comforts 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, vegetables) cannot be postponed and hence their demand is inelastic. Comforts have more elastic demand than necessities and less elastic than luxuries. Commodities are also categorized 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. ***Weightage in the Total Consumption.*** Another factor that influences the elasticity of demand is the proportion of income which consumers spend on a particular commodity. If proportion of income spent on a commodity is large, its demand will be more elastic. On the contrary, if the proportion of income spent on a commodity is small, its demand is less price-elastic. Classic examples of such commodities are salt, matches, books, pens, toothpastes, etc. These goods claim a very small proportion of income. Demand for these goods is generally inelastic because increase in the price of such goods does not substantially affect the consumer's budget. Therefore, people continue to purchase almost the same quantity even when their prices increase.

4. ***Time Factor in Adjustment of Consumption Pattern.*** Price-elasticity of demand depends also on *the time consumers need to adjust their consumption pattern to a new price*: the longer the time available, the greater the price-elasticity. The reason is that over a period of time, consumers are able to adjust their expenditure pattern to price changes. For instance, if the price of high quality cell phones is decreased, demand will not increase immediately unless people possess excess purchasing power. But over time, people may be able to adjust their expenditure pattern so that they can buy a laptop at a lower (new) price. Consider another example. If price of petrol is reduced, the demand for petrol does not increase immediately and significantly. Over time, however, people get incentive from low petrol prices to buy automobiles resulting in a significant rise in demand for petrol.

5. ***Range of Commodity Use.*** *The range of uses of a commodity* also influences the price-elasticity of its demand. The wider the range of the uses of a product, the higher the elasticity of demand for the decrease 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 increase in its price. For instance, milk can be taken as it is and in the form of curd, cheese, ghee and butter-milk. The demand for milk will therefore be highly elastic for decrease in price. Similarly, electricity can be used for lighting, cooking, heating and for industrial purposes. Therefore, with decrease in its price, demand for electricity has a greater elasticity. However, for the increase in price, such commodities have a lower price-elasticity because the consumption of a normal good cannot be cut down substantially beyond a point when the price of the commodity increases.

6. *Proportion of Market Supplied.* The elasticity of market demand also depends on the *proportion of the market supplied at the ruling price*. If less than half of the market is supplied at the ruling price, price-elasticity of demand will be higher than 1 and if more than half of the market is supplied, $e < 1$.

8.3.7 Application of Price Elasticity

Having explained the concept and measurement of elasticity, we discuss now the application of price elasticity of demand with respect to (a) manoeuvring price to maximize sales revenue and (b) determination of optimal price for profit maximization.

1. Manoeuvring of Price. Price manoeuvring means changing price of the product to achieve business objective. The concept of elasticity of demand plays a crucial role in business-decisions regarding manoeuvring of prices for the benefit of the firm. For instance, when cost of production is increasing, the firm would want to pass the rising cost on to the consumer by raising the price. Firms may decide to change the price even without any change in the cost of production. But, whether raising price following the rise in cost or otherwise proves beneficial or not depends on at least two factors:

(*a*) The price-elasticity of demand for the product, i.e., how high or low is the proportionate change in its demand in response to a certain percentage change in its price; and

(*b*) Price-elasticity of demand for its substitute, because when the price of a product increases, the demand for its substitutes increases automatically even if their prices remain unchanged.

Raising the price will be beneficial only if (*i*) demand for a product is less elastic; and (*ii*) demand for its substitute is much less elastic. Although most businessmen are intuitively aware of the elasticity of demand of the goods they make,5 the use of precise estimates of elasticity of demand will add precision to their business decisions.

2. Determination of Optimum Price. Another and a rather very important application of price elasticity is that it can be used, at least theoretically, to determine the optimum price with the objective of profit maximization for a firm facing downward sloping demand curve. How optimum price, i.e., profit maximizing price, can be determined by using the price elasticity is explained below. The necessary condition for profit maximization is given as *MR – MC*. That is, profit is maximized at the level of price and output at which

$$MR = MC \quad \ldots(8.18)$$

As has been shown in Eq. (8.17),

5. Mansfield, Edwin, *op. cit.*, p. 52.

$$MR = P\,[1 + (1/\,ep)] \quad \ldots(8.19)$$

By substituting Eq. (8.19) for *MR* in Eq. (8.18), the profit maximizing condition can be expressed as

$$P\,[1 + (1/\,ep)] = MC \quad \ldots(8.20)$$

The optimal price can be worked out from Eq. (8.20) as follows.

$$P = MC\,/\,[1 + (1/\,ep)] \quad \ldots(8.21)$$

Given the Eq. (8.21). if point elasticity of demand curve and marginal cost *(MC)* of a firm are known, the optimal price can be easily determined. For example, suppose point elasticity of demand at point on demand curve is estimated as *ep* = _ 2 and firm's *MC* = `50. In that case, optimal price can be worked out as follows.

$$P = 50\,/\,[1 + (1/-2)]$$

= `100

It may thus be concluded that the concept of price elasticity of demand can be used to manipulate the price to maximize the revenue of the firm given the demand function and to find the optimal price for profit maximization.

8.4 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 is the percentage change in its quantity demanded due to the change in the price of its substitute, coffee. The formula for measuring cross-elasticity of demand is the same as that of the price elasticity with a difference. For example, cross-elasticity of demand for tea ($e_{t,c}$) can be measured by the formula given below.

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

The cross-elasticity of demand for tea with respect price of coffee can be expressed technically as follows.

$$e_{t,c} = \frac{P_c}{Q_t} \cdot \frac{\Delta Q_t}{\Delta P_c} \qquad \ldots(8.22)$$

Similarly, cross-elasticity of demand for coffee with respect to change in the price of tea is measured as follows.

$$e_{c,t} = \frac{P_t}{Q_c} \cdot \frac{\Delta Q_c}{\Delta P_t} \qquad \ldots(8.23)$$

The same formula is used to measure the cross-elasticity of demand for a good with respect to a change in the price of its complementary goods. Electricity to electrical gadgets, petrol to automobiles, butter to bread, sugar to tea and coffee, are the examples of complementary goods.

It is important to note that when two goods are substitutes for one another, their demand has positive cross-elasticity because increase in the price of one good increases the demand for its substitute. And, the demand for complementary goods has negative

cross-elasticity, because increase in the price of a complementary good decreases the demand for the main *good*.

Uses of Cross-Elasticity

The concept of cross-elasticity has both theoretical and practical uses.

Theoretically, an *important use* of cross-elasticity is to define substitute goods. If cross-elasticity between any two goods is positive, the two goods may be considered as substitutes of one another. 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 one another: the higher the negative cross-elasticity, the higher the degree of complementarity.

Practically, the concept of cross-elasticity is of vital importance in pricing decisions, i.e., in changing prices of products having substitutes and complementary goods. If cross-elasticity in response to the price of substitutes is greater than one, it would be inadvisable to increase the price; rather, reducing the price may prove beneficial. In case of complementary goods also, reducing the price may be helpful in maintaining the demand in case the price of the complementary good is rising. Besides, if accurate measures of cross-elasticities are available, the firm can forecast the demand for its product and can adopt necessary safeguards against fluctuating prices of substitutes and complements.

The application of cross elasticity of demand, especially with regard to substitute good, can be explained clearly with an example. Suppose two firms, *A* and *B,* produce two goods *X* and *Y*, respectively, both the goods being substitutes for one another. The cross elasticity of demand for good *X* has been estimated at 1.5 and for good *Y* at 1.25. Given the cross elasticities, suppose firms intend to go for price competition. A question arises here: Is it advantageous for the firms to go for price competition?

Answer to this question can be found by assessing the effect of price change on the demand for their products. Suppose firm *A* cuts down its price by 10 per cent. As a result, the demand for *B's* product *Y* decreases by 10 × 1.5 = 15 per cent, as its demand get shifted to good *X*. Now, let the firm *B* react and cut down its price by 10 per cent. As a result, demand for good *X* decreases by 10 × l.25 = 12.5 percent as the substitution effect. It means firm *B* regains 12.5 per cent of its lost market (15 per cent). But firm *A* has still a gain of 15% –12.5 % = 2.5 percent. The final conclusions that emerge from this analysis are (*i*) firm *A* may initiate the price competition but not the firm *B,* and (*ii*) if firm *A* reduces its price, firm *B* has to react by reducing its own price, by a higher rate, if necessary.

8.5 INCOME-ELASTICITY OF DEMAND

Apart from the price of a product and its substitutes, consumer's income is another basic determinant of demand for a product. As noted earlier, the relationship between quantity demanded and consumers income is of positive nature, unlike the negative price-demand relationship. The demand for most goods and services increases with increase in consumer's income and *vice versa*. *The responsiveness of demand to the changes in income is known as* ***income-elasticity of demand.***

Income-elasticity of demand for a product, say *X*, (i.e., *ey*) may be measured as:

$$e_y = \frac{\%\ \text{change in demand for product } X}{\%\ \text{change in consumer's income } Y} = \frac{\Delta X_q / X_q}{\Delta Y / Y}$$

$$e_y = \frac{Y}{X_q} \cdot \frac{\Delta X_q}{\Delta Y} \quad \text{...(8.24)}$$

(where X_q = quantity of *X* demanded; *Y* = disposable income; ΔX_q = change in quantity of *X* demanded; and ΔY = change in income)

Obviously, the formula for measuring income-elasticity of demand is the same as that for measuring the price-elasticity. The only change in the formula is that the variable 'income' (*Y*) is substituted for the variable 'price' (*P*). Here, income refers to the disposable income, i.e., income net of taxes. All other formulae for measuring price-elasticities may by adopted to measure the income-elasticities, keeping in mind the difference between the independent variables and the purpose of measuring income-elasticity.

To estimate income-elasticity, let us suppose, for example, that the government announces a 10 per cent dearness allowance to its employees. As a result average monthly salary of government employees increases from `20,000 tò22,000. Following the pay-hike, monthly petrol consumption of government employees increases from 150 litre per month to 165 litre. The income-elasticity of petrol consumption can now be worked out as follows. In this case, $\Delta Y$ = `22,000 – `20,000 = `2,000, and ΔQ (oil demand) = 165 litre – 150 litre = 15 litre. By substituting those values in Eq. (8.24), we get

$$e_y = \frac{20{,}000}{150} \times \frac{15}{2{,}000} = 1$$

It means that income elasticity of petrol consumption by government employees equals 1. In simple words, $e_y = 1$ means that a one per cent increase in income results in a one per cent increase in petrol consumption.

Unlike price-elasticity of demand, which is always negative,[6] income-elasticity of demand is always positive[7] because of a positive relationship between income and quantity demanded of a product. But there is an *exception* to this rule. Income-elasticity of demand for an inferior good is negative, because of the inverse substitution effect. The demand for inferior goods decreases with increase in consumer's income. The reason is that when income increases, consumers switch over to the consumption of superior substitutes, i.e., 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; non-vegetarians buy more of meat and less of potato, and travellers travel more by plane and less by train.

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 of the three categories, viz., necessities, comforts and luxuries have different elasticities. The general pattern of

6. Except in case of Giffen goods.

7. With the exception of inferior goods.

income-elasticities of different goods for increase in income and their effect on sales are given in Table 8.2. As Table 8.2 shows, income elasticity of essential goods in less than 1. It is so because of Enget's law[8] Income elasticity of 'comforts' equals 1. And, in case of luxury goods, $e_y > 1$.

Table 8.2 *Income-Elasticities*

Consumer goods	*Co-efficient of income-elasticity*	*Effect on sales with change in income*
1. Essential goods	Less than one ($e_y < 1$)	Less than proportionate change in sale
2. Comforts	Almost equal to unity ($e_y \cong 1$)	Almost proportionate change in sale
3. Luxuries	Greater than unity ($e_y > 1$)	More than proportionate increase in sale

Income-elasticity of demand for different categories of goods may, however, vary from household to household and from time to time, depending on the choice and preference of the consumers, levels of 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 frequency of rise in income is high, income-elasticities will conform to the general pattern.

Uses of Income-Elasticity in Business Decisions

While price and cross elasticities of demand are of greater significance in price management aimed at maximizing the total revenue in the short run, income-elasticity of a product is of a greater significance in production planning and management in the long run, particularly during the period of a business cycle. The concept of income-elasticity can be used in estimating future demand provided that 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 a change in personal incomes is expected, other things remaining the same. It also helps in avoiding over-production or under-production. In forecasting demand, however, only the relevant concept of income and data should be used. It is generally believed that the demand for goods and services increases with increase in GNP, depending on the marginal propensity to consume. This may be true in the context of aggregate national demand, but not necessarily for each product. It is quite likely that increase in GNP flows to a section of consumers who do not consume the product in which a businessman is interested. For instance, if the major proportion of incremental GNP goes to those who can afford a car, the growth rate in GNP should not be used to calculate income-elasticity of demand for bicycles. Therefore, the income of only a relevant class or income-group should be used. Similarly, where the product is of a regional nature, or if there is a regional division of market between the producers, the income of only the relevant region should be used in forecasting the demand.

8. Engle's law states that as income increases, demand for necessities increases less than proportionately, and income elasticity of demand for necessities decreases with increase in income though it remains positive.

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 'normal goods'. On the other hand, goods whose income-elasticities are negative beyond a certain level of income are termed 'inferior goods'.

8.6 ADVERTISEMENT OR PROMOTIONAL ELASTICITY OF SALES

The expenditure on advertisement and on other sales-promotion activities does help in promoting sales, but *not at the same degree at all levels of the total sales* and total ad-expenditure. The concept of advertisement elasticity is useful in determining the optimum level of advertisement expenditure. The concept of *advertisement elasticity* assumes a greater significance in deciding on advertisement expenditure, particularly when there is competitive advertising by the rival firms. Advertisement elasticity (*eA*) of sales is measured as

$$eA = \frac{\%\text{ change in sales}}{\%\text{ change in Ad-expenditure}}$$

$$e_A = \frac{\Delta S / S}{\Delta A / A} = \frac{\Delta S}{\Delta A} \cdot \frac{A}{S} \qquad \ldots(8.25)$$

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

Suppose, for example, a company increases its advertising expenditure from `10 million tò12 million, and as a result, its sales increase from 5,000 units to 6,000 units. In this case $\Delta A$ = 12 million – 10 million = `2 million, and ΔS = 6,000 – 5,000 = 1000 units. By substituting these values in ad-elasticity formula (8.25), we get

$$eA = \frac{1000}{2} \times \frac{10}{5000} = 1$$

It means that a one per cent increase in ad-expenditure increases sales by 1 per cent.

Interpretation of Advertisement Elasticity The advertisement elasticity of sales varies between $eA = 0$ and $eA = \infty$ depending on the nature of the product, the level of market supplied, the trend in consumers' income, the compatitive strength of the competitors, etc. The interpretation of some measures of advertising elasticity is given below.

Elasticities

Interpretation

$eA = 0$

Sales do not respond to the advertisement expenditure.

$eA > 0$ but < 1

Increase in total sales is less than proportionate to the increase in advertisement expenditure.

$eA = 1$

Sales increase in proportion to the increase in advertisement expenditure.

$eA > 1$

Sales increase at a higher rate than the rate of increase of advertisement expenditure.

Determinants of Advertisement Elasticity

Some important factors that determine the level of ad-elasticity are the following.

(*i*) ***The level of total sales.*** In the initial stages of sale of a product, particularly of one which is newly introduced in the market, advertisement elasticity is greater than unity. Beyond a point of market supplied, however, sales increase, but ad-elasticity decreases. For instance, once potential market is supplied, the function of advertisement is to create additional demand by attracting more consumers to the product. But, consumer's respond to advertisement is generally low. 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. Simultaneous advertisement by the rival firms reduces sales of firm.

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

8.7 ELASTICITY OF PRICE EXPECTATIONS

Sometimes, mainly during the period of price fluctuations, consumer's price expectations play a much more important role than any other factor in determining the demand for a commodity. The concept of price-expectation-elasticity was devised and popularized by J.R. Hicks in 1939. The price-expectation-elasticity refers to the expected change in future price as a result of change in current prices of a product. The elasticity of price-expectation is defined and measured by the general formula given below.

Δf

P / P

ΔP

e

f

f

c

P

$x =$

$=$

.

…(8.26)

Δc

P / *c*

P

Δ *c*

P

f

P

Here, *Pc* = price in the recent past; Δ *Pc* = the current change in present price; *Pf* =

expected future price; and Δ *Pf* = expected change in future price.

For example, suppose current price of a storable commodity increases from `100 tò120. And, consumers of the commodity anticipate price of the commodity to increase in future from `120 tò150. In that cases.

150 −120

e

120

x = 120 −100

100

30 20

30 100

=

/

=

×

= 1.25

120 100 120 20

It means 1% change in present price will cause 1.25% change in future price.

The coefficient *ex* gives the measure of expected percentage change in future price as a result of 1 per cent change in present price. If *ex* > 1, it indicates that future change in price will be greater than the present change in price, and *vice versa*. If *ex* = 1, it indicates that the future change in price will be proportionately equal to the change in the current price.

The concept of elasticity of price-expectation is very useful in formulating future pricing policy. For example, if *ex* > 1, it indicates that sellers will be able to sell more in the future at higher prices. Thus, businessmen may accordingly determine their future pricing policy.

8.8 APPLICATION OF ELASTICITIES IN BUSINESS DECISION-MAKING

In the preceding section, we have discussed the concept and the method of measuring demand elasticities with respect to its determinants. We have also shown how various elasticities can be used in assessing the effect of change in each major determinant of demand on the demand for a product. In this section, we show how demand elasticities can be used in business decision-making by a firm, especially in regard to estimating the change in its total demand caused by a simultaneous change in most or all demand determinants, the independent variables.

The independent variables can be classified under two categories: (*i*) controllable variables, and (*ii*) uncontrollable variables. *Controllable variables* include price of the firm's own product, advertising expenses, and quality of the product. *Uncontrollable variables* include price of the substitute product, consumers' income, pricing strategy of the competitors, competitors' advertisement expenditure and their sales promotion strategy, growth of population, buyers' price expectations, etc. However, in estimating the future demand for its product, the firm has to identify all the major determinants of its product, both controllable and uncontrollable. While the firm can plan to make changes in the controllable variables, it can only anticipate the changes in uncontrollable variables or rely on changes reported by other agencies. In any case, the firm has to take into account the anticipated change in the variables beyond its control.

Let us suppose that a computer company, while estimating the demand function for its PC, identifies the following variables as most important determinants of demand for its computers.

- Product price *(P)*, i.e., price of its own PC
- Consumers' income *(I)*
- Price of the substitute brands *(Ps)*
- Advertisement *(A)*

Note that only two of these demand determinants, viz., price of firm's own PC and advertisement expenditure, are under the control of the firm and others are uncontrollable. Let us suppose also that the company has estimated the demand function for its PC as given below.

$$Qc = 50 - 1.5\ Pc + 0.5\ Y + 2.0\ Ps + 0.8\ A \quad \ldots(8.27)$$

where

Qc = number of company's PC demanded per unit of time

Pc = price of company's computer

Y = computer buyers' income

Ps = price of the substitute PC brands

A = ad-expenditure by the company

Suppose that, at a point in time, the numerical value of independent variables in Eq. (8.27) are given as follows: $Pc = 40$, $Y = 60$, $Ps = 30$ and $A = 25$. By substituting these values in the demand function, we get

$$Qc = 50 - 1.5\ (40) + 0.5\ (60) + 2.0\ (30) + 0.8\ (25) \quad \ldots(8.28)$$

$$= 50 - 60 + 30 + 60 + 20 = 100 \text{ thousand}$$

Thus, given the demand function in Eq. (8.27), the current annual demand for company's PC turns out to be 100,000 per time unit.

Given the current demand, the company is planning, as a matter of business policy, to make the following changes.

- increase its PC price by 10% and
- increase its ad-expenditure by 20%.

The company plans these changes in anticipation of an increase in PC-users' income (Y) by 8% per annum and no change in competitors' price.

Prior to implementing its plan, the company would like to know whether it would be advisable to make the planned changes in its price and ad-expenditure? An answer to this question can be found by assessing the impact of planned and anticipated changes on the demand for its PC. Since demand function for the company's PC is known [see Eq. (8.27)], the impact of planned and anticipated changes on the demand for its PC can be easily obtained by using demand elasticities.

What we need therefore is to first work out *price elasticity* (Ep), *income elasticity* (Ey) , *cross elasticity* (Es) and *advertisement elasticity* (EA) of demand for company's PC. These elasticities at point (i.e., point elasticities) are worked out as follows.

Recall, for example, the (price) elasticity formula.

$$E_P = \frac{\Delta Q}{\Delta P} \times \frac{P}{Q}$$

In this Ep formula, $\Delta Q/\Delta P$ is given by the estimated parameter (– 1.5) in the estimated demand function (8.28). So to find *(Ep),* we need to multiply –1.5 by *P/Q.* Note that in the estimated demand function, $Pc = 40$ and $Qc = 100$. Thus,

$$Ep = -1.5\ (P/Q) = -1.5\ (40/100) = -0.6$$

By using the same method, income-elasticity (Ey), cross-elasticity (Es) and advertisement-elasticity (Ey) can also be worked out.

$$Ey = 0.5\ (Y/Q) = 0.5\ (60/100) = 0.3$$

$$Es = 2.0\ (Ps/Q) = 2\ (30/100) = 0.6$$

$$Ea = 0.8\ (A/Q) = 0.8\ (25/100) = 0.2$$

Now that elasticities are measured, the impact of planned and anticipated changes in the independent variables on company's PC demand can be estimated by adjusting its existing demand of 100,000 units with effects of changes made in price and ad-expenditure. The anticipated demand for PC can be estimated as follows.

$$Qc = 1{,}00{,}000 - 0.6\ (10) + 0.3\ (8) + 0.6\ (0) + 0.2\ (20)$$

$$= 1{,}00{,}000 - 6 + 2.4 + 0 + 4 = 1{,}00{,}400$$

The demand estimate shows that if the company implements its plan of raising the price and ad-expenditure, the demand for its PC would increase only by 100,400 – 100,000 = 400 units. This is not a significant increase in the demand. The low increase in total demand is mainly because of decline in demand by 600 due to a 10% increase in PC price. The company would better be advised *not to increase the price* of its PC and rely on increase in PC users' income and on advertisement. In that case, the company will not lose a market of 600 PC. That is, if the company refrains from increasing the price of its PC, the total increase in demand for its product will be of the order of 1000 PC, which is, of course, a considerable increase in demand.

8.9 SOME ESTIMATES OF DEMAND ELASTICITIES

In this section, we present a summary of some estimates of demand elasticities, made in the United States and Britain.

Table 8.3 *Price Elasticities of Demand for Selected Products in the US*

Commodity	*Price elasticity of demand*	*Income elasticity of demand*
Tomatoes	4.60	—
Restaurant meals	1.63	1.40
Movies	0.87	—
Glassware	1.34	—
Foreign Air Travel	0.77	—
Taxi Service	1.24	—
Shoes	0.70	1.40
Radio & TV Service	1.19	—
Auto Repair	0.36	—
Furniture	1.01	1.48
Medical Insurance	0.31	0.92
Housing	1.00	—
Gasoline and oil	0.14	0.48
Alcohol	0.92	

1.54

Owner Occupied Housing

—

1.49

Source: Houthakker H. and L.D. Taylor, *Consumer Demand in the United States: Analysis and Projections*, Mass., Harvard University Press, 2nd Edn. quoted from Mansfield, E., *op. cit.*

Table 8.4 *Price Elasticities of Demand for Car Models in US*

1. Chevrolet Impala

Relative price elasticity

– 14.79

Cross-elasticity w.r.t.

Pontiac Catalina

19.30

2. Pontiac Catalina

Relative price elasticity

– 16.99

Cross elasticity w.r.t.

chevrolet Impala

5.09

3. Plymouth Fury

Relative price elasticity

– 4.59

Cross elasticity w.r.t.

Chevrolet Impala

4.22

Pontiac Catalina

0.49

Ford Calaxy

6.82

Source: Irvine, F.O., Jr., “Demand Equations for Individual New Car Models Estimated Using Transition Prices with Implications for Regulatory Issues.” *South Eco. Jl.*, (4.9), January 1983.

Table 8.5 *Demand Elasticities of Electricity in US*

Variable

Residential use

Commercial use

Industrial use

Electricity price

– 0.794

– 0.916

– 1.404

Per capita income

0.714

1.249

– 0.450

Price of natural gas

0.159

– 0.193

– 0.293

Source: Halvorsen R., “Demand for Electricity Energy in the United States,” *South Eco. Jl.*, (43), April 1976.

Table 8.6 *Price and Income Elasticities of Some Selected Consumer Goods in Britain*

Consumer

Price

Income

Goods

Elasticity

Elasticity

Food

0.00

0.21

Housing

0.31

0.03

Fuel

0.28

1.67

Drinks and Tobacco

0.60

1.22

Transport and Communication

1.21

1.23

Source: Deaton and Muellbauer, 1980.

SUMMARY

• Elasticity of demand refers to the degree of responsiveness of demand for a product to the change in its determinants, specifically price of the product, price of the substitutes, price of the complementary goods, income of the consumers, advertisement of the product, and future price expectations.

• Given these demand determinants, elasticity of demand is measured as (*i*) price elasticity, (*ii*) cross elasticity, (*iii*) income elasticity, (*iv*) advertisement elasticity, and (*v*) price expectation elasticity.

• A general method for measuring demand elasticity (*Ed)* is given below.

% change in demand for the product (*X*)

d

E =

% change in demand determinant factor (*F*)

The same method is used to measure all kinds of demand elasticity with specific reference to the different demand determinants.

• Price elasticity of demand refers to the degree of responsiveness of demand for a product to the change in its price. Price-elasticity of demand is measured by the following method.

% change in demand for the product (*X*)

d

E =

% change in product price (*x*

P)

The general formula used for measuring price elasticity is given follows.

Δ

E =

x

Q × *xP*

d

Δ *x*

P

x

Q

• Price elasticity measured at a point of the demand curve is called point elasticity of demand and elasticity measured between any two points of the demand curve is called arc elasticity.

• Given the demand function for a product, price elasticity of demand can be measured straightaway by changing the variable (price) in the demand function. (For details, see section 8.3.3).

• Price elasticity has important applications in pricing decisions. It can be used to estimate the change in total revenue and marginal revenue of the firm. As a simple rule, decreasing price is beneficial if elasticity (*Ed)* is greater than 1 and increasing price is beneficial if elasticity is greater than 1.

• Cross elasticity of demand refers to the responsiveness of demand for a product to the change in the price of its substitutes and complements. Formula for measuring cross elasticity is the same as one used for measuring price elasticity. In case of cross elasticity, the price of substitute and complementary good is used instead of the price of the product whose elasticity is measured.

• From business decision point of view, cross elasticity is taken in view by the firms while changing the price of their own product.

• Income elasticity of demand refers to the responsiveness of demand for a product to the change in the income of the consumers of the product.

• Similarly, advertisement elasticity refers to the responsiveness of demand for a product to the advertisement of the product, and price expectation elasticity refers to responsiveness of demand for a product to the advertisement of the product.

REVIEW QUESTIONS

1. What is meant by the elasticity of demand? How is the elasticity of demand measured?

2. What are the different kinds of demand elasticities? What is the importance of different kinds of demand elasticites in price management?

3. Define and distinguish between:

(*a*) Arc elasticity and point elasticity, (*b*) Price elasticity and cross-elasticity, and

(*c*) Income elasticity and price elasticity.

4. When prices of both substitutes and complements of a commodity, say *X*, rise, what happens to the demand for commodity *X*?

(*a*) rises,

(*b*) falls,

(*c*) remains constant, or

(*d*) all of the above possibilities exist

5. List the major purposes of demand analysis from the standpoint of business management. Can managers manipulate all the variables which affect demand?

6. (*a*) Distinguish between linear and non-linear demand functions.

(*b*) What is the difference between the following demand functions?

(*i*) $Q_x = 1 - 5 P_x$;

(*ii*) $Q_x = 100 - 2 P 2 x$

(*iii*) $Q_x = Ap - b$ and

(*iv*) $Q_x = (a/p + c) - b$

7. What is meant by the elasticity of demand? Suppose price elasticity co-efficient for a commodity is estimated at –2. What does it mean?

8. Which of the following commodities has the most inelastic demand and why?

(*a*) Soap, (*b*) Salt, (*c*) Penicillin, (*d*) Cigarettes and (*e*) Ice-cream.

9. Explain the following concepts separately:

(*i*) Income elasticity of demand,

(*ii*) Price elasticity of demand and

(*iii*) Elasticity of price expectations.

What useful information do these concepts of elasticity provide to management?

10. Suppose the demand function for a commodity is given as

$Q = 12 - P$

(*a*) Find the demand and marginal revenue schedules,

(*b*) Plot the *AR* and *MR* schedules,

(*c*) Find marginal revenue when $P = 10$, 6 and 2, and

(*d*) Estimate the elasticity co-efficient of the demand curve, when the total revenue is at the maximum.

11. Define *elasticity of price expectation* (*Ee*). In the context of an environment of business recession, state briefly the implication of:

(*i*) $Ee > 1$, (*ii*) $Ee = 1$, (*iii*) $0 < Ee > 1$, (*iv*) $Ee = 0$ and (*v*) $Ee < 0$.

12. A publishing company plans to publish a book. It finds from the sales data of other publishers of similar books that the demand function for the book can be expressed as $Q = 5000 - 5P$. Find out:

(*a*) Demand schedule and demand curve,

(*b*) Number of books sold when *P* = `25,

(*c*) Price for selling 2500 copies,

(*d*) Price for zero sales,

(*e*) Point-elasticity of demand at pricè20, and

(*f*) Arc elasticity for a fall in price from `25 tò20 and for a rise in price from `20 tò25.

13. Suppose the demand function for a product is given as *Q* = 500 – 5 *P*. Find out:

(*i*) Quantity demanded at pricè15,

(*ii*) Price to sell 200 units,

(*iii*) Price for zero demand, and

(*iv*) Quantity demanded at zero price.

14. Which of the following statements is true?

(*i*) If price elasticity = 1, *MR* = 0

(*ii*) If price elasticity > 1, *MR* > 0

(*iii*) If price elasticity < 1, *MR* < 0

(**Ans.** All Three)

15. Suppose individual demand schedules for *A*, *B* and *C* are given as follows:

Price (`)	*A′ s* *demand*	*B′ s* *demand*	*C′ s* *demand*
5	80	40	20
10	40	20	10
15	20	10	5
20	10	5	0
25	0	0	0

Find

(*i*) market demand schedule,

(*ii*) market demand curve,

(*iii*) elasticity when price falls from `15 tò10, and

(*iv*) elasticity when price rises from `10 tò15.

FURTHER READING

Baumol, William J. *Economic Theory and Operations Analysis* (Prentice-Hall of India Pvt. Ltd., New Delhi, 1985).

Bilas, RichaRd a, *Microeconomic Theory* (McGraw-Hill, New York, 1971).

BRigham, EugEnE F. and JamEs, l. PaPPas, *Managerial Economics* (Dryden Press, Hinsdale, Illinois).

douglas, Evans J., *Managerial Economics: Theory and Practice and Problems* (Prentice-Hall Inc., New Jersey).

dWivEdi, d.n., *Microeconomics: Theory and Applications* (Vikas Publishing House, Delhi, 2003).
hiRschEy, maRk, *Economics for Managers* (Thomson South-Western, Delhi, 2006), Chs. 4 and 5.
mansFiEld, E., W.B. allEn, n.a. dohaRty and k. WEigElt, *Managerial Economics: Theory, Application, and Cases* (W.W. Nortion & Co., NY, 2nd Edn., New York, 2002).
PEtERson, h. gRaig and cRis, W. lEWis, *Managerial Economics* (Pearson Education, Delhi, 1999).
salvatoRE, dominick, *Managerial Economics in Global Economy* (Thompson—South-Western, 4th Edn. 2001).
thomas, chRistoPhER R and chaRlEs, mauRicE s. *Managerial Economies: Concepts and Application* (Tata McGraw-Hill Publishing Company, New Delhi, 8th Edn. 2005), Ch. 2.

CHAPTER

9 Demand Forecasting

CHAPTER OBJECTIVES

The objective of this chapter is to describe the method of estimating and forecasting demand for some future date. This chapter tells you:

- Why firms need to go for demand forecasting
- What steps are followed in demand forecasting
- What are the various methods of demand forecasting
- Different survey methods of demand forecasting, their advantages and limitations
- The various statistical methods of demand forecasting, their requirement and application

9.1 INTRODUCTION

The theory of demand discussed in preceding chapters deals with how current demand is determined whereas a major concern of businessmen, especially the big ones, is 'what would be the future demand for their product?' This question arises because the knowledge about future demand for the firm's product helps a great deal in forward planning for production, acquiring inputs (man, material and capital), managing finances and chalking out future pricing strategy, etc. For this purpose, businessmen make their own estimates or even a 'guesstimate' of the future demand for their product or take the help of specialized consultants or market research agencies to get the demand forecast for their product.

There are a variety of methods used for demand forecasting, that are used depending on the purpose and perspective of forecasting. In this chapter, we discuss the methods of estimating future demand, i.e., methods of demand forecasting. Let us first look at the need for demand forecasting in some detail.

9.2 WHY DEMAND FORECASTING

The business world is characterized by risk and uncertainty and, therefore, most business decisions are made under the condition of risk and uncertainty. One way to reduce the adverse effects of risk and uncertainty is to acquire knowledge about the future demand prospects for the product. The information regarding the future demand for the product is obtained by demand forecasting. *Demand forecasting is predicting the future demand for firm's product*. A raliable forecast of the future demand for the product helps a great deal

in the following areas of business management.

- Determining the production target,
- Planning and scheduling production,
- Acquiring inputs (labour, raw material and capital),
- Making provision for finances,
- Formulating pricing strategy,
- Planning advertisement.

Demand forecasting assumes greater significance where large-scale production is involved. Large-scale production requires a good deal of forward planning as it involves a long gestation period. The information regarding future demand is also essential for the existing firms to be able to avoid 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 products because they will have to acquire inputs and plan their production accordingly. The firms are hence required to estimate the future demand for their products. Otherwise, their functioning will be shrouded with 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 or seasonal demand or the demand of a routine nature. Their past experience and business skills may be sufficient for the purpose of planning and production. But, firms working on a large scale find it extremely difficult to obtain a fairly accurate estimate of future market demand.1 In some cases, it is very difficult to obtain information needed to make even short-term demand forecasts and thus it is extremely difficult to make long-term forecasts. Under such conditions, it is not possible for the firm 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 an approximate estimate of the demand prospects. Also, demand forecast plays an important role in planning for acquiring of 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 essential element of decision-making by the large-scale firms.

9.3 STEPS IN DEMAND FORECASTING

The demand can be forecast reliably only when forecast is made systematically and scientifically and when it is fairly reliable. The following steps are generally taken to make systematic demand forecasting.

(i) Specifying the objective. The objective or the purpose of demand forecasting must be clearly specified. The objective may be specified in terms of (*a*) short-term or long-term demand, (*b*) industry demand for a product or for firm's own product, (*c*) the whole or only a segment2 of the market for its product, or (*d*) firm's market share. The objective

1. Eugene, F. Brigham, and James L. Pappas, *Managerial Economics* (The Dryden Press, Hinsdale. Illinois, 1976), p. 129.

2. The market segment may be defined in terms of geographical region, income-groups of the consumers or a particular section of the society (e.g., libraries and students segment of market for textbooks).

of demand forecasting must be determined before the process of forecast is started. This has to be the first step.

(ii) Determining the time perspective. Depending on the firm's objective, demand may be forecast for a short period, i.e., for the next 2–3 years, or for a long period. In demand forecasting for a short period, many of the demand determinants can be taken to remain constant or not to change significantly. In the long run, however, demand determinants may change significantly. Therefore, the time perspective of demand forecasting must be specified as it helps in making choice of appropriate demand determinant.

(iii) Making choice of method for demand forecasting. As we will see below, there are a number of different methods of demand forecasting. However, all methods are not suitable for all kinds of demand forecasting because the purpose of forecasting, data requirement of a method, availability of data and time frame of forecasting vary from method to method. The demand forecaster has therefore to choose a suitable method keeping in view his purpose and requirements. The choice of a forecasting method is generally based on the purpose, experience and expertise of the forecaster. It depends also to a great extent on the availability of required data. The choice of a suitable method saves not only time and cost but also ensures the reliability of forecast to a great extent.

(iv) Collection of data and data adjustment. Once method of demand forecasting is decided on, the next step is to collect the required data—primary or secondary or both. The required data is often not available in the required mode. In that case, data needs to be adjusted—even massaged, if necessary—with the purpose of building data series consistent with data requirement. Sometimes the required data has to be generated from the secondary sources.

(v) Estimation and interpretation of results. As mentioned above, the availability of data often determines the method, and also the model to be used for demand forecasting. Once required data is collected and forecasting method is finalized, the final step in demand forecasting is to make the estimate of demand for the predetermined years or the period. Where estimates appear in the form of an equation, the result must be interpreted and presented in a usable form.

9.4 METHODS OF DEMAND FORECASTING: AN OVERVIEW

There are various methods of demand forecasting. As pointed out above, the choice of method for forecasting demand depends on the purpose and kind of demand forecasting and availability of required data. The various methods of demand forecasting are listed below.

1. Survey Methods

(*i*) Consumer Survey – direct interview

(*ii*) Opinion Poll Methods

2. Statistical Methods

(*i*) Trend Projection

(*ii*) Barometric Methods

(*iii*) Econometric Methods

All these methods have different kinds of sub-methods. A detailed list of methods of demand forecasting is given in a Chart on the next page. The process of demand forecasting by different methods are discussed here elaborately.

ecasting

Demand For

Techniques of

Chart:

9.5 SURVEY METHODS

Survey methods are the most common and simple methods of estimating current demand and projecting future demand for a product. Survey methods are generally used when the purpose is to make short-run forecast of demand for a product. Under survey methods, the required information is collected through a survey of consumers/users. The survey seeks information of consumers' future plan to buy the product for which demand has to be forecast. Surveys are conducted by two methods.

(*i*) Consumer Survey Method—the Direct Consumer Survey, and

(*ii*) Opinion Poll Method.

The choice of any of these methods depends on the status of the availability of primary data and time and money that firms are willing to spend on the survey. If primary data on demand for a product is not available in any form and money and time are not constraints, then *direct consumer survey* is the only option. But, in case some relevant information is already available with experts and time and money are the constraints, the firms may go for *opinion poll of experts*. The process of conducting these kinds of surveys is discussed below.

9.5.1 Consumer Survey Method – Direct Interview

Under this method, the potential consumers of the product are directly interviewed to collect the required data on their demand for the product. Depending on purpose, time and cost, consumer surveys are conducted by three methods: (*a*) Complete enumeration, (*b*) Sample survey, and (*c*) End-use survey method.

In brief, 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 under reference they would be willing to buy at different prices over a given period, say, one year. This method is known as *direct interview method*. 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.

These consumer survey methods are used under different conditions and for different purposes. Their advantages and disadvantages are described below.

(*a*) Complete Enumeration Method. The complete enumeration method is used when market size is small and all consumers can be contacted by the surveyors. In this method, almost all present and potential users of the product are contacted and are 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 majority 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

n

D

$\sum q$

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

i

$i = 1$

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

This method has certain **limitations .** It can be used successfully only in case of those products whose consumers are concentrated in a certain region or locality. In case of a

widely dispersed market, this method may not be physically possible or may prove to be very costly in terms of both money and time. Besides, the demand forecast through this method may not be reliable for the following reasons.

(*i*) Consumers themselves may not have the exact idea of their demand in future and hence may be unable or unwilling to tell the exact quantity;

(*ii*) Even if they answer, their answers to hypothetical questions may be hypothetical—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. Sample survey method is used when population of the target market is very large and expanded over a large area. Under sample survey method, only a sample of potential consumers or users is selected for interview. Consumers to be surveyed are selected from the relevant market through a sampling method. 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.

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

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).

An important **advantage** of this method is, it is very simple, less costly and less time-consuming compared to the comprehensive survey method. This method is generally used to estimate short-term demand of business firms, government departments and agencies and also of 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 budgets of their expenditure can provide reliable information about their purchases.

Besides, the sample survey method can be used to verify the demand forecast made by using quantitative or statistical methods. Although some authors[3] 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, expectations, etc.) is not required and where consumer's behaviour is subject to frequent changes.

The sample survey method is the most widely used survey method to forecast demand. But this method has certain **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.

3. For example, Brigham, Eugene F. and James L. Pappas, *op. cit.* , pp. 548–49.

(*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 demand 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 that 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. The different stages of the end-use method and problems associated with each stage are described here briefly.

First Stage. 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 rule of this method of forecasting. Difficulty arises because published data on the end-users are rarely available. In case relevant and adequate data are not available, the managers need to have a thorough knowledge of the product and its uses. Such knowledge and experience need to be supplemented by consultations and discussions with manufacturers or their associations, 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 users 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.

Second Stage. 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 involve social, moral and ethical values, e.g., in case of consumption of drugs, alcohol or running a dance bar. 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 applied. 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 be derived mainly from the information collected through the questionnaires.

Third Stage. 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 that use the product and the likely output targets.

Fourth Stage. The ***fourth*** and ***final 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. 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 of demand forecasting, the future demand can be estimated only at the aggregate level. This is because past data are seldom available the types and sizes of the product demanded by the economy. Even if projections are made by using the past data, either by the trend method, or by regression techniques or by historical analogies, it produces 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.

9.5.2 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*) Delphi method, and

(*c*) Market studies and experiments.

(*a*) Expert-Opinion Method. Firms having a good network of sales representatives can put them on to the work of assessing the demand for the target 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 reactions 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 predict the future demand by using their experience and expertise. This method is also known as *opinion poll method*.

Limitations 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 and expertise to analyze the market and their experience.

Secondly, demand estimates may involve the subjective judgement of the expert assessor, which may lead to over or under-estimation—a risk element.

Finally, the assessment of market demand is usually based on inadequate information available to the sales representatives as they 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. Therefore, reliability of demand forecast becomes doubtful.

(*b*) Delphi Method.[4] 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 to arrive at a compromise estimate of future demand. The process is simple.

Under Delphi method, the task of projecting demand for a product is assigned anonymously to a group of market experts with a group leader. The task is assigned anonymously in the sense that experts are not aware of each other. Market experts make their own demand projection individually for the product and submit their report to the group leader. The group leader interchanges the demand projection reports between the experts anonymously seeking their opinion on the accuracy and reliability of projection made by the other experts. The group leader seeks also the suggestions for the modification of projections made by the other experts. The comments, remarks and suggestions made by the experts on the demand projection made by experts are again interchanged between the experts. The experts may revise their own estimates on the basis of suggestions and projections made by other experts, if they feel so. The group leader may find a compromise estimate of projections to be taken as final estimate of demand forecast. The compromise forecast is taken as the final demand forecast.

The ***advantages*** of Delphi method of forecasting are (*i*) it provides opportunity to experts to forecast the demand systematically, (*ii*) it overcomes the disadvantages of face-to-face discussion by the panel of experts as it puts pressure of compromising unwillingly, and (*iii*) it offers a fairly reliable demand forecast as it is based on systematic approach.

However, some critics of Delphi method point out ***drawbacks*** of this method. They argue that Delphi method is not scientific as it is based on intuitive judgments and hence it provides unreliable demand projection. This argument, however, does not negate the advantages of Delphi Method because individual experts use scientific methods, not intuitive judgment, in forecasting demand and final estimate is based on the a consensus of the experts.

It may be concluded that Delphi method has the merit of being a last resort in projecting demand under complex conditions. It may be noted, as a proof, that the empirical studies conducted in the USA have shown that unstructured opinions of the experts is the most widely used forecasting technique. 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.

(*c*) Market Studies and Experiments. An alternative method of collecting necessary information regarding current and also future demand for a product 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

4. The origin of "Delphi method" is traced to Greek mythology. In ancient Greece, Delphi was an oracle of Apollo and served as a medium for consulting deities. In modern times, Delphi method was developed by Olaf Helmer at the Rand Corporation of the US, as a method of obtaining a consensus of panelists without direct interaction between them. (J.R. Davis and S. Chang, *Managerial Economics* (Prentice-Hall, N.J., 1986, p. 191).

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.[5] After such changes are introduced in the market, the consequent changes in the demand over a period of time (a week, a fortnight, or a month) are recorded. On the basis of data collected, elasticity coefficients are computed. The elasticity coefficients are then used along with the variables of the demand function to assess the future demand for the product.

Alternatively, market experiments can be replaced by *consumer clinics* or *controlled laboratory experiments*. 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 certain serious limitations and disadvantages that 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.

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

Thirdly, 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.

Fourthly, changes in socio-economic conditions during the field experiments, such as local strikes or lay-offs, aggressive advertising 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.'[6]

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 that 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 mattresses—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 same 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.[7]

5. Brigham, Eugene F. and James, L. Pappas, *op. cit.* , p. 135.

6. Webb, Samuel C., *Managerial Economics,* (Houghton Miffln Company, Boston, 1976), p. 156.

7. Dean, J., *Managerial Economics*, (Englewood Cliffs, N.J., Indian Edn., 1960), p. 181.

9.6 STATISTICAL METHODS

In the foregoing sections, we have described survey and experimental methods of estimating and forecasting demand for a product on the basis of information supplied by the consumers themselves and on-the-spot observation of consumer behaviour. In this section, we will explain **statistical methods** which utilize historical (time-series) and cross-sectional 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*) 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 because forecasts are based on observed facts, and

(*iv*) Estimation involves smaller time and money cost.

Three kinds of statistical methods are used for demand projection.

(1) Trend Projection Methods,

(2) Barometric Methods, and

(3) Econometric Method.

These statistical methods are described here briefly.

9.6.1 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 fairly justifiable 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."8

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

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

(*a*) Graphical method,

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

(*c*) Box-Jenkins method.

8. Balakrishna, S., *Techniques of Demand Forecasting for Industrial Products*, p. 4.

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 2015, 2016 and 2017. For this purpose, it uses time-series data on its total sales over the past 10 years. Suppose its time-series sales data are given as in Table 9.1.

Table 9.1 *Time-Series Data on Sale of Bread*

Year

2005 2006 2007 2008 2009 2010 2011 2012 2013 2014

Sales of Bread (' 000 tonnes) 10

12

11

15

18

14

20

18

21

25

Let us first use the graphical method and project demand for the year, 2009.

(*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 Fig. 9.1 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 (marked *M* and *S*), we can forecast an approximate sale of 26,200 tonnes in 2017.

Fig. 9.1 *Trend Projection*

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 2014, and extend the trend line beyond point *P* towards *O*, and predict a sale of 30,000 tonnes of bread in 2017. On the other hand, a conservative analyst may consider the fluctuating nature of sales data and expect the total sale in 2017 to remain the same as in 2014 as indicated by the line *PC*. One may even predict a fall in the sale to 25,000 tonnes, if one over-emphasizes the fluctuating nature of sales in one's judgement. This is indicated by the line *PN*. It implies that trend projection method may yield three or mroe than three estimates. So the reliability of the demand projection becomes doubtful.

(*b*) Trend Fitting 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.9 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 Fig. 9.1) or by trying different forms of trend equations for the best fit.

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

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

$$S = a + bT \qquad \ldots(9.1)$$

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

The co-efficients *a* and *b* are estimated by solving the following two equations based on the principle of least square.

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

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

The terms included in Eqs. (*i*) and (*ii*) are calculated using sales data given in Table 9.1 and presented in Table 9.2.

By substituting numerical values given in Table 9.2 in Eqs. (*i*) and (*ii*), we get

$$164 = 10a + 55b \qquad \ldots(iii)$$

$$1024 = 55a + 385b \qquad \ldots(iv)$$

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

$$S = 8.26 + 1.48T$$

Table 9.2 *Estimation of Trend Equation*

Year	*S (Sales)*	*T (Time)*	T^2	*ST (S × T)*

2005	10	1	1	10
2006	12	2	4	24
2007	11	3	9	33
2008	15	4	16	60
2009	18	5	25	90
2010	14	6	36	84
2011	20	7	49	140
2012	18	8	64	144
2013	21	9	81	189
2014	25	10	100	250
$n = 10$	$\Sigma S = 164$	$\Sigma T = 55$	$\Sigma T^2 = 385$	$\Sigma ST = 1024$

9. The statistical technique used to find the trend line, i.e., the 'least square method', has already been discussed in Chapter 5, see 5.4.2.

Having estimated the parameters of the trend equation, it is quite easy to project the sales for 2015, 2016 and 2017 respectively. The calculation procedure is given below.

2015

S 2

=

8.26 + 1.48 (11)

=

24,540 tonnes

2016

S 3

=

8.26 + 1.48 (12)

=

26,020 tonnes

2017

S 4

=

8.26 + 1.48 (13)

=

27,500 tonnes

Treatment of the Abnormal Years An abnormal year is one in which sales are abnormally low or high. Time series data on sales may reveal, more often than not, abnormal years. 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 analyzed 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 the total sale (or any dependent variable) has increased over the past years at an increasing rate or at a constant percentage rate per time unit, then the appropriate trend equation to be used is an exponential trend equation of any of the following forms.

(1) If trend equation is given as

$Y = aebT$

…(9.2)

then its semi-logarithmic form is used

$\log Y = \log a + bT$

…(9.3)

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

(2) If trend equation takes the following form

$Y = aTb$

…(9.4)

then its double logarithmic form is used.

$\log Y = \log a + b \log T$

…(9.5)

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

(3) Polynomial trend of the form

$Y = a + bT + cT$ 2

…(9.6)

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 lies in 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, unlike regression analysis, this method 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 elasticities) that 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. Box-Jenkins method10 of forecasting is used only for short-term projections and predictions. Besides, this method is suitable for forecasting demand with only stationary time-series sales data. Stationary time-series data is one that does not reveal a long-term trend. In other words, Box-Jenkins technique can be used only in those cases in which time-series analysis depicts monthly or seasonal variations recurring with some degree of regularity.

When sales data of various commodities are plotted, in casse of many commodities, it will show a seasonal or temporal variation in sales. For examples, 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 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. Box-Jenkins technique is used for predicting demand where time-series sales data reveals this kind of seasonal variation.

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

(*i*) autoregression model, (*ii*) moving average model, and

(*iii*) autoregressive-moving average model.

The autoregressive-moving average model is the final form of the Box-Jenkins model. The three models are, in fact, the three stages of Box-Jenkins method. The purpose of the three models of 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 briefly here with the purpose of introducing Box-Jenkins method to the reader rather than providing the entire methodology.11

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 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 data 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 **final step** is to use the models to predict sales in the intended period.

The Box-Jenkins models are described here briefly in the same sequence.

10. A new method of demand forecasting was suggested by G.E.P. Box and G.M. Jenkins in their book, *Time Series Analysis: Forecasting and Control* (Holdan-Day, San Francisco, 1970). That is why this mehtod is named after their names.

11. Computer programs on Box-Jenkins method are available for use.

(*i*) *Autoregressive 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.

$$Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \ldots + a_n Y_{t-n} + e_t \quad \ldots(9.7)$$

This model states that the value of Y in period t depends on the values of Y in periods $t - 1$, $t - 2$ ... $t - n$. 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, ... a_n are different from zero, it reveals seasonality in data. This completes the second step.

The model (9.7), however, does not specify the relationship between the value of Y_t and residuals (e_t) of previous periods. Box-Jenkins method uses moving average method to specify the relationship between Y_t and e_t, the values of residuals in previous years. This is the third step. Let us now look at the moving average model of 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.

$$Y_t = m + b_1 e_{t-1} + b_2 e_{t-2} \ldots + b_p e_{t-p} + e_t \quad \ldots(9.8)$$

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

There is alternative model of moving average method discussed below.

(iii) Autoregressive-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.

$$Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \ldots + a_n Y_{t-n} + b_1 e_{t-1} + b_2 e_{t-2} + \ldots + b_p e_{t-p} + e_t \quad \ldots(9.9)$$

Box-Jenkins method of forecasting demand is a sophisticated and highly complicated method. This method is, however, impracticable without the aid of computer.

Moving Average Method: An Alternative Technique

As noted above, the moving average model of Box-Jenkins method is a part of a complicated technique of forecasting demand in time t on the basis of its past time values. There is a simple, rather a naïve, yet useful method of using moving average to forecast demand. This 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 year t; X_{t-1}, $_{t-2}$ … $_{t-n}$ = demand or sales in previous years; N = number of preceding years.

According to this method, the likely demand for a product in year 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 9.3 and we want to forecast demand for refrigerators for the year 2015.

Table 9.3 *Sale of Refrigerators: 2000-2006*

Year	2008	2009	2010	2011	2012	2013	2014
Sales ('000)	12	14	14	15	15	17	18

Given this sales data, demand for the year 2015 will be computed as follows.

$$D_{2007} = \frac{1}{7}(18 + 17 + 15 + 15 + 14 + 14 + 12) = 15$$

Thus, the demand for refrigerators for 2015 is forecast at 15,000 units. Now suppose that the actual sales of refrigerators in the city in 2015 turns out to be 18,000 refrigerators against the forecast figure of 15,000. Given the actual sales figure for 2015, the demand for 2016 can be forecast as

1

D 2008 = 7 (18 + 18 + 17 + 15 + 15 + 14 + 14) = 15.86

Note that, in the moving average method, the sale of 2015 is added and the sale of 2008 (the last of the preceding years) is excluded from the formula. The reason is, that the demand has to be forecast on the basis of sales in the past 7 years.

The moving average method is simple and can be used to make short-term forecasts. However, this method has a serious **limitation**, which has to be borne in mind while using it. 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-estimation 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.

9.6.2 Barometric Method of Forecasting

The barometric method of forecasting follows the method that 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. It was, however, abandoned as it had failed to predict the Great Depression of the 1930s.12 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.13

It may be noted at the outset that the barometric technique was developed by the NBER 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, allotment of land by the Delhi Development Authority (DDA) to the Group Housing Societies (a lead indicator) indicates higher demand prospects for building materials—cement, steel, bricks, etc.

12. Lange, O., *Introduction to Econometrics*, 2nd Edn. (Oxford Pergamon Press, 1962), pp. 85–95.

13. A summary of use and findings of this method can be had from R. Davis and Semoon Chang, *Principles of Managerial Economics* (Prentice-Hall, NJ, 1986).

Barometric method of demand forecasting includes two methods of choosing the economic indicators demand.

(*i*) Lag-lead indicators, and

(*ii*) Diffusion index.

(i) Load-lag Indicators. 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 lead lag indicators used in this method are classified as:

(*a*) leading indicators,

(*b*) coincidental indicators, and

(*c*) lagging indicators.

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

The *coincidental indicators*, on the other hand, are the ones that move up or down simultaneously with the level of general economic activities. 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; and (*iv*) sales recorded by the manufacturing, trading and the retail sectors.

The *lagging indicator series,* consist of those indicators that 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 various indicators are chosen 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; and

(*vi*) Smoothness of the series.

(ii) Diffusion Index. The problem of choice in lead-lag indicators may arise because some of the indicators appear in more than one class of indicators. Also, 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 is the percentage of rising indicators. A diffusion index copes up with the problem of differing signals given by the 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 one is the diffusion index of the leading series. However, there are problems of identifying the leading indicator for the variables under study.

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.

9.6.3 Econometric Methods

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. We explain here briefly the use of econometric methods 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 explanatory economic variables are so interrelated and interdependent 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.

(1) Regression method, and

(2) Simultaneous equations model.

These methods are explained here briefly.

(1) Regression Method14. 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, one needs to estimate the demand function for a product. Recall that in estimating a demand function, demand is a 'dependent variable' and the variables that determine the demand are called 'independent' or 'explanatory' variables. For example, demand for pizza in a city may be said to depend largely on 'per capita income' of the city and its population. Here demand for pizza 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

14. Students are advised to have a review of Chapter 5.

to depend largely on the population of the city. Under this condition, the demand functions for salt and sugar contain only one independent variable, i.e., population. In that case, the demand functions for these goods 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 independent variables like commodity's own price, price of its substitutes, household incomes, population, etc. Such demand functions are called ***multi-variable demand functions.*** For a single-variable demand function, 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.

(a) Simple or Bivariate Regression Technique

The simple regression technique has already been discussed in Chapter 5. Recall that 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. Let us now discuss the *simple regression technique* of forecasting demand for a commodity.

Suppose we have to forecast demand for sugar for a country for the year 2015 on the basis of 7-year data given in Table 9.4. When this data is graphed, it produces a continuously rising trend in demand for sugar with rising population. This shows a linear trend. Now, country's demand for sugar in 2015 can be obtained by estimating a regression equation of the form

$$Y = a + bX \quad \text{...(9.10)}$$

where Y is sugar consumed, X is population, and a and b are the two parameters.

For an illustration, consider the hypothetical data on a country's annual consumption of sugar given in Table 9.4.

Table 9.4 *Annual Consumption of Sugar*

Year	*Population (millions)*	*Sugar Consumed (million tonnes)*
2008	10	40
2009	12	50
2010	15	60
2011	20	70
2012	25	80
2013	30	90
2014	40	100

Equation (9.10) can be estimated by using the 'least square' method. The procedure is the same as shown in Table 9.2. That is, the parameters a and b can be estimated by solving the following two linear equations:

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

$$\Sigma X_iY_i = \Sigma X_i a + bX_i^2 \quad \ldots(ii)$$

The procedure of calculating the terms in Eqs. (*i*) and (*ii*) above is presented in Table 9.5.

Table 9.5 *Calculation of Terms of the Linear Equations*

(Figures in million)

Year	*Population* (*X*)	*Sugar consumed* (*Y*)	X^2	*XY*
1999–2000	10	40	100	400
2000–01	12	50	144	600
2001–02	15	60	225	900
2002–03	20	70	400	1400
2003–04	25	80	625	2000
2004–05	30	90	900	2700
2005–06	40	100	1600	4000
$\Sigma n = 7$	ΣX 2 $i = 152$	$\Sigma Y_i = 490$	$\Sigma X_i = 3994$	$\Sigma X_iY_i = 12000$

By substituting the values from Table 9.5 into Eqs. (*i*) and (*ii*), we get

$490 = 7a + 152b$

…(*iii*)

$12{,}000 = 152a + 3994b$

…(*iv*)

By solving Eqs. (*iii*) and (*iv*), we get

$a = 27.44$ and $b = 1.96$

By substituting values for *a* and *b* in Eq. (9.10), we get the estimated regression equation as

$Y = 27.44 + 1.96X$

...(9.11)

Given the regression Eq. (9.11), the demand for sugar in 2015 can be easily projected if population for 2015 can be known. Supposing population for 2015 is projected to be 50 million, the demand for sugar in 2015 can be estimated as follows.

$Y = 27.44 + 1.96(50) = 126$ 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 for the test of the reliability of estimates, consult a standard book on statistics).

(b) Multi-variate Regression

The multi-variate regression equation is used where demand for a commodity is considered to be a function of more than one explanatory variables. The application of multiple regression analysis is 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 from the determinants of demand, viz., price of the product, price of its substitutes, consumers’ income and their tastes and preferences. For estimating the demand for durable consumer goods, (e.g., TV sets, refrigerators, houses, etc.), the other explanatory variables that 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 are collected, the *third and a very important 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 equation with the help of statistical techniques. The multi-variate equation cannot be easily estimated manually. Therefore, the equation has to be estimated with the help of a 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 linear equation, the most common form of equation used for estimating demand is given below.

$Qx = a - bPx + cY + dPy + jA$

…(9.12)

where Qx = quantity demanded of commodity X; Px = price of commodity X; Y = consumers’ income, Py = 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 Px, Y, Py and A, respectively.

In a linear demand function, quantity demanded is assumed to change at a constant rate with a change in each of the independent variables $P x$, Y, Py and A. The parameters (regression co-efficients) 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.

$$Qs = 50 - 0.75\,Ps + 0.1\,Y + 1.25\,Py + 0.05\,A \quad \ldots(9.13)$$

The numerical values in this equation express the quantitative relationship[15] between demand for sugar and the variables with which it is associated. More precisely, regression co-efficients 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 tonne change in sugar demand and a change of one rupee in income leads to a 0.1 tonne change in sugar demand, and so on.

Power Function In linear Eq. (9.12), the coefficients of independent variables are assumed to be constant and independent of change in other variables. For example, it assumes that the parameter of price (*P*) is independent of change in income (*Y*) or other independent variables, and so on. However, one can find cases in which it is theoretically and also empirically found that the coefficients of the independent variables are 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

15. Estimated values of parameters, – 0.75, 0.1, 1.25 and 0.05 are the regression co-efficients of demand with respect to *Px*, *Y*, *Py* and *A*, respectively.

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

$$Q_x = a\,P_x^{-b}\,Y^{c}\,P_y^{d}\,A^{j} \quad \ldots(9.14)$$

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

$$\log Qx = \log a - b \log Px + c \log Y + d \log Py + j \log A \quad \ldots(9.15)$$

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 co-efficients. After regression co-efficients 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 made by using statistical methods 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 reliable related 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 function 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 co-efficients of independent variables. What is more important, therefore, is to carefully ascertain the theoretical relationship between the dependent and the independent variables.

9.6.4 Simultaneous Equations Model

Recall that regression technique of demand forecasting consists of a single equation. In contrast, the simultaneous equations model of forecasting involves 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 that are beyond the scope of this book16. We will, therefore, restrict ourselves here only to the basic features of this method of forecasting. Let us look at the steps that are taken to build the simultaneous equation model.

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.

Endogenous variables are those that are determined within the model. These 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*.

Exogenous variables are those that are determined outside the model. These 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 in case of demand forecasting 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 to be even generated from the available primary or secondary sources.

After the model is developed and necessary data are collected, the ***third 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.

The application of simultaneous equation model is illustrated with a simple example. Suppose a simple macroeconomic model is given as follows.

$$Yt = Ct + It + Gt + NXt \quad \ldots(9.16)$$

where

Yt = Gross national product,

Ct = Total consumption expenditure,

It = Gross private investment,

Gt = Government expenditure,

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

and subscript t represents a given time unit.

16. For detailed discussion on the use of econometric methods in business decision, see J.W. Elliott, *Econometric Analysis for Management Decisions* (Homewood, Irwin, 1973).

Equation (9.16) is an identity, that may be explained with a system of simultaneous equations. Suppose in Eq. (9.16)

$$Ct = a + bYt \quad \ldots(9.17)$$

$It = 20$

...(9.18)

$Gt = 10$

...(9.19)

$NX = 5$

...(9.20)

In the above system of equations, Yt and Ct are endogenous variables and It, Gt and NXt are exogenous variables. Eq. (9.17) is a regression equation that has to be estimated. Equations (9.18), (9.19) and (9.20) show the values of exogenous variables determined outside the model.

Suppose we want to predict the value of Yt and Ct simultaneously. Suppose also that when we estimate Eq. (9.17), we get

$Ct = 100 + 0.75\ Yt$

...(9.21)

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

$Yt = Ct + 20 + 10 + 5 = Ct + 35$

Since $Ct = 100 + 0.75\ Yt$, by substitution, we get

$Yt = 100 + 0.75\ Yt + 35$

then

$Yt – 0.75\ Yt = 100 + 35$

$0.25\ Yt = 135$

and

$Yt = 135/0.25 = 540$

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

$Ct = 100 + 0.75\ Yt = 100 + 0.75\ (540) = 505$

Thus, the predicted values are

$Yt = 540$ and $Ct = 505$

Thus,

$Yt = 505 + 20 + 10 + 5 = 540$

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.

However, this method is theoretically superior to the regression method. The main advantage of this method is that it is capable of capturing the effect 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.

9.7 CONCLUDING REMARKS

There are several methods and techniques available for forecasting demand for a product. All the methods have their own limitations and advantages, merits and demerits, in varying degrees. The applicability and usefulness of a method depends on the purpose of forecasting and availability of reliable and relevant data. The analyst should, therefore, choose a method or a technique of demand forecasting that is relevant to the purpose, convenient to handle, applicable to the available data and also inexpensive.

It may be added that mere possession of the right tools does not necessarily ensure an accurate forecast. Equally important is the analyst’s own judgement. In fact, the role of analyst’s judgement cannot be over-emphasized. It is required at all stages of demand forecasting. Any estimate relating to a future period in itself is a judgement and involves a series of judgements. Analyst’s judgement is required in the first place to ensure that the method used in making the estimate is appropriate to the task. Also, when a statistical series is generated, its interpretation requires judgement.

Besides, several components and elements used in forecasting are often in conflict and are so balanced that they leave little room for choice. Nevertheless, one or the other has to be judged as being appropriate for use under the given circumstances.

Furthermore, events relating to economic activity take place sometimes suddenly and sometimes in an uneven manner. Without adequate information about the current situation, there are bound to be errors of judgement in statistical projections. Most people, if they lack the correct perspective, cannot fathom the significance of even current information unless it is properly related to all concerned activities. Therefore, interpreting the current information or data and making proper use of the same for future projection also calls for judgement. Finally, is should be appreciated that forecasting is merely an attempt to utilize the generally accepted methods or techniques for knowing the future demand for a product. By the very nature of the problem, there can be no guarantee of accuracy in any specified methodology or system of indicators. Hence, it would be unreasonable to be dogmatic about the results obtained or to say that any one particular method is superior to others. As time proceeds, opportunities are provided for testing the validity of any forecast. A person who has done a forecast should be prepared to recede from a position previously taken whenever it is known that economic and other conditions that formed the basis for the previous forecast have changed. The only sound procedure is to correct the previous forecast in as objective a manner as possible. There is need for continual revision so that forecast gets closer to reality.

9.8 SOME CASE STUDIES OF DEMAND FORECASTING

In this section, we briefly describe some studies in demand forecasting. Not many case studies with Indian data are available on the various methods of demand forecasting. The case studies that are available mostly use the regression method. The cases described below include some on consumer goods and some on intermediate goods.

9.8.1 Consumer Goods

Eggs Eggs are one of the popular items of food for non-vegetarians and semi-vegetarians. Sidhu[17] estimated demand function for eggs for Ludhiana district of Punjab for various occupational groups in rural and urban areas. We describe here only the results for all groups combined. In his aggregated demand functions he considered the following variables:

1. Quantity of eggs consumed (the dependent variable),
2. Size and composition of family,
3. Family income,

17. Sidhu, D.S., *Demand and Supply of Eggs: An Economic Analysis* (S. Chand & Co., New Delhi, 1974). The study was made for 1968-69.

4. Occupation, and
5. Number of earning members in the family.

However, in his annual demand function he included only two variables viz., (*i*) quantity of eggs consumed, and (*ii*) per capita disposable income, for lack of data and problems of specification. He estimated the following forms of demand function:

(*i*) $Y = a + bX$ (Linear)

(*ii*) $Y = aX^b$ (Exponential)

The estimated functions respectively, are:

(*i*) 3.0085 + 0.0619 X

$R^2 = 0.6569$

(13.5301) (0.0030)

(*ii*) – 2.0119 + (1.2108) X

$R^2 = 0.5276$

(0.2739) (0.0769)

The linear function gave a 'consistently better fit to the data'. But, for the urban households in which both husband and wife were employed the 'Cobb-Douglas' form, i.e., exponential form, of function gave a better fit.

Sidhu has also calculated the income-elasticities of demand for eggs, based on both linear and exponential demand functions.

The estimated income elasticities (e_y) are

(*i*) $e_y = 0.9876$ (Linear), (*ii*) $e_y = 1.2108$ (Exponential)

Both these elasticities are statistically significant. Now, if per capita income projections are available, the demand for eggs can be forecast for the successive years.

Soap Balakrishna[18] has estimated demand functions for various durable and non-durable consumer goods and has, by using different methods, forecast demand thereof for the late sixties and the early seventies. We pick up only a few to illustrate the practical forecast. For the purpose of forecasting the future demand for soap in India, the two variables which affect the consumption significantly, namely, growth in population and increase in per capita income can be chosen for regression. Unfortunately, the true consumption levels for soap in the past years are not available. Table 9.6 shows the consumption built up only on the basis of indigenous production from the organized sector and from imports.

Regression equation $Y = -425.5541 + 1.1756\, x_1 + 6.4544\, x_2$

where Y is consumption of soap; x_1 is population; and x_2 is per capita income.

Coefficient of multiple correlation $R_{1.23} = 0.848$.

In India, quite a substantial portion of the demand for soap is met from the production in the small-scale sector and therefore any projection based on the data furnished in Table 9.6 cannot be truly meaningful. Nevertheless, a multiple regression equation was obtained with the data contained in Table 9.6 as shown below.

18. Balakrishna, S. *Technique of Demand Forecasting for Industrial Product.*

Table 9.6 *Apparent Consumption of Soap in India* [19]

(Tonnes)

Years	*Production*	*Import*	*Total (2 + 3)*	*Export*	*Apparent consumption (4 – 5)*
(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
1950-51	77,255	174.5	77,429.5	282.3	77,147.2
1951-52	87,747	122.3	87,869.3	1,606.2	86,263.1
1952-53	86,772	117.0	86,889.0	—	86,889.0
1953-54	82,492	96.7	83,588.7	634.3	82,954.4
1954-55	90,765	62.8	90,827.8	566.0	90,261.8
1955-56	104,304	100.2	104,402.2	472.0	

103,930.2
1956-57
115,198
251.6
115,449.6
310.5
115,139.1
1957-58
112,689
110.3
112,799.3
293.9
121,006.7
1958-59
127,195
43.2
127,238.2
365.9
126,872.3
1959-60
134,799
45.0
134,844.0
464.5
134,379.5
1960-61
143,805
21.7
143,826.7
628.1
143,198.6
1961-62
147,922
13.5
147,935.5
591.6
147,343.9
1962-63
151,721
1.4
151,722.4
563.1
151,159.3
1963-64
164,468
0.5
164,468.5
640.4
163,828.1
1964-65
164,402
1.5
164,403.5
920.6
163,482.9
1965-66
164,130
1.2

164,131.2

1,932.0

162,199.2

The analysis of variance for the multiple regression is as follows:

Source of Variation	*Degree of freedom*	*Sum of squares*	*Mean sum of squares*	*F. Ratio*
Due to regression	2	2233.26	1116.63	
Residual about regression	5	872.66	174.53	6.3979
Total	7	3105.9		

Since *F* ratio is significant at 5 per cent level, one may accept the hypothesis that the regression of *Y* on x_1 and x_2 is jointly linear. The coefficient of multiple correlation 0.848 indicates that the variables together explain 72 per cent of all the variance in consumption of soap during the period under review. Since the variables are logically consistent so as to be related to the dependent variable, soap, projections have been made using the above equation and the results are given below:

Years	*Thousand tonnes*
1965–66	133.25
1970–71	170.42
1975–76	192.62

The projection for the year 1965-66 as given above is lower than the result obtained by the trend method.

Projection on Per Capita Basis. A rational method for projecting the consumption levels onto the future is on per capita basis. By and large, it is reasonable to assume that

19. Reproduced from Balakrishna, *op. cit.*

the future per capita consumption level of non-durable consumer goods will not be less than that obtained for the current year. In that case the anticipated increase in population will alone determine the aggregate consumption in the future. The future population levels would have, of course, to be determined first by the trend method. The projections given above are based on this assumption and procedure.

9.8.2 Intermediate Goods

Steel. Table 9.7 furnishes the available data on the consumption of finished steel in India from 1950-51, 1964–65. The following techniques of forecasting were applied to forecast the demand for steel on the basis of its past apparent consumption and the various economic indicators.

(a) Trend method. In the trend method, both the linear and non-linear models have been attempted. The two equations based on two models, their parameters and the projections for the various years are now given as follows:

Equations: (*a*) *Linear Equation:* $Y = 471.54 + 286.07\,t$ ($Y = a + bt$)

(*b*) *Non-linear Equation:* $Y = 1791.3\, t\, 3 + 223.3\, t + 22.6$

Steel Demand Forecast

(Million tonnes)

Year	(*a*) *Linear*	(*b*) *Non-linear*
1965-66	—	5.630
1970-71	5.907	9.230
1975-76	7.337	14.190

In 1965-66, the actual consumption of finished steel was 4.9020 million tonnes, consisting of 4.45 million tonnes from domestic production and 0.492 million tonnes of imports less exports of 0.045 million tonnes. It is obvious that the trend method has failed to give reliable results in this case even with regard to a foreseabe future year. The explanation may lie in the fact that constraints in foreign exchange have depressed imports of steel in the recent years. Nevertheless, the fact that the trend method has not given a close result even for 1965–66 and also the wide divergence between the results by the two sets of equations in the trend method call for further probe by other methods.

(b) Regression method. Taking up the regression method next, two independent projections—one with a single variable (industrial output) and the other with national income gave the following results. The correlation in both the cases is 0.94, which is quite high.

Year	*Regression on industrial output (million tonnes)*	*Regression on national income*
1965-66	4.095	4.134
1970-71	5.757	5.656
1975-76	7.972	7.916

The correlation with industrial output is significant as the industries sector accounts for nearly 60 to 65 per cent of the total consumption of steel in India and, therefore,

20. This does not take into account the steel content of imported machinery and other finished goods made from steel.

the results obtained on this basis should be more realistic. But, the forecast obtained by this regression is nearly a million tonnes short compared to the actual consumption in 1965–66. The estimates for the future years by the regression method are also lower than those obtained by the trend method.

Table 9.7 *Consumption of Finished Steel in India*

('000 tonnes)

Years	*Indigenous Production*	*Imports*	*Total of Cols. 2 + 3*	*Exports and Re-exports*	*Apparent Consumption*

1	*2*	*3*	*4*	*5*	*6*
1950-51	1044.7	154.0	1198.7	7.7	1191.0
1951-52	1100.4	57.7	1198.1	7.7	1154.9
1952-53	1105.0	68.5	1173.5	3.2	1171.1
1953-54	1103.1	156.7	1259.8	2.4	1256.4
1954-55	1263.4	293.3	1556.0	5.0	1551.7
1955-56	1296.8	659.2	1956.7	2.4	1953.6
1956-57	1370.0	1114.4	2484.4	0.1	2484.3
1957-58	1361.1	679.3	2041.2	0.3	2040.9
1958-59	1354.2	294.7	1658.9		

Neg.
1648.9
1959-60
1907.1
366.4
2273.5
0.3
2273.2
1960-61
2459.1
476.5
2935.6
1.0
2934.6
1961-62
3006.2
389.7
3395.1
2.3
3393.6
1962-63
3986.7
242.0
4228.7
7.1
4221.6
1963-64
4297.7
301.4
4599.1
13.2
4585.9
1964-65
4431.8
410.2
4842.0
n.a.
4842.0

The demand for finished steel was forecast by using multiple regression model. For the purpose, data on the two variables, namely, industrial output and construction, (which are closely related and which account for more than 80 per cent of the total consumption of steel), together with data on national income were used. The multiple correlation coefficient and the regression equation estimated with the available data are given below:

(*i*) *Regression equation*

$y = 7.91752\, x\ 1 + 15.26547\, x\ 2 + 1.0092\, x\ 3 - 6635888$

where y is the consumption of steel

x 1 = index of industrial output

x 2 = index of construction activity

x 3 = national income

(*ii*) *Coefficient of multiple correlation* = 0.9255

Sum of squares

Degree of freedom

Mean sum of squares

Regression

5,128,218.09

3

179,406.03

Residual about regression

858,781.09

8

107,347.63

F = 15.924 is highly significant. The coefficient of multiple correlation r 1.123 indicates that the three variables together explain 74 per cent of all the variance in steel demand during the years that have been considered for establishing the correlation. The estimated future consumption using the above regression equation is given below.

Years	*Steel consumption in million tonnes*
1965-66	4,406
1970-71	6,590
1975-76	9,777

It is interesting to observe that the demand estimate for 1965–66 by multiple regression is closer to the actual consumption in that year than those obtained either by the trend or by regression method on a single variable, like industrial output or national income.

(c) End-use method. The end-use study that was carried out in 1963 for steel consumption,21 the aggregate figures, for 1965–66 and 1970–71 were as follows. The demand for 1975–76 was not estimated in that study.

Kind of steel	*1965-66*	*1970-71*
Finished steel	6.902	135.94
Crude steel	9.735	18.287

There is a difference of two million tonnes between the end-use estimate for 1965-66 and the actual consumption of finished steel in that year. A probe into the actual output in 1965-66, of the industries and other sectors that consume steel, would reveal the factors accounting for the shortfall. It was found that the shortfall was mostly in the industrial sector.

SUMMARY

• Chapter 9 deals with the need for and methods of demand forecasting. Forecasting demand is estimating and predicting the future demand for a product.

• Demand forecasting plays an important role in making business decisions in regard to (*i*) determining the production target, (*ii*) planning and scheduling production, (*iii*) acquisition of inputs, and (*iv*) managing the required finances.

• The steps that are in general followed for demand forecasting are (*i*) specifying the production target, (*ii*) determining the time perspective, (*iii*) making choice of method of demand forecasting, (*iv*) determining the method of data collection, and (*v*) demand estimation and prediction.

• There are two broad methods of demand forecasting: (*i*) survey methods, and (*ii*) statistical methods. The choice of method depends on the purpose and availability of required data.

• Under survey methods data of consumer demand are collected by (*a*) consumer survey, and (*b*) opinion poll of the market experts. In case of consumer survey, three methods are used; (*i*) complete enumeration of consumers, (*ii*) sample survey method, and (*iii*) end-use method.

• Under opinion poll methods, two different methods are used: (*i*) collecting expert opinion, and (*ii*) market experiment. Expert opinion is collected through a sample of experts or through Delphi method – collecting expert opinion and getting the opinion counterchecked.

21. “Reappraisal of Steel Demand”, 1963, National Council for Applied Economics Research, New Delhi.

• Statistical methods include (*i*) trend projection method, (*ii*) biometric method, and (*iii*) econometric method. Trend projection method uses the past data and projects the future possible demand accordingly. Sales data are used also to estimate the sales function which is used to project the future demand. Trend projection method is most widely used method of demand projection.

• Barometric and econometric methods are used in case demand for a product depends on a variety of interrelated and interdependent economic variables and also on non-economic factors. These methods are highly technical. (For details, see the relevant sections).

REVIEW QUESTIONS

1. What is the purpose of demand forecasting? What are the necessary steps that need to be taken for forecasting demand for a product?

2. What are the methods used for forecasting demand? Discuss in detail the survey methods of forecasting demand for consumer goods. What are their advantages and limitations?

3. What is the purpose of applying the end-use method of demand forecasting? What are the necessary steps taken in forecasting demand for an industrial input? What are the advantages and limitations of this method?

4. What are the opinion poll methods of demand forecasting? What is the difference between the ‘expert opinion poll’ method and ‘Delphi method’? Which of the two methods is preferable under what conditions?

5. What is the purpose of using statistical methods of demand forecasting? Discuss the trend projection method of demand forecasting. What are the advantages and limitations of this method?

6. What is Box-Jenkins method of demand forecasting? What are the techniques used in this method? Discuss the moving average method of the Box-Jenkins approach?

7. Explain the regression method of demand forecasting. What is the difference between bivariate and multivariate regression techniques? Which of these regression techniques will you use forecast demand for passenger cars and why?

8. Suppose you are required to estimate future demand for a branded mobile phone. What independent variables will you include in the demand function to forecast the car demand? Explain the reason for selecting the independent variables.

9. Since the Delhi government is facing socio-political problem due to sharp rise in onion price from `30 p/kg tò90 p/kg. So the government plans to import onion to meet the domestic demand for one year. The authorities want you to forecast annual demand for onion. What kind of demand function will you use to estimate the annual demand for onion? Give reasons for your choice.

10 “The concept of elasticity of demand and demand forecasting are versatile tools of economic analysis.” Discuss the validity of this statement with appropriate examples.

11. Plot the following data on a graph and find the trend equation for sales:

Year	Total sales (units)
1970	1150
1971	1020
1972	3050
1973	3000
1974	950
1975	3060
1976	4030

12. Demand function for the product of a shoe-manufacturing company is given as $Q = -0.70\ P + 0.45\ A$ (where P = price per pair and A = advertisement cost per unit). The company sells 50,000 pairs of shoes per annum at `600 per pair. What will be the annual sales if the company spends

`1 lakh on advertisement? (Note that advertisement cost increases price.)

13. You are given the following data:

Estimate the regression equation $Y = a + bX$.

X	3	6	8	10	13	13	13	14
Y	8	6	10	12	12	14	14	20

(**Ans .** $Y = 0.8125 + 3.875\,X$)

14. Which independent variables are relevant, in your opinion, for forecasting demand for (*a*) cement, (*b*) toothpastes, (*c*) electricity and (*d*) text books?

15. What are the different techniques of survey methods? Under what conditions are complete enumeration and sample survey methods chosen?

16. Explain barometric method of demand forecasting. What is the difference between lead and lag indicators?

17. Suppose an Economic Research Centre has published data on GDP and demand for refrigerators as given below.

Year	*2000*	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*
GDP (bill. `)	20	22	25	27	30	33	35
Refrigerators (mill. units)	50	60	80	80	90	100	120

(*a*) Estimate regression equation $R = a + bY$, where R = No. of refrigerator sold and Y = GDP.

(*b*) Forecast demand for refrigerators in the years 2007 and 2008. The research centre has projected GDP for 2007 and 2008 at `38 billion and `40 billion respectively.

[**Ans.** (*a*) $R = -3.108 + 0.4145\,Y$ (*b*) R 2007 = 12.68 million, R 2008 = 13.51 million]

FURTHER READING

Ashton, DAviD and LesLie simister, (eds.), *The Role of Forecasting in Corporate Planning* (Staples Press, London, 1970), Chs. 1–3, 8–9.

BrighAm, eugene F., and JAmes, L. PAPPAs, *Managerial Economics* (The Dryden Press, Hinsdale, Illinois, 1976).

DougLAs, evAn J., *Managerial Economics: Theory, Practice and Problems* (Prentice Hall, Inc., N.J., 1970), Ch. 5.

sALvAtor, Dominick, *Managerial Economics in a Global Economy* (Thompson S.W., 7th Edn., 2012), Chs. 4 and 5.

reDDy, mehenDer J., *Demand Forecasting* (Light and Life Publications, New Delhi, 1981).

WoLFe, h.D., *Business Forecasting Methods* (Holt, Rinehart and Winston, Inc., 1966).

Part III

PRODUCTION AND COST ANALYSIS

CHAPTER 10 Theory of Production

CHAPTER OBJECTIVES

The objecitve of this Chapter is to discuss elaborately the theory of production. By going through this chapter, you will learn:

- Meaning of production
- The concept of fixed and variable inputs
- Meaning and application of production function
- Short-run laws of production – the laws of returns to variable input
- Long-run laws of production – the laws of returns of scale
- Meaning and properties of Isoquants – the tool of production analysis
- Meaning and determination of optimum combination of labour and capital
- Application of different production functions to explain the laws of returns to scale

10.1 INTRODUCTION

Once business firms take decision on 'what to produce' the major issues that arise are 'how to produce' and 'how much to produce'. These issues arise because achieving optimum efficiency in production and minimizing cost for a given production is one of the prime concerns of the business managers. In fact, the very survival of a firm in a competitive market depends on their ability to produce at a competitive cost. Therefore, managers of business firms endeavour to minimize the production cost of a given output or, in other words, maximize the output from a given quantity of inputs. In their effort to minimize the cost of production, the fundamental questions that managers are faced with are:

(*i*) How can production be optimized with given resources?

(*ii*) How does output respond to change in quantity of inputs?

(*iii*) How does technology matter in reducing the cost of production?

(*iv*) How can the least-cost combination of inputs be achieved?

(*v*) Given the technology, what happens to the rate of return when more plants are added to the firm?

The theory of production provides a theoretical answer to these questions through abstract models built under hypothetical conditions. The production theory may therefore not provide

solutions to the real life problems. But it does guide to provide tools and techniques to analyze the real-life production conditions and to find solutions to the practical business problems.

This chapter presents a detailed discussion on the theory of production. Production theory deals with quantitative relationships—technical and technological relations—between inputs (especially labour and capital) and output. Let us first discuss some basic concepts used in production analysis.

10.2 SOME BASIC CONCEPTS

This section presents a brief discussion on the basic concepts and terminology used in the analysis of the theory of production.

Meaning of Production

In general sense of the term, 'production' means transforming inputs (labour, capital, raw materials, time, etc.) into an output with value added. This concept of production is however limited to only ‘manufacturing’. In economic sense, the term ‘production’ means a process by which resources (men, material, time, etc.) are transformed into a *different* and more useful commodity or service. In other words, a process by which men, material, capital and time are converted into value added products is called *production*.

In economic sense, *production process* may take a variety of forms other than manufacturing. For example, transporting a commodity in its original form from one place to another where it can be consumed or used in the process of production is *production*. For example, a sand dealer collects and transfers the sand from the river bank to the construction site and a coal-miner does virtually nothing more than digging coal out of coal mines and transporting to the factories. Similarly, fishermen only catch and transport fishes from sea, river and fisheries to the fish market. Their activities too are ‘production’. Transporting men and materials from one place to another is also a productive activity: it produces *service*. Storing a commodity for future sale or consumption is also ‘production’. Wholesaling, retailing, packaging, assemblin g are all productive activities. These activities are just as good examples of production as manufacturing. Cultivation is the earliest form of productive activity.

Besides, production process does not necessarily involve physical conversion of raw materials into tangible goods. Some kinds of production involve an intangible input to produce an intangible output. For example, in the production of legal, medical, social and consultancy services, both inputs and outputs are intangible; lawyers, doctors, social workers, consultants, musicians, orchestra players, ‘bar girls’, etc., are all engaged in producing intangible goods.

Input and Output

An *input* is a good or service that goes into 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.”1 An *output* is any good or service that comes out of production process.

The term ‘inputs’ needs some more explanations. Production process requires a wide variety of inputs, depending on the nature of product. But, economists classified inputs as (*i*) labour, (*ii*) capital, (*iii*) land, (*iv*) raw materials and (*v*) entrepreneurship. *Technology* and *time* are also treated as inputs in the modern concept of production.

1. Baumol, W.J., *Economics Theory and Operations Analysis, op. cit.*, p. 267.

Fixed and Variable Inputs

Inputs are classified as

(*i*) *fixed inputs*, and

(*ii*) *variable inputs*.

Fixed and variable inputs are defined in economic sense and also in technical sense. The two concepts of input are explained below in economic as well as in technical sense.

Fixed Inputs In **economic sense**, a fixed input is one whose supply is inelastic in the short-run. Therefore, all of its users together cannot buy more of it in the short-run. In **technical sense,** a fixed factor is one that remains fixed (or constant) for a certain level of output.

Variable Input A ***variable input*** is defined as one whose supply in the short-run is elastic, e.g., labour and raw material, etc. All the users of such factors can employ a larger quantity in the short-run as well as in the long-run. Technically, a variable input is one that changes with the change in output. In the long-run, all inputs are variable.

Short-run and Long-run

Production of a good involves timer. 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, machinery, etc.) is fixed or is inelastic and are used in a fixed quantity. 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.

On the other hand, ***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, all the inputs are variable. Therefore, in the long-run, production of a commodity can be increased by employing more of both variable and fixed inputs.

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

The economists use another term, i.e., ***very long-run*** which refers to a period in which the technology of production is also subject to change or can be improved. In the very long-run, the production function also changes. The technological advances result in a larger output from a given quantity of inputs per unit of time.

10.3 PRODUCTION FUNCTION

Production function is a mathematical presentation of input-output relationship. More specifically, a production function states the technological relationship between inputs and output in the form of an equation, a table or a graph. In its general form, it specifies the inputs required for the production of a commodity or service. In its specific form, it states the extent of quantitative relationships between inputs and output. Besides, the production function represents the technology of a firm or of an industry. For example, suppose production of a product, say X, depends on labour (L) and capital (K), then production function is expressed in equation form as

$$Q_x = f(L, K)$$

A real-life production function is generally very complex. It includes a wide range of inputs, viz., (*i*) land and building; (*ii*) labour including manual labour, engineering staff and production manager, (*iii*) capital, (*iv*) raw material, (*v*) time, and (*vi*) technology. All these variables enter the actual production function of a firm. The long-run production function is generally expressed as

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

where LB = land and building L = labour, K = capital, M = raw materials, T = technology and t = time.

The economists have however reduced the number of variable inputs used in a production function to only two, viz., *labour* (L) and *capital* (K), for the sake of convenience and simplicity in the analysis of input-output relations. It has logical reasoning also. A production function with two variable inputs, K and L, is expressed as

$$Q = f(K, L)$$

The reasons for excluding other inputs are following.

Land and *building* (LB), as *inputs,* are constant for the economy as a whole, and hence they do not enter into the aggregate production function. However, land and building are not a constant variable for an individual firm or industry. In the case of individual firms, land and building are lumped with 'capital'.[2]

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 size 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 number of houses constructed. To consider another example, in car manufacturing of a particular brand or size, the quantity of steel, number of the engine, and number of tyres and tubes are fixed per car. Since in case of 'raw materials' the relationship between inputs and output is fixed, the output can be easily estimated given the quantity of inputs. Therefore, raw materials are left out of production function.

So is the case, generally, with *time and space*. Also, technology (T) of production remains constant over a period of time. That is why, in most production functions, only *labour* and *capital* are included.

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 mathematical form of a production function. It is this form of production function that is most commonly used in production analysis.

To illustrate the algebraic form of production function, let us suppose that a coal mining firm employs only two inputs—capital (K) and labour (L)—in its coal production activity. So the coal output depends on the number of labour and capital units employed to produce coal. As such, the general form of its production function may be expressed symbolically as

$$Q_C = f(K, L) \quad \text{...(10.1)}$$

where Q_C = the quantity of coal produced per time unit; K = capital; and L = labour.

The production function (10.1) implies that quantity of coal produced depends on the quantity of capital (K) and labour (L) employed to produce coal. Increasing coal production

2. Koutsoyiannis, A. *op. cit.*, p. 70.

will require increasing *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 a *short-run* or a *long-run*.

By definition, as noted above, short-run is a period in which supply of capital is *inelastic.* In the short-run, therefore, the firm can increase coal production by increasing only labour since the supply of capital in the short run is fixed.[3] *Long-run* is a period during which supply of both labour and capital becomes elastic, i.e., it increases over time. In the long-run, therefore, the firm can employ more of both capital and labour.

Accordingly, there are two kinds of production functions:

(*i*) *Short-run production function;* and

(*ii*) *Long-run production function.*

The two kinds of production functions are described here breifly.

(i) Short-run Production Function. A short-run production function is a *single variable* function. The single variable factor is labour (*L*), capital (*K*) remaining constant, expressed as

$Q = f(L, K)$, where *K* indicates constant capital.

…(10.2)

An estimated short-run production function may take any of the following forms depending on input-output relationship.

(*a*) Linear function: $Q = a + bL$

(*b*) Quadratic function: $Q = a + bL - cL^2$;

(*c*) Cubic function: $a + bL + cL^2 - dL^3$; or

(*d*) Power function: $Q = aL^b$ ($b > 1$).

(ii) Long-run Production Function: In the long-term production function, both *K* and *L* are treated as variable factors and the function takes the following form.

$Q = f(K, L)$

As mentioned above, a production function can be expressed in the form of an equation, a graph or a table, though each of these forms can be converted into its other forms. We illustrate here how a production function in the form of an equation can be converted into its tabular form. Consider, for example, the Cobb-Douglas production function[4]—the most famous and widely used production function—given in the form of an equation as

$Q = AK^aL^b$

…(10.3)

(where K = Capital, L = Labour, and A, a and b are parameters, and $b = 1 - a$)

Production function (10.3) gives the *general form* of Cobb-Douglas production function. The numerical values of parameters *A, a* and *b,* can be estimated by using actual factory data on production, capital and labour. Suppose numerical values of parameters are estimated as $A = 50$, $a = 0.5$ and $b = 0.5$. Once numerical values are known, the Cobb-Douglas production function can be expressed in its *specific emperical form* as follows.

3. Supply of capital may, of course, be elastic in the short-run for an individual firm under perfect competition but not for all the firms put together. Therefore, for the sake of convenience in explaining the laws of production, we will continue to assume that, in the short-run, supply of capital remains inelastic.

4. The Cobb-Douglas production function was constructed first by Paul H. Douglas in his book *The Theory of Wages* (Macmillan, NY, 1924). The function was developed further by C.W. Cobb and Paul H. Douglas in their joint paper "A Theory Production" in *Am. Eco. Rev.,* March 1928. Since then this production function is known as Cobb-Douglas Production Function.

$Q = 50K^{0.5}L^{0.5}$

This production function can be used to obtain the maximum quantity *(Q)* that can be produced with different combinations of capital *(K)* and labour *(L).* The maximum quantity that can be produced from different combinations of *K* and *L* can be worked out by using the following formula.

Q = 50 *KL* or *Q* = 50 *K L*

For example, suppose $K = 2$ and $L = 5$. Then

Q = 50 2 5 = 158

and if $K = 5$ and $L = 5$, then

Q = 50 5 5 = 250

Similarly, by assigning different numerical values to *K* and *L*, the resulting output can be worked out for different combinations of *K* and *L* and a tabular form of production function can be prepared. Table 10.1 shows the maximum quantity of a commodity that can be produced by using different combinations of *K* and *L*, both varying between 1 and 10 units.

Table 10.1 *Production Function in Tabular Form*

Table 10.1 shows the units of output that can be produced with different combinations of capital and labour. The figures given in Table 10.1 can be graphed in a three-dimensional diagram.

Before we proceed, it is ***important*** to note here that four combinations of *K* and *L* given in Table 10.1—10 *K* + 1 *L*, 5 *K* + 2 *L*, 2 *K* + 5 *L* and 1 *K* + 10 *L*—produce the same output, i.e., 158 units. When these combinations of *K* and *L* producing the same output are joined by a line, it produces a curve as shown in the table. This curve is called 'Isoquant'. An isoquant is a very important tool used to analyze input-output relationship. More follows on isoquant in section 10.5.1.

Having defined and explained the formation of production function, we now move on to explain the laws of production, first with one variable input and then with two variable inputs, i.e., the short-run and long-run laws of production, respectively. We will then illustrate the laws of production with the help of production function.

10.4 SHORT-RUN LAWS OF PRODUCTION: PRODUCTION WITH ONE VARIABLE INPUT

The laws of production state the relationship between output and input. In the short-run, input-output relations are studied with one variable input (labour), other inputs (especially, capital) held constant. The laws of production under these conditions are called the 'Laws of Variable Proportions' or the 'Laws of Returns to a Variable Input'. In this section, we explain the 'laws of returns to a variable input'.

10.4.1 The Law of Diminishing Returns to a Variable Input

The Law of Diminishing Returns. The law of diminishing returns states that *when more and more units of a variable input are used with a given quantity of fixed inputs, the total output may initially increase at increasing rate and then may be at a constant rate, but it will eventually increase at diminishing rates.* That is, the marginal increase in total output decreases eventually when additional units of a variable factor are used, given quantity of fixed factors.

Assumptions. The law of diminishing returns is based on the following assumptions:

(*i*) labour is the only variable input, capital remaining constant;

(*ii*) labour is homogeneous;

(*iii*) the state of technology is given; and

(*iv*) input prices are given.

To illustrate the law of diminishing returns, let us assume (*i*) that a firm (say, the coal mining firm in our earlier example) has a set of mining machinery as its capital (*K*) fixed in the short-run, and (*ii*) that it can employ only more mine-workers to increase its coal production. Thus, the short-run production function for the firm will take the following form.

$Qc = f(L)$, *K* constant

Now let us suppose that labour and production data of the coal mining firm are given as in cols. (1) and (2) of Table 10.2. When regression technique is applied to estimate the production function, it produces an empirical production function as given below.

$$Qc = -L^3 + 15L^2 + 10L \qquad \ldots(10.4)$$

Given the production function (10.4), we may assign different numerical values to *L* in the function and work out a series of *Qc*, i.e., the quantity of coal that can be produced with different number of workers. For example, if $L = 5$, then by substitution, we get

$$Qc = -5^3 + 15 \times 5^2 + 10 \times 5 = -125 + 375 + 50 = 300$$

What we need now is to analyse the input-output relationship with the objective of taking production decision. To this end, we need is to work out *marginal productivity of labour* (*MPL*) to find the trend in the contribution of the marginal labour and *average productivity of labour (APL)* to find the average contribution of labour.

Marginal Productivity of Labour (MPL) can be obtained by differentiating the production function (10.4). Thus,

$$MPL = \frac{\partial Q}{\partial L} = -3L^2 + 30L + 10 \quad \text{...(10.5)}$$

By substituting numerical value for labour (*L*) in Eq. (10.5), *MPL* can be obtained at different levels of labour employment. However, this method can be used only where labour is perfectly divisible and $\partial L \to 0$. Since, in our example, each unit of $L = 1$, calculus method cannot be used.

Alternatively, where labour can be increased by a minimum of one unit, *MPL* can be obtained as

$$MPL = TPL - TPL_{-1}$$

The *MPL* worked out by this method is presented in Col. 3 of Table 10.2.

Average Productivity of Labour (APL) can be obtained by dividing the production function (10.4) by *L*. Thus,

$$AP_L = \frac{-L^3 + 15L^2 + 10L}{L} = -L^2 + 15L + 10 \quad \text{...(10.6)}$$

Now *APL* can be obtained by substituting the numerical value for *L* in Eq. (10.6). *APL* obtained by this method is given in Col. 4 of Table 10.2.

Table 10.2 *Three Stages of Production*

No. of Workers (N)	*Total Product (TPL) (tonnes)*	*Marginal Product* (MPL)*	*Average Product (APL)*	*Stages of Production (based on MPL)*
(1)	*(2)*	*(3)*	*(4)*	*(5)*
1	24	24	24	I
2	72	48	36	Increasing
3	138			

66
46
returns
4
216
78
54
5
300
84
60
6
384
84
64
7
462
78
66
II
8
528
66
66
Diminishing
9
576
48
64
returns
10
600
24
60
11
594
– 6
54
III
12
552
– 42
46
Negative returns

* $MPL = TP_n - TP_{n-1}$. *MPL* calculated by differential method will be different from that given in Col. 3.

The information contained in Table 10.2 is presented graphically in panels (*a*) and (*b*) of Fig. 10.1. Panel (*a*) of Fig. 10.1 presents the total product curve (*TPL*) and panel (*b*) presents marginal product (*MPL*) and average product (*APL*) curves. The *TPL* schedule demonstrates the laws of returns to variable input (*L*). As the curve *TPL* shows, the total output increases at an increasing rate till the employment of the 5th worker, as indicated by the increasing slope of the *TPL* curve. (See also Col. 3 of the table). Employment of the 6th worker contributes as much as the 5th worker. Note that beyond the employment of the

6th worker, although *TPL* continues to increase (until the 10th worker), the rate of increase in *TPL* (i.e., *MPL*) begins to fall. This shows the operation of the law of diminishing returns.

The three stages in production.

Table 10.2 and Fig. 10.1 present the three usual stages in the application of the laws of returns to variable input.

In **Stage I**, *TPL* increases at increasing rate. This is indicated by the rising *MPL* till the employment of the 5th and 6th workers. Given the production function (10.4), the 6th worker produces as much as the 5th worker. The output from the 5th and the 6th workers represents an intermediate stage of constant returns to the variable factor, labour.

In **Stage II**, *TPL* continues to increase but at diminishing rates, i.e., *MPL* begins to decline. This

Fig. 10.1 *Total, Average and Marginal Products*

stage in production shows *the law of diminishing returns to the variable factor*. Total output reaches its maximum level at the employment of the 10th worker. Beyond this level of labour employment, *TPL* begins to decline. This marks the beginning of **Stage III** in production.

To conclude, **the law of diminishing returns** can be stated as follows. Given the employment of the fixed factor (capital), when more and more workers are employed, the return from the additional worker may initially increase but will eventually decrease.

Factors Behind the Laws of Returns. As shown in Fig. 10.1, the marginal productivity of labour (*MPL*) increases is Stage I, whereas it decreases in Stage II. In other words, in Stage I, Law of Increasing Returns to variable input is in operation and in Stage II, the law of Diminishing Returns is in application. The reasons which underly the application of the laws of returns in Stages I and II may be described as follows.

One of the important factors causing increasing returns to a variable factor is the ***indivisibility of fixed factor*** (capital). The size of a capital is given and, given the technoloyg, it requires an optimum number of labour. If labour is less than optimum number, capital cannot be divided to suit the number of workers. Therefore, if labour is less than its optimum number, capital remains underutilized. Let us suppose that optimum capital-labour combination is 1:6. If capital is indivisible and less than 6 workers are employed, then capital would remain underutilized. When more and more workers are added, utilization of capital increases. As a result, productivity of capital increases. This gives increasing returns to variable input, labour.

The *second* and the most important reason for increase in labour productivity is the ***division of labour***, i.e., assigning works to labour according to their skill. This becomes possible with the employment of additional labour. Devision of labour according to their skill and specialization increases productivity of labour.

Thus increase in productivity of both the factors, capital and labour, increases the output at increasing rate.

Once the optimum capital-labour ratio is reached, employment of additional labour amounts to underutilization of labour. So the productivity of marginal labour decreases. Also, with increasing number of workers, capital remaining the same, capital-labour ratio goes on decreasing. As a result, productivity of labour begins to decline. This marks the beginning of the second stage—the stage of decreasing returns.

10.4.2 Application of the Law of Diminishing Returns

The law of diminishing returns to the variable input is *an empirical law,* frequently observed in various production activities. This law, however, may not apply uniformally to all kinds of productive activities since it is not as true as the law of gravitation. In some productive activities, the law of diminishing returns may operate quickly, in some its operation may take a little longer time and in some others, it may not appear at all. This law has been found to operate in agricultural production more regularly than in industrial production. The reason is, in agriculture, natural factors play a predominant role whereas man-made factors play the major role in industrial production. 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 decrease eventually.

The Law of Diminishing Returns and Business Decisions. The law of diminishing returns as presented graphically has a relevance to the business decisions. The graph can help in identifying the rational and irrational stages of operations. It can also tell the business managers the number of workers (or other variable inputs) to apply to a given fixed input so that, given all other factors, output is maximum. As Fig. 10.1 exhibits, capital is presumably underutilized in Stage I. So a firm operating in Stage I is required to increase labour, and a firm operating in Stage III is required to reduce labour, with a view to maximizing its total production. From the firm's point of view, setting an output target in Stages I and III is irrational because setting output target in stage I means underutilization of capital and setting output target in Stage III means accepting negative productivity of labour. The only meaningful and rational stage from the firm's point of view is Stage II in which the firm can find answer to the question 'how many workers to employ' to maximize production, given the labour productivity. Figure 10.1 shows that the firm should employ a minimum of 7 workers and a maximum of 10 workers even if labour is available free of cost. This means that the firm has a limited choice—ranging from 7 to 10 workers. How many workers to employ against the fixed capital and how much to produce can be answered, only when the price of labour, i.e., wage rate, and product price are known. This question is answered below.

10.4.3 Optimum Employment of Labour

It may be recalled from Fig. 10.1 that an output maximizing coal-mining firm would like to employ 10 workers since at this level of employment, the output is maximum. The firm can, however, employ 10 workers only if workers are available free of cost. But

labour is not available free of cost—the firm is required to pay wages to the workers. The objective of the firm to maximize profit, not output. Therefore, the question arises as to how many workers would the firm employ—10 or less or more than 10—to maximize its profit. A simple answer to this question is that the number of workers to be employed depends on the output that maximizes the firm's profit, given the product price and the wage rate. This point can be proved as follows.

As has already been discussed in Chapter 2, profit is maximum where

$MC = MR$

The same rule can be applied to determine the optium level of employment, with some modification, of course. In our example here, since labour is the only variable input, marginal cost (MC) equals marginal wages (MW), i.e., $MC = MW$.

As regards MR, in case of factor employment, the concept of ***Marginal Revenue Productivity (MRP)*** is used. The marginal revenue productivity is the value of product resulting from the marginal unit of variable input (labour). In specific terms, marginal revenue productivity (MRP) equals marginal physical productivity (MPL) of labour multiplied by the price (P) of the product, i.e. ,

$MRP = MPL \times P$

For example, suppose that the price (P) of coal is given at `10 per quintal. Now, $MRP$ of a worker can be known by multiplying its $MPL$ (as given in Table 10.2) by `10. For example, MRP of the 3rd worker (see Table 10.2) equals 66 × 10 = `660 and of the 4th worker, 78 × 10 = `780. Likewise, if the entire column (MPL) is multiplied by `10, it gives us a table showing marginal revenue productivity of workers. Given the $MRP$ table the firm can find the labour employment at which $MRP = MW$. For example, let us suppose that wage rate (per time unit) is given at `660. Given the wage rate, the profit maximizing firm will employ only 8 workers because at this employment, MRP = wage rate = MRP of 8th worker; 66 × 10 = `660. If the firm employs the 9th worker, his $MRP$ = 48 × 10 = `480 < `660. Clearly, the firm loses `180 on the 9th worker. And, if the firm employees less than 8 workers, it will not maximize profit. This leads to a theoretical conclusion that optimum employment of labour is determined where $MRP = MW$.

Graphic Illustration

The theory of optimum employment of labour is illustrated graphically in Fig. 10.2. When

relevant series of *MRP* is graphed,
it produces an *MRP* curve as
shown in Fig. 10.2. Similarly, the
MRP curve for any input may be
drawn and compared with *MC* (or
MW) curve. Labour being the only
variable input, in our example,
let us suppose that wage rate
in the labour market is given at
OW (Fig. 10.2). When wage rate
is constant, average wage (*AW*)

Fig. 10.2 *Determination of Labour Employment*

equals the marginal wage (*MW*),

in the Short-Run

i.e., *AW* = *MW*, for the entire range of employment in the short-run. When *AW* = *MW*, the cost of labour is shown by a straight horizontal line, as shown by the line *AW* = *MW*. With the introduction of *MRP* curve and *AW* = *MW* line (Fig. 10.2), a profit maximizing firm can easily find the maximum number of workers that can be optimally employed against a fixed quantity of capital.

As Fig. 10.2 shows, the *MRP* curve and *AW* = *MW* line intersect at point *P*, where *MRP* = *MW*. Point *P*, therefore, determines the optimum level of employment of labour at *ON*. The optimum employment refers to the number of labour to be employed to maximize profit. A profit maximizing firm would therefore employ only *ON* number of workers. Given the number of workers, the total output can be known by multiplying *ON* with average labour productivity (*AP*).

10.5 PRODUCTION WITH TWO VARIABLE INPUTS

In the preceding section, we have discussed the ***short-term laws of production***, i.e., technological relationship between inputs and output assuming labour to be the only variable input, capital held constant. In this section, we proceed to discuss the **long-term laws of production,** i.e., the nature of relationship between inputs and output under the condition that both the inputs, capital and labour, are variable factors. In the long-run, supply of both the inputs is supposed to be elastic and, therefore, firms can use larger quantities of both labour and capital. With larger employment of capital labour, the scale of production increases. The nature of changing relationship between changing scale of inputs and output is refered to the ***laws of returns to scale***. The laws of returns to scale are generally explained through the *production function* and *isoquant curve* technique. The most common and simple tool of analysis is isoquant curve technique. We will, therefore, first introduce and elaborate on this tool of analysis. The laws of return to scale will then be explained through isoquant curve technique. The laws of returns to scale through production function will be explained in the next section.

10.5.1 Isoquant: The Tool of Analyses

The term ‘isoquant’ has been derived from the Greek word *iso* meaning ‘equal’ and Latin word *quantus* meaning ‘quantity’. The ‘isoquant curve’ is, therefore, also known as ‘Equal Product Curve’ or ‘Production Indifference Curve’. An ***isoquant curve*** can be defined as *the locus of points representing various combinations of two inputs—capital and labour—yielding the same output*. An ‘isoquant curve’ is analogous to an ‘indifference curve’, with two points of distinction: (*a*) an indifference curve is made of two consumer goods while an isoquant curve is constructed of two producer goods (labour and capital), and (*b*) an indifference curve represents an unmeasured level of satisfaction whereas an isoquant represents the quantity of output of a commodity.

Isoquant curves are drawn on the basis of the following assumptions:

(*i*) there are only two inputs, viz., labour (*L*) and capital (*K*), to produce a commodity *X*;
(*ii*) both *L* and *K* and product *X* are perfectly divisible;
(*iii*) the two inputs— *L* and *K*—can substitute each other but at a diminishing rate as they are imperfect substitutes; and
(*iv*) the technology of production is given.

Fig. 10.3 *Isoquant Curves*

Given these assumptions, it is technically possible to produce a given quantity of commodity *X* with various combinations of capital and labour. The factor combinations are so formed that the substitution of one factor for the other leaves the output unaffected. This technological fact is presented through an isoquant curve (*IQ* 1 = 100) in Fig. 10.3. The curve *IQ* 1 all along its length represents a fixed quantity, 100 units of product *X*. This quantity of output can be produced with a number of labour-capital combinations. For example, points *A*, *B*, *C*, and *D* on the isoquant *IQ* 1 show four different combinations of inputs, *K* and *L*, as given in Table 10.3, all yielding the same output—100 units. Note that movement from *A* to *D* indicates decreasing quantity of *K* and increasing number of *L*. This implies substitution of labour for capital such that all the input combinations yield the same quantity of commodity *X*, i.e., *IQ* 1 = 100.

Table 10.3 *Capital Labour Combinations and Output*

Points	*Input Combinations*	*Output*
	K + *L*	
A	*OK* 4 + *OL* 1	= 100
B	*OK* 3 + *OL* 2	= 100
C	*OK* 2 + *OL* 3	= 100
D	*OK* 1 + *OL* 4	= 100

10.5.2 Marginal Rate of Technical Substitution (MRTS)

An isoquant curve as explained above, shows various combinations of two inputs—labour and capital—that can technically produce the same quantity of a commodity. Given the shape of the isoquants in Fig. 10.3, the movement along the isoquants means substituting one input for the other – capital for labour or labour for capital. Given the technology, the inputs are so substituted for one another that output remains the same. The rate at which one input is substituted for the other is called the ***marginal rate of technical substitution*** *(MRTS)*. The ***MRTS*** is a very important concept used in determining the shape and properties of isoquants and also in analyzing the production with two variable inputs. Therefore, let us look at the concept of the ***MRTS*** in detail before we proceed to discuss the properties of isoquants.

In simple words, ***MRTS*** *is the rate at which a marginal unit of labour can be substituted for a marginal unit of capital or other way round, total output remaining the same.* The ***MRTS*** gives the slope of the isoquant at different levels of input combinations.

The *MRTS* is measured as follows.

$Kc - Kp - \Delta K$

MRTS =

=

= Slope of the isoquant

…(10.7)

c

$L - Lp$

ΔL

where subscript *C* means current and *p* means previous.

It is ***important*** to note that the ***MRTS*** determined the shape of the isoquants. If two inputs are *perfect substitutes,* the isoquant takes the form of straight line sloping downward to the right and in case of *imperfect substitutes*—the most general case—the isoquant takes a *curvilinear form* convex to origin, as shown in Fig. 10.3. Let us now look at the basic properties of isoquants.

10.5.3 Properties of Isoquants

Isoquants, the production indifference curves, have certain specific properties. Incidentally, Isoquants have the same properties as the consumer indifference curves. The properties of isoquants are discussed below with respect to two production inputs.

(i) Isoquants have a negative slope. The isoquants have a negative slope in its *economic* region5. An isoquant has a negative slope because of substitution between two inputs and measure of the *MRTS* carries a negative sign. For example, look at the isoquants given in Fig. 10.4.

Fig. 10.4 *Curvilinear Isoquants*

Given the isoquants, consider the movement from point *A* to point *B* on isoquant IQ_1 .

As discussed above, *MRTS* between points *A* and *B* can be measured as

MRTS = K3 – K4 / L2 – L1

$= -\Delta K_1 / \Delta L_1$

5. Economic region refers to the range of isoquant curve over which labour and capital can be substituted for one another total output remaining the same. The concept of economic regions has been discussed in detail in the next section.

Note that *MRTS* carries a negative sign because $K_3 < K_4$ as marked in Fig. 10.4. Therefore, isoquants have a negative slope. This condition holds in case of both straight line and curvilinear isoquants. It is for this reason that an isoquant slopes downward to right or upward to left.

(ii) Isoquants are convex to origin. Isoquants are convex to origin in the sense that they tend to bend towards the point of origin. Isoquants are convex to origin for two reasons: (*i*) the two inputs – capital and labour – are imperfect substitutes, and (*ii*) the returns from inputs are subject to the law of diminishing returns, i.e., as units of an input increase, its marginal productivity decreases. For these reasons, *MRTS* goes on decreasing along the isoquant. For example, as shown in Fig 10.4, in case of isoquant IQ_1 *,* ΔK s are equal, i.e., $\Delta K_1 = \Delta K_2 = \Delta K_3$ while the corresponding ΔL s substituting *K* go on increasing, i.e., $\Delta L_1 < \Delta L_2 < \Delta L_3$. It implies that the units of labour to replace a given unit of capital go on increasing . As a result, the *MRTS* decreases along the isoquant. That is,

$$\frac{\Delta K_1}{\Delta L_1} > \frac{\Delta K_2}{\Delta L_2} > \frac{\Delta K_3}{\Delta L_3} \quad \ldots(10.8)$$

The *MRTS decreases* because of the law of diminishing marginal productivity of inputs. As per the law, the marginal productivity of an input decreases when its quantity increases. It our example here, the marginal productivity of labour goes on decreasing with increase in labour and marginal productivity of capital goes on increasing with decrease in capital. Therefore, more and more units of labour are required to replace each successive unit of capital in order to maintain the level of output. Since *MRTS* decreases along the isoquant, isoquants are convex to origin.

(iii) Isoquants are non-intersecting and non-tangential. Another important feature of isoquants as a tool of analysis is that they do not intersect nor are the isoquants tangent to one another. If isoquants intersect or are tangent, the laws of production get violated as it leads to two untenable facts: (*i*) given the technology, a combination of two inputs, can produce two different quantities - larger and smaller, and (*ii*) a given quantity of a commodity can be produced with a smaller and a larger combination of inputs. This is untenable so long as marginal productivity of an input is greater than zero. These two points can be verified from Fig. 10.5.

Fig. 10.5 *Intersecting Isoquants*

Note that two isoquants intersect at point *M*— a point common to both the isoquants, 0 = 100 and $Q_2 = 200$. At point *M* the input combination is given as ML_1 of capital *plus* OL_1 of labour, i.e., ML_1(K) + OL_1 *(L)*. Since this input combination is common to both the isoquants, it means that the same input combination can produce 100 units as well as 200 units. This is technically not possible.

(iv) Upper isoquants represent higher level of output. Between any two isoquants, the upper one represents a higher level of output than the lower one. The reason is, an upper isoquant represents a larger input combination, in terms of one or both inputs and a larger input combination yields a larger output. Therefore, upper isoquant represents a higher level of output. For instance, IQ_2 in Fig. 10.6 will always indicate a higher level of output than IQ_1. For, any point at IQ_2 consists of more of either capital or labour or both. For example, consider point *a* on *IQ*

Fig. 10.6 *Comparison of Output at Two Isoquants*

1 and compare it with any point at IQ_2. The point *b* on IQ_2 indicates more of capital (*ab*) labour remaining the same; point *d* has more of labour (*ad*) capital remaining the same; and point *c* has more of both, capital and labour. Therefore, IQ_2 represents a higher level of output (200 units) than IQ_1 indicating 100 units.

10.5.4 Isoquant Map and Economic Region of Production

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 Fig. 10.7. Each isoquant shows various combinations of two inputs that can be used to produce a given quantity of output. An upper isoquant is formed by a greater quantity of one or both the inputs than the input combination 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. 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.

Economic region. Economic region is that area of production plane in which substitution between two inputs is technically feasible without affecting the output. This area is marked by locating the points on the isoquants at which *MRTS* = 0. *A zero MRTS implies that* further substitution between inputs is technically not feasible. It also determines the minimum quantity of an input that 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 Fig. 10.7. 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 (*MP*)

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

Fig. 10.7 *Isoquant Map*

quantity of output.

10.5.5 Optimal Combination of Inputs:

The Least-Cost Combinations of Inputs

A profit maximizing firm seeks to minimize its cost for a given output or to maximize its output for a given total cost. The logic of isoquant says that a given output can be produced with different input-combinations. On the other hand, given the input prices and its resources, a firm can buy many combinations of inputs. But all the combinations do not minimize the cost for a given output—only one of the input combinations conforms to the least-cost criterion. The firm is therefore required to find the combination of inputs (*K* and *L*) that minimizes the cost of production for a given output. In this section, we explain how a firm can find the least-cost combination of inputs.

To begin with, let us consider the information contained in Fig. 10.8. As the figure 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 *j*, *k*, and *l* represent three different combinations of *K* and *L*: (*i*) *OK* 3 + *OL* 1 at point *j*, (*ii*) *OK* 2 + *OL* 2 at point *k*, and (*iii*) *OK* 1 + *OL* 3 at point *l*. These three combinations can produce 100 units of *X*. Therefore, any of these combinations may be technically chosen for producing 100 units of *X*, but total cost will be different at each point. For, given the input prices—interest and wages—the total cost of production varies from point to point, and only one of the combinations at the isoquant *I* 1 gives the minimum cost, and not necessarily either *j*, *k* or *l*.

Similarly, upper isoquants represent a higher level of output that can be produced with a higher and different combinations of inputs, with varying total cost, of course.

Fig. 10.8 *Input Combination*

The firm's problem is how to find the input combination that minimizes the total cost for a given level of output.

Budgetary Constraint and Budget Line

The above problem can be solved by combining the firm's production and cost functions. The production function is represented by the isoquants. To construct the cost function, let us assume that the firm decides to incur a total cost *TC*, on both *K* and *L* and that *Pk* and *Pl* are the unit costs of *K* and *L*, respectively. Given these conditions, the firm's cost function may be expressed as

$TC = K \cdot Pk + L \cdot Pl$

...(10.9)

From Eq. (10.9), the quantity of capital, *K*, and labour, *L*, that can be purchased out of the total cost *TC*, can be easily obtained as shown below.

K

TC P

=

l

−

L and *L* = *K*

P

P

k

k

Equation 10.9 is the firm's *budget equation*. This equation yields a line, as shown in Fig. 10.9, which represents the alternative combinations of *K* and *L* that can be purchased out of the total cost *TC*. This line is known as ***isocost***. **Isocost** is also known as ***isocline, budget line,*** or the ***budget constraint line.***

The derivation of isocost is shown in Fig. 10.9. Consider the isocost $K_1 L_1$. This line is drawn on the assumption that a firm has the option of spending its total cost *TC*, either on *K* or *L*, or on both. If the firm spends its total resources on capital and nothing on labour, it would be able to buy OK_1 units of capital and the firm would be at point K_1 in Fig. 10.9. Similarly, if the firm spends its total resources on labour and nothing on capital, the firm would be at point L_1. The total capital (OK_1), with *L* = 0, and total labour (OL_1) with *K* = 0, can be worked out as follows.

TC P

OK

−

1 =

k · *L* , (where *L* = *O*)

P

P

l

l

or *OL*

TC P

1 =

k

−

K , (where *K* = *O*)

P

P

l

l

These measures of capital and labour are shown at points K_1 and L_1 respectively. The line connecting points K_1 and L_1 gives us the isocost line or the budget line. It shows the whole range of combinations of *K* and *L* that can be bought, given the total cost and factor prices.

Given the factor prices, if total cost increases, say, to $TC + \Delta TC$, larger quantities of both *K* and *L* can be bought, making the isocosts shift upwards to the right, as shown by $K_2 L_2$ and $K_3 L_3$. It is

Fig. 10.9 *Isocosts*

important to note that the slope of the isocosts (i.e., – D *K*/D *L*) gives the *marginal rate of exchanges* (*MRE*) between *K* and *L*. Since factor prices are constant, marginal rate of exchange remains constant.

Having introduced the isocosts, we may now combine isoquants and isocosts to determine the *optimal input-combination* or the *least-cost combination of inputs*. There are two conditions for the least-cost combination of inputs: (*i*) the first order or necessary condition, (*ii*) the second order or supplementary condition.

(i) The First Order Condition. The first order condition of the least-cost input-combination can be expressed in physical or quantity terms as well as in value terms. Let us discuss it first in physical terms.

Least-cost criterion in physical terms. Given the inputs, *K* and *L*, the first order or necessary condition in physical terms requires that

$$\frac{-\Delta K}{\Delta L} = \frac{MPl}{MPk} \quad \ldots(10.10)$$

where $\Delta K/\Delta L$ is the marginal rate of exchange (*MRE*) between *K* and *L*, and *MPl* / *MPk* is the ratio of marginal productivity of *L* and *K*.

In Eq. (10.10), – D *K*/D *L* = slope of the isocost, and *MPl* / *MPk* = slope of the isoquant. *It implies that the least-cost combination is given by the point where isoquant is tangent to the isocost.* The least-cost combination of *K* and *L* is graphically shown in Fig. 10.10. Let us suppose that, given the firm's total cost (*C*) and input prices, *PK* and *PL*, firm's isocost is given by the line K_2L_2 marked as $Q_2 = 200$.The isoquant $Q_2 = 200$ is tangent to isocost, K_2L_2 at point *P*. At this point, the combination of *K* and *L* equals *OM* of *K plus ON* of *L*. This combination of *K* and *L* is optimal as it satisfies the least-cost criterion, i.e., – D *K*/D *L* = *MPl* / *MPk*.The ***first order criterion*** of the least-cost input-combination can also

Fig. 10.10 *Least-Cost Combination of Inputs*

be expressed in value terms by the ratio of marginal physical-product of *K* and *L* and the ratio of input prices. Thus, the first order condition can be written as

$$\frac{MP_l}{P_l} = \frac{MP_k}{P_k} \quad \text{or} \quad \frac{MP_l}{MP_k} = \frac{P_l}{P_k} \quad \ldots(10.11)$$

where *MPl* and *MPk* are marginal products of labour and capital respectively, and *Pl* and *Pk* are prices of labour and capital respectively.

(ii) The Second Order Condition. The second order or supplementary condition requires that the first order condition be fulfilled at the highest possible isoquant. Note that the first order condition is satisfied also at points *A* and *D*, the points of intersection between $Q_1 = 100$ and isocost K_2L_2 (Fig. 10.10), as at these intersection points

—$\Delta K/\Delta L = MPl/MPk$. But, points *A* and *D* are not on the highest possible isoquant showing the highest possible output. Therefore, these points do not satisfy the second order condition. The second order condition is satisfied at point *P*. It can be seen in Fig. 10.10 that points *A*, *D* and *P* satisfy the first order condition but only point *P* satisfies the second order condition. Thus, both first and second order conditions are satisfied at point *P*. Therefore, point *P* determines the *optimum input combination*.

That point *P* determines the optimum input combination can be proved in non-technical terms. Note that while point *P* is associated with an output of 200 units, points *A* and *D*, being on a lower isoquant, are associated with an output of 100 units. It means that given the total cost, a firm can produce 100 units as well as 200 units. Therefore, a cost-minimizing firm chooses input combination at point *P* rather than at points *A* or *D*. Note that 100 units of output can be produced at a lower total cost represented by the isocost K_1L_1.

Least-cost criterion in value terms. The physical criterion can be translated in value terms by multiplying the factor exchange ratio with factor prices and marginal rate of technical substitution (*MRTS*) with product price (*P*). In fact, factor price ratios are the same as the reciprocal of factor ratios, i.e., $Pl/Pk = \Delta K/\Delta L$ and

MP

MP P

MRP

l

(

l .)

MRTS =

l

=

=

MP

MP P

MRP

…(10.12)

k

(

k .)

k

In Eq. 10.12, *MRP* = marginal revenue productivity of the actor, and = product price.

Thus, the least-cost criterion expressed in quantity terms can be converted in value terms, i.e., in terms of input and output prices as given below.

P

MRP

l

l

=

P

MRP

k

k

MRP

MRP

l

k

=

or

P

P

l

k

…(10.13)

It may be inferred from Eq. (10.13) that least-cost or optimal input combination requires that the *MRP* ratio of inputs should be equal to their price ratio.

Effect of Change in Input Price

We have shown above the determination of the least-cost combination of inputs assuming constant input prices. But, in reality, input prices do not remain constant—they continue to change intermittently. When input prices change, it necessitates change in the least-cost input-combination for cost minimization, 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, when input prices change at different rates in the same direction, or they change in the opposite direction or price of only one input changes, price of the other input remaining constant, then the relative prices of the inputs change. A change in relative input-prices changes both input-combination and the level of output. The change in input-combination results from the *substitution effect* of change in relative prices of inputs. A 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 cheaper input for the costlier one. This is known as the *substitution effect* of change in the relative input prices.

The effect of change in input prices on the input-combination is illustrated in Fig. 10.11. Suppose given the *Pk* and *Pl*, and the total cost, firm's isocost is given by *KL* and the firm's optimum input-combination is given by point *E*, in Fig. 10.11. Let us suppose that *Pl* decreases (*Pk* remaining constant) so that the isocost *KL* rotates to the position of isocost *KW*, that is tangent to isoquant *I* 2 at point *N*. At this point, firm's new optimum combination of inputs is *OK* 1 + *OL* 3. Thus, as a result of decrease in *Pl*, the firm reduces its *K* by *K* 1 *K* 2 and increases *L* by *L* 1 *L* 3. This change in input combination is *price effect*. The price effect combines *substitution* and *budget effects*. The price and budget effects can be separated in the following manner.

In order to measure the budget effect first, let us find out how much additional labour the firm will employ if its resources increase so that the firm reaches the isoquant *I* 2, *input prices remaining the same*. This can be established by drawing an isocost parallel to *KL* and tangent to *I* 2, as shown by isocost *K*¢ *L*¢. The isocost *K*¢ *L*¢ is tangent to isoquant *I* 2 at point *M*.

Fig. 10.11 *Substitution Effect and*

It means that if *P*

Input Combination

k and *Pl* remain constant and firm's budgetary resources increase, it will settle at point *M* and its optimum input-combination will be *OK* 3 of *K* + *OL* 2 of *L*. This combination may be said to have resulted from the *budget effect* or resources effect, or the output effect. If we deduct the *budget effect* on labour from the price effect, we get the *substitution effect*, i.e.,

Substitution effect = Price effect – Budget effect

Since price effect = *L* 1 *L* 3, and budget effect = *L* 1 *L* 2, and

Substitution effect = *L* 1 *L* 3 – *L* 1 *L* 2 = *L* 2 *L* 3

Thus, we find that as a result of change in price of an input, input combination of the firm changes: the firm employs more of cheaper input (*L*) and less of the costlier one (*K*). Besides, the level of output also changes. If price of an input decreases, the level of output increases, and *vice versa*.

This concludes our brief discussion on the traditional production theory, production function, laws of variable proportions, law of returns to scale, and the choice of least-cost input combination. These aspects have been explained in physical terms—physical quantities of inputs and outputs. In the next chapter, we will discuss the theory of cost—the monetary aspects of production theory.

10.5.6 Elasticity of Factor Substitution

We have explained earlier (section 10.5.2) the concept of the marginal rate of technical substitution (*MRTS*) and have noted that *MRTS* decreases along the isoquant. The *MRTS* refers only to the slope of an isoquant, i.e., the ratio of marginal changes in inputs. It does not reveal the substitutability of one input for another—labour for capital or otherwise with changing combination of inputs due to change in input prices. The economists have devised a method of measuring the degree of substitutability of factors, called the ***Elasticity of Factor Substitution***. The elasticity of substitution (σ) is formally defined as *the ratio of the percentage change in the capital-labour ratio (K/L) to the percentage change in marginal*

Fig. 10.12 *Graphic Derivation of Elasticity of Substitution*

rate of technical substitution (*MRTS*).

It is measured as follows.

$$\sigma = \frac{\text{Percentage change in } K/L}{\text{Percentage change in } MRTS}$$

or

$$\sigma = \frac{\partial(K/L)/(K/L)}{\partial(MRTS)/(MRTS)} \quad \ldots(10.14)$$

Since all along an isoquant, capital-labour ratio (*K/L*) and *MRTS* change in the same direction, the value of σ is always positive. Besides, the elasticity of factor substitution (σ) is ‘a pure number, independent of the units of the measurement of *K* and *L*, since both the numerator and the denominator are measured in the same units’.

The concept of elasticity of factor substitution, as given in Eq. (10.14), is graphically presented in Fig. 10.12. The movement from point *A* to *B* on the isoquant *IQ*, gives the ratio of change in *MRTS*. The rays *OA* and *OB* represent two techniques of production with different factor intensities. While line *OA* indicates capital intensive technique, line *OB* indicates labour intensive technique. The shift from *OA* to *OB* gives the change in factor intensity. The ratio between the two factor intensities measures the *elasticity* of substitution.

The value of substitution elasticity depends on the curvature of the isoquants. It varies between 0 and ∞, depending on the nature of production function. It is, in fact, the production function that determines the curvature of the various kinds of isoquants. For example, in case of fixed-proportion production function yielding *L*-shaped isoquant, σ = 0. If production function is such that the resulting isoquant is linear, σ = ∞ . And, in case of a homogeneous production function of degree 1 of the Cobb-Douglas type, σ = 1.

10.5.7 Expansion Path: The Path of Increasing Production Scale

The firms in general tend to expand their production in the long run. A cost-minimizing firm expands its production by ensuring the optimum combination of inputs. The root through which firm expands in cost minimizing output is called the ***expansion path.*** The *expansion path* may be defined as the locus of points indicating various cost minimizing combination of inputs at the different levels of production. The derivation of the *expansion path* has been discussed in two section.

In Section 10.5.5, we have discussed the conditions for optimum (least-cost) combination of inputs and have illustrated optimum input combination in Fig. 10.10. The optimum combination of inputs is reproduced in Fig. 10.13. Given the firm's original budget line as K_1L_1 and isoquant as IQ_1, the firm finds its optimum combination of inputs at point *A,* where $MRTS = Pk/PL$. At point *A,* the optimum combination of input is determined at *AL* of capital and *OL* of labour. This input combination is based on the assumption that the firm has a limited resource (investible fund) to be spent on labour and capital, given by the budget line K_1L_1. In the long run, however, the firm can manage larger investible funds and hire more of labour and capital to increase its scale of production.

When firm's resources increase, input prices remaining constant, its budget line shifts upward from K_1L_1 to K_2L_2, remaining parallel to the previous budget line. As a result, firm's production level shifts to a higher isoquant and equilibrium point shifts from point *A* to point *B.* When the firm invests more, its budget line shifts further upward, say, to K_3L_3, and it moves an upper isoquant IQ_3. Its optimum combination of inputs to

Fig. 10.13 *Expansion Path*

maximize its output shifts from point *B* to point *C*. As Fig. 10.13 shows, as firm's resources increase, the firm's production level goes on shifting from lower isoquants to upper isoquants. This indicates expansion of production. By joining equilibrium points *A, B* and *C,* one gets the ***expansion path.*** The expansion path shows graphically the root of increase in production with cost minimization.

10.5.8 Other Forms of Isoquants: A Digression

We have discussed above the meaning, purpose and features of convex isoquants—the most widely used tool of production analysis in traditional economic theory. The shape of an isoquant, however, depends on the degree of substitutability between the factors in the production function. A *convex isoquant* assumes a continuous substitutability between capital and labour but at a diminishing rate. The economists have, however, observed other degrees of substitutability between *K* and *L* and have demonstrated the existence of three other kinds of isoquants.

1. Linear Isoquants. In case two inputs—labour and capital—are perfect substitutes for one another then the isoquant takes a linear form. A linear isoquant is presented by the line *AB* in Fig. 10.14. A linear isoquant implies perfect substitutability between the two inputs, *K* and *L*. The isoquant *AB* indicates that a given quantity of a product can be produced by using only capital or only labour or by using both.

A linear isoquant also implies that the *MRTS* between *K* and *L* remains *constant* throughout.

The mathematical form of the production function exhibiting perfect substitutability of factors is given as follows. Suppose a

Fig. 10.14 *Linear Isoquant*

production function is given as

$Q = aK + bL$

...(10.15)

The production function (10.15) means that the total output, Q, is simply the weighted sum of K and L. The slope of the isoquant resulting from this production function is given by – b/a. This can be proved as shown below.

Given the production function (10.15),

∂Q

∂Q

MP

$= a$

$= b$

$k = \partial K$

and $MPL = \partial L$

MP

$- b$

Since

$MRTS = MPL$ and

$L =$

MP

MP

a

K

K

$- b$

Therefore,

$MRTS = a$ = slope of the isoquant

The production function exhibiting perfect substitutability of factors, however, may be rarest of the rare cases in the real world production process.

2. L-Shaped Isoquants: Isoquants with Fixed Factor-Proportion. When a production function assumes a fixed proportion between K and L, the isoquant takes 'L' shape, as shown by isoquants Q 1 and Q 2 in Fig. 10.15. Such an isoquant implies zero substitutability between K and L. Instead, it assumes perfect complementarity between K and L. The *perfect complementarity* 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. It also implies that if the quantity of an input is increased and the quantity of the other input is held constant, there will be no change in output. The output can be increased only by increasing both the inputs proportionately. For example, if output is to be increased to Q 2, K has to be increased by K 1 K 2 and labour by L 1 L 2.

It means that if output has to be doubled, then both the inputs have to be doubled. The line OB shows that there is only one factor combination for a given output. This kind of technological relationship between K and L gives a fixed proportion production function.

A ***fixed-proportion production function***, called Leontief function, is given as

$Q = f(K, L) = min\ (aK, bL)$

...(10.16)

where '*min*' means that Q equals the lower of the two terms, aK and bL. That is, if $aK > bL$, $Q = bL$ and if $bL > aK$, then $Q = aK$. 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 a linear production function, the fixed-factor-proportion production function has a wide range of application in the real world. One can find many techniques of production in which a fixed proportion of labour and capital is fixed. For example, to run a taxi or to operate a photocopier, one needs only one labour. 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.

3. Kinked Isoquants or Linear Programming

Isoquants. The fixed proportion production

Fig. 10.15 *The L-Shaped Isoquant*

function and resulting isoquant as shown in Fig. 10.15 assumes that there is only one technique of production. 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 machinery available to produce a commodity. Each machine requires a fixed number of to make its full utlization. This number varies from machine to machine. For example, 40 persons can be transported from one place to another by two methods: (*i*) by hiring 10 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. Handlooms and power looms are other examples of two different factor proportions. One can find many such technology of production in manufacturing industries.

Let us suppose that for producing 10 units of a commodity, *X*, there are four different techniques of production available. Each techniques has a different fixed factor-proportion, as given in Table 10.4.

Table 10.4 *Alternative Techniques of Producing 100 Units of X*

S. No.	*Technique*	*Capital*	+	*Labour*	*Capital/labour ratio*
1	*OA*	10	+	2	10:2
2	*OB*	6	+	3	6:3
3	*OC*	4	+	6	4:6
4	*OD*	3	+	10	3:10

The four hypothetical production techniques, as presented in Table 10.4, are presented graphically in Fig. 10.16. The ray *OA* represents a production process having a fixed factor-proportion of 10 *K*:2 *L*. Similarly, the other three production processes having fixed

capital-labour ratios 6 *K*:4 *L*, 4 *K*:6 *L* and 3 *K*:10 *L* have been shown by the rays *OB*, *OC* and *OD* respectively. Points *A*, *B*, *C* and *D* represent four different production

techniques. By joining the points,
A, *B*, *C* and *D*, we get a **kinked**
isoquant, *ABCD*.

Each of the points on the
Kinked Isoquant represents a
combination of capital and labour
that can produce 100 units of
commodity *X*. If there are other
processes of production, many
other rays would be passing
through different points between

Fig. 10.16 *Fixed Proportion Techniques*

A and *B*, *B* and *C* , and *C* and *D*,

of Production

increasing the number of kinks on
the isoquant *ABCD*. The resulting isoquant would then resemble the typical isoquant.
The kinked isoquant is used basically in linear programming. It is, therefore, also
called ***linear programming isoquant*** or *activity analysis isoquant*.

10.6 LAWS OF RETURNS TO SCALE

Having introduced the isoquants—the basic tool of analysis—we now return to the theory of long-run production, specifically the laws of returns to scale. The laws of returns to scale explain the nature of changed in output in response to a proportional and simultaneous change in inputs. Increasing inputs proportionately and simultaneously is, in fact, an expansion of the scale of production.

When a firm expands its scale of production, i.e., it increases both the inputs in a certain proportion, then there are three technical possibilities of increase in production:

(*i*) Total output may increase more than proportionately;

(*ii*) Total output may increase proportionately; and

(*iii*) Total output may increase less than proportionately.

Accordingly, there are three kinds of laws of returns to scale:

(*i*) The law of increasing returns to scale;

(*ii*) The law of constant returns to scale, and

(*iii*) The law of diminishing returns to scale.

So far as the sequence of the laws of 'returns to scale' is concerned, the law of increasing returns to scale is followed by the law of constant returns and then by the law of diminishing returns to scale. This is the most common sequence of the laws.

Let us now explain the laws of returns to scale through isoquants for a two-input and single output production system.

10.6.1 Law of Increasing Returns to Scale

When inputs, *K* and *L*, are increased at a certain proportion and output increases more than proportionately, it exhibits the law of *increasing returns to scale*. For example, if quantities of both the inputs, *K* and
L, are successively doubled and the
resultant output is more than doubled,
then the law of returns to scale is said
to be in operation. The increasing
returns to scale is illustrated in Fig.
10.17. In Fig. 10.17, lines *OB* and
OC represent the *expansion path*. The
movement from point *a* to *b* on the
line *OB* means doubling the inputs. It
can be seen in Fig. 10.17 that input-
combination increases from 1 *K* + 1 *L*
to 2 *K* + 2 *L*. As a result of doubling
the inputs, output is more than
doubled: it increases from 10 to 25

Fig. 10.17 *Increasing Returns to Scale* units, i.e., an increase of more than double. Similarly, the movement from point *b* to point *c* indicates 50% increase in inputs as a result of which the output increases from 25 units to 50 units, i.e., by 100%. Clearly, output increases more than the proportionate increase in inputs. This kind of relationship between the inputs and output exemplifies the law of *increasing returns to scale*.

The Factors Behind Increasing Returns to Scale

The factors that lead to increasing returns to scale are known as *internal economics of scale* [6]. There are at least three plausible factors causing increasing returns to scale.

(i) Technical and managerial indivisibilities. Certain inputs, particularly mechanical equipments and managers, used in the process of production are available in a given size. Such inputs cannot be divided into parts to suit small scale of production. For example, half a turbine cannot be used and one-third or a part of a composite harvester and earth-movers cannot be used to produce the relevant product.

In regard to managerial manpower, if scale of production is small, managers remain under-employed because half of a production manager cannot be employed, if part-time employment is not acceptable to the manager. Because of indivisibility of machinery and managers, given the state of technology, they have to be employed in a minimum quantity even if scale of production is much less than the capacity output. Therefore, when scale of production is expanded by increasing all the inputs, the productivity of indivisible factors increases exponentially because of technological advantage. This results in increasing returns to scale.

(ii) 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 suitable to a particular job and

6. The ‘economies and diseconomies of scale’ are discussed in detail in the next chapter.

of a composite machinery increases productivity of both labour and capital per unit of inputs. Their cumulative effects contribute to the increasing returns to scale. Besides, employment of specialized managerial personnel, e.g., administrative manager, production managers sales manager and personnel manager, contributes a great deal in increasing production.

(iii) Dimensional relations. Increasing returns to scale is also a matter of dimensional relations. For example, when the length and breadth of a room ($15' \times 10' = 150$ sq. ft.) are doubled, then the size of the room is more than doubled: it increases to $30' \times 20' = 600$ sq. ft. When diameter of a pipe is doubled, the flow of water is more than doubled. In accordance with this dimensional relationship, when the labour and capital are doubled, the output is more than doubled and so on.

10.6.2 Law of Constant Returns to Scale

When the increase in output is proportionate to the increase in inputs, it exhibits *constant returns to scale*. For example, if quantities of both the inputs, *K* and *L*, are doubled and output is also doubled, then the returns to scale are said to be constant. Constant returns to scale are illustrated in Fig. 10.18. The lines *OA* and *OB* are ‘expansion paths’ indicating two hypothetical techniques of production with optimum capital-labour ratio. The isoquants marked $Q = 10$, $Q = 20$ and $Q = 30$ indicate the three different levels of output. In the figure, the movement from points *a* to *b* indicates doubling both the inputs—increasing capital from 1 *K* to 2 *K* and labour from 1 *L* to 2 *L*. When inputs are doubled, output is also doubled, i.e., output increases from 10 to 20.

Fig. 10.18 *Constant Returns to Scale*

Similarly, the movement from point *b* to *c* indicates a 50 per cent increase in both labour and capital. This increase in inputs results in an increase of output from 20 to 30 units, i.e., a 50 per cent increase in output. In simple words, a 50 per cent increase in inputs leads to a 50 per cent increase in output. This relationship between a proportionate change in inputs and the same proportional change in outputs may be summed up as follows.

1 *K* + 1 *L* ^ 10

2 *K* + 2 *L* ^ 20

3 *K* + 3 *L* ^ 30

This kind of relationship between inputs and output exhibits ***constant returns to scale.***

The constant returns to scale are attributed to the **limits of the economies of scale**. With expansion in the scale of production, economies arise from such factors as indivisibility of fixed factors, greater possibility of speciali-zation of capital and labour, use of more efficient techniques of production, etc. But there is a limit to the economies of scale. When economies of scale reach their limits and diseconomies are yet to begin, returns

to scale become constant. The constant returns to scale take place also where factors of production are perfectly divisible and where technology is such that capital-labour ratio is fixed. When the factors of production are perfectly divisible, the production function is homogeneous of degree 1 showing constant returns to scale.

10.6.3 Decreasing Returns to Scale

The firms are faced with *decreasing returns to scale* when a certain proportionate increase in inputs, *K* and *L*, leads to a less than proportionate increase in output. For example, when inputs are doubled and output is less than doubled, then decreasing returns to scale is in operation. The decreasing returns to scale is illustrated in Fig. 10.19. As the figure shows, when the inputs *K* and *L* are doubled, i.e., when capital-labour combination is increased from 1 *K* + 1 *L* to 2 *K* + 2 *L*, the output increases from 10 to 18 units. This means that when capital and labour are increased by 100 per cent, output increases by only 80 per cent. That is, increasing output is less that the proportionate increase in inputs. Similarly, movement from point *b* to *c* indicates a 50 per cent increase in the inputs. But, the output increases by only 33.3 per cent. This exhibits *decreasing returns to scale*.

Fig. 10.19 *Decreasing Return to Scale*

Causes of Diminishing Return to Scale

The decreasing returns to scale are attributed to the **diseconomies of scale**. The economists find that the most important factor causing diminishing returns to scale is ‘the diminishing return to management’, i.e., managerial diseconomies. As the size of the firms 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 plant may not double the coal output because of limitedness 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 in the ocean when fishing is carried out on an increased scale.

10.7 LAWS OF RETURNS TO SCALE THROUGH PRODUCTION FUNCTION

The laws of returns to scale may be explained more precisely through a production function. Let us assume a production function involving two variable inputs (*K* and *L*) and one commodity *X*. The production function may then be expressed as

$$Qx = f(K, L) \qquad \text{...(10.17)}$$

where Qx denotes the quantity of commodity *X*.

Let us also assume that the production function is homogeneous. A production function is said to be homogeneous when all the inputs are increased in the same proportion and the proportion can be factored out. And, if all the inputs are increased by a certain proportion (say, k) and output increases in the same proportion (k), then production is said to be homogeneous of degree 1. This kind of production function may be expressed as follows.

$kQx = f(kK, kL)$

...(10.18)

or

$= k(K, L)$

A ***homogeneous production function*** of degree 1, as given in Eq. (10.18), implies ***constant returns to scale.*** Eq. (10.18) shows that increase in inputs, K and L, by a multiple of k, increases output, Qx by the same multiple (k). This means constant returns to scale.

The constant returns to scale may not be applicable at all the levels of increase in inputs. Increasing inputs K and L in the same proportion may result in increasing or diminishing returns to scale. In other words, it is quite likely that if all the inputs are increased by a certain proportion, output may increase more or less than proportionately. For example, if all the inputs are doubled, the output may increase by less than or more than double. Then the production function may be expressed as

$hQx = f(kK, kL)$

...(10.19)

where h denotes h-times increase in Qx, as a result of k-times increase in inputs, K and L. The proportion h may be greater than k, equal to k, or less than k. Accordingly, it reveals the three laws of returns to scale:

(*i*) If $h = k$, production function reveals constant returns to scale.

(*ii*) If $h > k$, it reveals increasing returns to scale.

(*iii*) If $h < k$, it reveals decreasing returns to scale.

This aspect has been elaborated in the following section.

10.7.1 Degree of Production Function and Returns to Scale

In case of a homogeneous production function of degree 1 [Eq. (10.18)], k has an exponent equal to 1, i.e., $k = k^1$. It means that if k has an exponent equal to 1, the production function is homogeneous of degree 1. But, a production function may not be homogeneous of degree 1. It may be homogeneous of a degree less than 1 or greater than 1. It means that the exponent of k may be less than 1 or greater than 1. Let us assume that exponent of k is r, where r π 1. A production function is said to be of degree r when all the inputs are multiplied by k and output increases by a multiple of k^r. That is, if

$f(kK, kL) = k^r(K, L) = k^r Q$

...(10.20)

then function (10.20), is homogeneous of degree r.

From the production function (10.20), we can again derive the laws of returns to scale.

(*i*) If $k > 1$ and $r < 1$, it reveals decreasing returns to scale;

(*ii*) If $k > 1$ and $r > 1$, it reveals increasing returns to scale; and

(*iii*) If $k > 1$ and $r = 1$, it means constant returns to scale.

For example, consider a multiplicative form of production function, i.e.,

$Q = K^{0.25} L^{0.50}$

...(10.21)

If K and L are multiplied by k, and output increases by a multiple of h then

$hQ = (kK)^{0.25} (kL)^{0.50}$

By factoring out k, we get

$hQ = k^{0.25+0.50} [K^{0.25} L^{0.50}]$

$= k^{0.75} [K^{0.25} L^{0.50}]$

...(10.22)

In Eq. (10.22), $h = k^{0.75}$ and $r = 0.75$. This means that $r < 1$ and, thus, $h < k$. Production function (10.21), therefore, shows ***decreasing returns to scale.***

Now consider another production function given as

$Q = K^{0.75} L^{1.25} X^{0.50}$

...(10.23)

If K, L and X are multiplied by k, Q increases by a multiple of h then the production function can be expressed as

$hQ = (kK)^{0.75} (kL)^{1.25} (kX)^{0.50}$

By factoring out k, we get

$hQ = k^{(0.75+1.25+0.50)} [K^{0.75} L^{1.25} X^{0.50}]$

$= k^{2.5} [K^{0.75} L^{1.25} X^{0.50}]$

Here $h = k^{2.5}$ where $2.5 = r$. It means $r > 1$. So $h > k$. Therefore, function (10.23) gives *increasing returns to scale*. Similarly, if in a production function, $h = k^r$ and $r = 1$, the production function gives *constant returns to scale*.

10.7.2 Cobb-Douglas Production Function—

The Multiplicative Power Function

In the preceding section, the laws of returns to scale have been illustrated through a normal and a power function. In this section, we show the application of the most widely used production functions – the multiplicative *power function*. The most popular production function of this category is '*Cobb-Douglas Production Function* 7 of the form

$$Q = AK^a L^b \qquad \ldots(10.24)$$

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 following form.

$$Q = AK^a L^{1-a} \qquad \ldots(10.25)$$

Properties of Cobb-Douglas Production Function. A power function of this kind has several important properties.

First, the multiplicative form of the power function (10.16) can be changed into its log-linear form as

$$\log Q = \log A + a \log K + b \log L \qquad \ldots(10.26)$$

7. This production function, widely referred to in economic literature, was first constructed by Paul H. Douglas in his book, *The Theory of Wages* (Macmillan, N.Y., 1924). It was developed further by C.W. Cobb and P.H. Douglas in their paper "A Theory of Production", *Am. Eco. Rev.,* March 1928 (Suppl.) and was used by P.H. Douglas, 20 years later in his paper "Are There Laws of Production", *Am. Eco. Rev.* , March 1948.

In its logarithmic form, the function becomes simple to handle and can be empirically estimated using linear regression analysis.

Secondly, power functions are homogeneous and the degree of homogeneity is given by the sum of the exponents a and b. If $a + b = 1$, then the production function is homogeneous of degree 1 and implies ***constant returns to scale***.

Thirdly, parameters a and b represent the elasticity co-efficient of output for inputs K and L, respectively. The output elasticity co-efficient (ϵ) in respect of capital may be defined as proportional change in output as a result of a given change in K, keeping L constant. Thus,

$$\epsilon_k = \frac{\partial Q / Q}{\partial K / K} = \frac{\partial Q}{\partial K} \cdot \frac{K}{Q} \qquad \ldots(10.27)$$

By differentiating the production function $Q = AK^aL^b$ with respect to K and substituting the result in Eq. (10.27), we can find the elasticity co-efficient. We know that

$$\frac{\partial Q}{\partial K} = a AK^{a-1} L^b$$

By substituting the values for Q and $\partial Q/\partial K$ in Eq. (10.2), we get

$$\epsilon_k = a AK^{a-1} L^b \left(\frac{K}{AK^a L^b} \right) = a \qquad \ldots(10.28)$$

Thus, output-elasticity coefficient for K is ' a'. The same procedure may be adopted to show that b is the elasticity co-efficient of output for L.

Fourthly, constants a and b represent the relative share of inputs, K and L, in total output Q. The share of K in Q is given by

$\partial Q \cdot K$

∂K

Similarly, the share of L in Q is given by

$\partial Q \cdot L$

∂L

The relative share of K in Q an e obtained as

$a -1$

∂Q

1

b

$a\ AK\ L\ .\ K$

$K \cdot =$

$= a$

$a\ b$

∂L

Q

$AK\ L$

Similarly, it can be shown that b gives the relative share of L in Q.

Finally, Cobb-Douglas production function in its general form, $Q = Ka\ L\ 1– a$ implies

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:

$APL = A\ (K/ L)1– a$

$APK = A\ (L/ K)1$

(*ii*) Marginal Product of L and K:

$MPL = a.\ A\ (K/ L)\ a = a\ (Q/ L)$

$MPK = (a – 1)\ A\ (L/ K)\ a = (1 – a)\ Q/ K$

(*iii*) Marginal Rate of Technical Substitution:

MP

⎡ a

K

L

$MRTS$

⎤

=

=

.

L, k

MP

⎢⎣ $a\ L$ ⎥

−

K

1

⎦

10.7.3 CES Production Function

In addition to the Cobb-Douglas production function, there are several other forms of

production function, viz., 'constant elasticity substitution' (*CES*), 'variable elasticity

of substitution' (*VES*), Leontief-type, and linear-type. Of these, the *constant elasticity*

substitution (*CES*) production function8 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

$Q = A[\alpha\ K–\beta + (1 – \alpha)\ L–\beta]–1/\beta$

... (10.29)

or

$Q = A[\alpha L^{-\beta} + (1 - \alpha) K^{-\beta}]^{-1/\beta}$

($A > 0$, $0 < a < 1$, and $b > -1$)

where *L* = labour, *K* = capital, and *A*, a and b are the three parameters.

An important property of the *CES* production function 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 Eq. (10.29) can be written as follows.

$Q' = A[a(mK)^{-b} + (1 - a)(mL)^{-b}]^{-1/b}$

$= A[m^{-b}\{aK^{-b} + (1 - a)L^{-b}\}]^{-1/b}$

$= (m^{-b})^{-1/b} \times A[aK^{-b} + (1 - a)L^{-b}]^{-1/b}$

...(10.30)

As given in Eq. (10.29) the term $A[aK^{-b} + (1 - a)L^{-b}]^{-1/b} = Q$, By substitution, therefore, we get

$Q' = mQ$

Thus, the *CES* production function is homogeneous of degree 1.

Given the production function (10.29), the marginal product of capital (*K*) can be obtained as

1

δ *Q*

α ⎡ *Q* β+

⎤

=

.

K

*A*β ⎢ *K* ⎥

δ

⎣ ⎦

and of labour (*L*) as

8. *CES* production function was constructed by K. Arrow, H.B. Chenery, B.S. Minhas, and R.M. Solow, “Capital Labour Substitution and Economic Efficiency”, *Review of Economics and Statistics*, Vol. 43 (August 1961). This production function is also sometimes called *SMAC* production function.

1

δ *Q* 1− α ⎡ *Q* β+

⎤

=

.

K

*A*β ⎢ *L* ⎥

δ

⎣ ⎦

The rate of technical substittion (*RTS*) can be obtained as

β 1

α ⎡ *L* +

RTS

⎤

= 1 α ⎢ *K* ⎥

− ⎣ ⎦

Merits of CES Production Function. *CES* production function has certain advantages over the other functions:

(*i*) it is a more general form of production function;

(*ii*) it can be used to analyze all types of returns to scale, and

(*iii*) it removes many of the problems involved in the Cobb-Douglas production function.

Limitations. The *CES* production function has, however, its own *limitations*. Some economists claim that it is not a general form of production function as it does not stand the empirical test. In other words, it is difficult to fit this function to empirical data.

Also, 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 function does not provide a measure to separate the effects on the productivity of inputs.

SUMMARY

• Theory of production deals with input-output relationship. Prior to discussing the theory of production, it is important to understand the concepts used in production analysis.

• Production means transforming inputs into output with value added. Production without value addition is not production – it is only waste of resources.

• Inputs can be anything that can be used in the process of producing goods and services. In reality, inputs include all factors of production, viz., labour, capital, land, raw materials, and entrepreneurship. In theoretical analysis of input-output relationship, however, only two inputs – labour and capital – are treated as inputs.

• Production function is a mathematical expression of relationship between inputs and output. The general form of production function is expressed as $Q = f(L, K)$, where L = labour, and K = capital

• Depending on time perspective, a production function may be a short-run or a long-run production function. Short-run production function is expressed as $Q = f(L)$, capital (K) treated as a constant input. Long-run production function is expressed as $Q = f(L, K)$, both inputs L and K are treated as variable inputs.

• A short-run production function may take a form of (*i*) a liner function: $Q = a + bL;$ (*ii*) a quadratic function: $Q = a + bL - cL2$, or (*iii*) a cubic function: $Q = a + bL + cL2 - dL3$. A long-term production function is given by Cobb-Douglas production function as $Q = AKaLb$, where $b = 1 - a$.

• Given the nature of production function, there are two kinds of laws of production: (*i*) the law of diminishing returns – the short term law of production, and (*ii*) the laws of returns to scale – the long term laws of production.

• The law of diminishing returns states that, capital remaining constant, if labour is increased continuously, the resulting output may increase initially at increasing rate but ultimately at decreasing rate. Output increases initially at increasing rate because of (*i*) better utilization of indivisible capital, and (*ii*) division of labour according to their skill.

• The long-run laws of production, i.e., the laws of returns to scale, refer to input-output relation under the condition that both the inputs (*L* and *K*) increasing proportionately and simultaneously. The long-run input-output reflect three kinds of laws of returns to scale: (*i*) increasing to returns to scale, (*ii*) constant returns tom scale, and (*iii*) decreasing returns to scale.

• The laws of returns to scale are illustrated by using isoqant technique. An isoquant is a locus of points showing different combination of inputs yielding the same output. The properties of isoquants are (*i*) isoquants are convex to origin due to diminishing marginal rate of technical substitution (*MRTS*) between the inputs the rate at which one input is substituted for other at margin.

• Given their resources and factor prices, firm's budget line can be drawn to show the various combinations the firm can acquire. Given the isoquants and firm's budget line, firms find the optimum combination of inputs where *MRTS = factor price ratio = w/r = wages/capital rentals.*

• Once optimum combination of inputs is obtained, firms can increase labour and capital to increase their production. The response of output to increase in inputs is presented under the laws of returns to scale. As noted above, economists have disclosed three kinds of laws of returns to scale: (*i*) increasing to returns to scale, (*ii*) constant returns tom scale, and (*iii*) decreasing returns to scale.

• The law of increasing returns to scale can be stated as when firms increase inputs simultaneously and proportionately, output increases more than proportionately, i.e., output increases at increasing rates. Output increases at increasing rate because of economies of scale which consist of (*a*) output gains from increasing divisibility of technology and management, (*b*) advantage of specialization, and (*c*) dimensional complementarity of factors.

• The law of constant returns to scale can be stated as when inputs are increased in a certain proportion, output increases in the same proportional – it means at a constant rate. The law of constant returns to scale comes into operation when the benefits of internal and external economies tend to become constant.

• In case of the law of decreasing returns to scale, total output increases at decreasing rates with constant increase in inputs. The law of decreasing returns to scale comes into operation when economies of scale tent to decline.

• The laws of returns to scale can be explained by using the Cobb-Douglas production function. (For details, see section 10.7).

REVIEW QUESTIONS

1. What is meant by production? Distinguish between fixed and variable inputs.

2. Define production function. How is production function formulated? What purpose does a production function serve in production analysis?

3. State and illustrate Cobb-Douglas production function. What are the properties of Cobb-Doulas production function?

4. How is short run and long run determined in the context of production of a commodity? How are short-run laws of production different from the long-run laws of production?

5. Explain the laws of returns to variable proportions. What are the factors that cause increasing and decreasing returns to a variable input?

6. (*a*) What is meant by the marginal revenue product of a variable input?

(*b*) How is the optimum level of employment of a variable input determined?

7. Distinguish between short-run and long-run production function. Suppose a production function is given as $Q = 10\,Z, + 15\,L2 - L\,3$.

(*a*) What law of production is indicated by this production function?

(*b*) What number of labour will maximize the output?

8. What is meant by the laws of returns to scale? What are the factors responsible for increasing returns to scale? What are the reasons for decreasing returns to scale?

9. What is meant by isoquant? How are isoquants different from indifference curves? What are the properties of isoquants?

10. What is meant by marginal rate of technical substitution? Why does marginal rate of technical substitution decrease along an isoquant convex to origin?

1l. What are the technical conditions for the least-cost combination of inputs? Explain and illustrate the determination of the least-cost combination of inputs. Show also the change in least-cost combination of inputs with change in the prices of inputs.

12. What is meant by optimum combination if inputs? What are the technical conditions for the optimum combination of inputs? Explain and illustrate the determination of optimum combination of inputs by using isoquants and isocosts.

13. What is meant by economic region? What determines the scope of economic region? How is economic region determined in an indifference map?

14. What is meant by laws returns to scale? Explain the laws of returns to scale by using isoquants. What are the factors behind the increasing and decreasing returns to scale?

15. Distinguish between economies and diseconomies of scale. Why do firms enjoy economies of scale by increasing scale production in the first stage and diseconomies of scale in the later stages of production?

16. Supposing price of capital, Pk = `2 and price of labour, $Pl$ = `5 and $Q = 20$. Find

(*a*) slope of the isocost; and (*b*) equation for the isocost.

[**Ans.** (*a*) If labour is plotted along horizontal axis and capital along vertical axis, $\Delta Pl/\Delta Pk = 2.5$, (*b*) $20 = 2K + 5L$]

17. Suppose a short-run production function is given as follows:

$Q = 2\,L\,2 + 0.2\,L\,3$

where Q = output and L = variable input.

Find the following:

(*a*) marginal product function,

(*b*) average product function, and

(*c*) value of L that maximizes Q.

18. Determine whether the following production functions show constant, increasing or decreasing

returns to scale:

(*a*) $Q = L^{0.60} K^{0.40}$ (*b*) $Q = 5 K^{0.5} L^{0.3}$ (*c*) $Q = 4L + 2K$

19. Suppose a production function is given as

$Q = -L^3 + 5L^2 + 10L$

(*i*) Which law of production is revealed by this production function?

(*ii*) At what level of labour employment does the total production begin to decline?

[**Ans.** (*i*) The law of diminishing returns, (*ii*) Five workers]

20. (*a*) What is the marginal rate of technical substitution?

(*b*) Illustrate graphically the substitution effect of a change in relative prices of inputs.

21. Suppose a short-run production function is given as

$Q = 10L + 15L^2 - L^3$

where Q is output and L is labour employed per unit of time.

(*i*) Derive *MPL* and *APL* schedules;

(*ii*) Derive *MPL* functions;

(*iii*) Find the output at which *APL* = *MPL*; and (*iv*) Find L for producing 600 units of output.

(Compare your answers with figures in Table 10.1).

22. Suppose a Cobb-Douglas production function is given as

$Q = L^{0.5} K^{0.5}$

(*a*) Find the degree of production 'function',

(*b*) Find the law of production it reveals, and

(*c*) Find the output for 10 units of labour and 5 units of capital.

23. Show the effects of change in input prices on the isocost line. How is the optimum combination of inputs affected if (*a*) price of only one input decreases, and (*b*) prices of both the inputs decrease proportionately?

24. Which of the following gives the condition for the least-cost combination of inputs?

(*a*) *PL*/ *PK* = *MRTS*

(*b*) *PL*/ *PK* = *MPL*/ *MPK*

(*c*) *MPL*/ *PL* = *MPK*/ *PK*

(*d*) *MPK*/ *MPL* = $\Delta L / \Delta K$

(*e*) All of the above

25. Suppose production department of a company estimates its production function as given below.

$Q = K^{0.6} L^{0.4}$

(*a*) What kind of returns to scale does this production reveal?

(*b*) Suppose the power of L is raised to 0.6. What returns to scale does it give?

FURTHER READING

Baumol, William, J., *Economic Theory and Operations Analysis,* (Prentice-Hall India, New Delhi, 1985), 4th Edn., Chs. 9 and 10.

Bilas, richard A., *Microeconomic Thoery*, (Richard D. Irwin, Homewood, IV, 1972), Ch. 7.

Brigham, EugEnE F. and JamEs L. PaPPas, *Managerial Economics,* 2nd Edn., (Dryden Press, Hinsdale IV, 1972), Ch. 6, 8.

clowEr, roBErt W., and John F. duE, *Microeconomics,* (Dryden Press, Homewood IV, 1972), Ch. 8.

douglas, Evan J., *Managerial Economics: Theory, Practice and Problems,* (Prentice-Hall Inc., N.J., 1979), Ch. 6.

FErguson, C. E., *Microeconomic Theory*, 3rd Edn., (Richard D. Irwin, Homewood, III., 1972), Ch. 7.

hirschEy, mark, *Economics for Managers* (Thomas South Western, Indian Edition, 2007), Chs. 8 and 9.

kEating, Barry and wilson J. holton, *Managerial Economics: An Economic Foundation for Business Decisions* (Atamic Dog Publishing, USA, Indian Imprint by Biztantra, 2nd Edn., New Delhi, 2003), Ch. 7.

mansFiEld, Edwin, allEn, dohErty, w. BrucE, wEigElt , nEil a., *Managerial Economics: Theory and Application, and Cases*, (W.W. Norton, New York, 1970), Ch. 7, 5th Edn.

PEtErson, h. craig and cris w. lEwis, *Managerial Economics*, (Pearson Education, Delhi), 4th Edn., Ch. 6.

salvatorE, dominick, *Managerial Economics*, 4th Edn., (Oxford University Press, New Delhi, 7th Ed, 2012), Ch. 6.

sPEncEr, milton H., *Managerial Economics,* (Richard D. Irwin, Homewood IV, 1968).

stiglEr, gEorgE J., *The Theory of Price,* 3rd Edn., (Macmillan, New York, 1966), Chs. 6, 7.

watson, donald S., *Price Theory and Its Uses,* 3rd Edn., (Houghton Mifflin, Boston, 1968).

wEBB, sammuEl, C., *Managerial Economics, (Houghton Mifflin, Boston, 1976), Chs. 12, 13 and 15.*

CHAPTER

11 Theory of Cost and

Break-Even Analysis

CHAPTER OBJECTIVES

The objective of this chapter is to discuss elaborately the cost-output relationship, i.e., how cost of production changes with change in production, and the meaning and purpose of break-even analysis.

The content of this chapter includes:

• Accounting and analytical cost concepts

• Short-run cost functions, cost curves and cost-output relationship

• Long-run cost functions, cost curves and cost-output relationship

• Economies and diseconomies of scale and cost of production

• Learning curve – a modern approach to cost-output relations

• Break-even analysis and production decision

11.1 INTRODUCTION

In the previous chapter, we have discussed the short-run and long-run theories of production, i.e., the input-output relationship, in terms of quantities of input and output. Production involves use of inputs. Inputs (labour and capital) have a cost. So production involves cost of production. Given the input prices, theory of production provides clue to the behaviour of cost of production. In this chapter, we discuss the theory of cost, i.e., cost-output relationship. The theory of cost provides conditions for minimizing the cost of production. In addition, analysis of cost of production is very important in almost all kinds of business decisions, especially those related to the weak points of production management; determining the output level for cost minimization; determining the price of the product and dealers' margin; and estimating and projecting the cost of business operation.

This chapter is divided into three sections. Section 11.2 explains various cost concepts used in business decisions and section 11.3 analyzes cost-output relations. Break-even analysis is discussed in section 11.4.

11.2 COST CONCEPTS

The cost concepts that are relevant to business operations and decisions can be grouped on the basis of their nature and purpose under two overlapping categories: (*i*) cost concepts used for

accounting purposes, and (*ii*) analytical cost concepts used in economic analysis of business activities. We will discuss some important concepts of the two categories. It is important to note here that this classification of cost concepts is only a matter of analytical convenience.

11.2.1 Accounting Cost Concepts

1. *Opportunity Cost and Actual Cost* *Actual cost* is all paid out costs of the business firms to take the advantage of the best opportunity available to them. The opportunity cost is the opportunity lost for lack of resources. An opportunity to make income is lost because of scarcity of resources like land, labour, capital, etc. We know that resources available to any person, firm or society are scarce but have alternative uses with different returns. Income maximizing resource owners put their scarce resources to their most productive use and thus, they forego the income expected from the second best use of the resources. Thus, the *opportunity cost* may be defined as the returns expected from the second best use of the resources foregone due to the scarcity of resources. The opportunity cost is also called *alternative cost.* Had the resource available to a person, a firm or a society been unlimited, there would be no opportunity cost.

To explain and illustrate the concept of opportunity cost, suppose a firm has a sum of `100,000 for which it has only two alternative uses. The firm can buy either a printing machine or a photo copier, both having a productive life of 10 years. The firm expects an annual income of `20,000 from the printing machine, and `15,000 from the photo copier.

A profit maximizing firm would invest its money in the printing machine and forego the expected income from the photo copier. The opportunity cost of the income from printing machine is the foregone income expected from the photo copier, i.e., `15,000. In assessing the alternative cost, 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 of expected earnings firm printing machine, the *economic rent* of the printing machine is the excess of its earning over the income expected from the photo copier. That is, *economic profit* or *economic rent* of the printing machine equals `20,000 – `15,000 = `5,000. The implication of this concept for a businessman is that investing in the 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, it will be helpful in choosing the best investment avenue.

2. *Business Costs and Full Costs* *Business costs* include all the expenses that are incurred to carry out a business. The concept of business costs is similar to the actual or real costs. Business costs "include all the payments and contractual obligations made by the firm together with the book cost of depreciation on plant and equipment."[1] Business costs are used for calculating business profits and losses and for filing returns for income-tax and also for other legal purposes.

The concept of ***full cost***, includes business costs, opportunity cost and normal profit. The opportunity cost includes the foregone earning expected from the second best use of the resources, or the market rate of interest on the internal money capital and also the value of an entrepreneur's own services that are not charged for in the current business. Normal profit is a necessary minimum earning in addition to the opportunity cost, which a firm must receive to remain in its present occupation.

1. Watson, Donald S., *Price Theory and Its Uses*, (Houghton Mifflin Company, Boston, 1963), p. 126.

3. *Actual or Explicit Costs and Implicit or Imputed Costs* The ***Actual*** or ***Explicit costs*** are those which are actually incurred by the firm in payment for labour, material, plant, building, machinery, equipment, travelling and transport, advertisement, etc. The total money expenses, recorded in the books of accounts are, for all practical purposes, the *actual costs*. Actual cost comes under the accounting cost concept.

In contrast to explicit costs, there are certain other costs that 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***. Opportunity cost is an important example of implicit cost. 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 sets up his own business, he foregoes his salary as manager. This loss of salary is the opportunity cost of income from his own business. This is an implicit cost of his own business. Thus, implicit wages, rent, and implicit interest are the wages, rent and interest that an owner's labour, building and capital respectively, can earn from their second best use.

Implicit costs are not taken into account while calculating the loss or gains of the business, but they form an important consideration in deciding whether or not to retain a factor in its present use. The explicit and implicit costs together make the ***economic cost.***

4. *Out-of-Pocket and Book Costs* The items of expenditure that involve cash payments or cash transfers, both recurring and non-recurring, are known as ***out-of-pocket costs***. All the explicit costs (e.g., wages, rent, interest, cost of materials and maintenance, transport expenditure, electricity and telephone expenses, etc.) fall in this category. On the contrary, there are certain actual business costs that do not involve cash payments, but a provision is therefore made in the books of account and they are taken into account while finalizing the profit and loss accounts. Such expenses are known as ***book costs***. In a way, these are payments made by a firm to itself. Depreciation allowances and unpaid interest on the owner's own funds are the example of *book costs.*

11.2.2 Analytical Cost Concepts

The analytical cost concepts refers to the different cost concepts that are used in analysing the cost-output relationship with increase in inputs and output and also the cost concepts that figure in analysing the effect of expansion of production on the society as a whole.

1. *Fixed and Variable Costs. Fixed costs* are those that remain fixed in amount for a certain quantity of output. Fixed cost does not vary with variation in the output between zero and a certain level of output. In other words, costs that do not vary or remain constant for a certain level of output are treated as *fixed costs*. The fixed costs include (*i*) depreciation of machinery, building and other fixed assets, (*ii*) costs of managerial and administrative staff, (*iii*) maintenance of land, etc. The concept of fixed cost is associated with the short-run.

Variable costs are those which vary with the variation in the total output. Variable costs include cost of raw material, running cost of fixed capital, such as fuel, repairs, routine maintenance expenditure, direct labour charges associated with the level of output, and the costs of all other inputs that vary with output.

2. *Total, Average and Marginal Costs. Total cost* (*TC*) refers to the total outlays of money expenditure, both explicit and implicit, on the resources used to produce a given level of output.

It includes both fixed and variable costs. The total cost for a given output is measured as

TC = Total fixed cost + Total variable cost

***Average cost* (*AC*)** is of statistical nature—it is not actual cost. It is obtained by dividing the total cost (*TC*) by the total output (*Q*), i.e.,

$$AC = \frac{TC}{Q}$$

***Marginal cost* (*MC*)** is defined as the addition to the total cost on account of producing one additional unit of the product. Or, marginal cost is the cost of the marginal unit produced. Marginal cost is calculated as $TC_n - TC_{n-1}$ where n is the number of units produced. Using cost function, *MC* is obtained as the first derivative of the cot function.

$$MC = \frac{\partial TC}{\partial Q}$$

Total, average and marginal cost concepts are used in the economic analysis of firm's production and in pricing decisions. These cost concepts are discussed in further detail in the following section.

3. *Short-Run and Long-Run Costs.* Short-run and long-run cost concepts are related to variable and fixed costs, respectively, and often figure in economic analysis cost-output relationship.

Short-run refers to the time period during which scale of production remains unchanged. The costs incurred in the short-run are called short-run costs. It includes both the variable and the fixed costs. From analytical point of view, short-run costs are those that vary with the variation in output in short-run, the size of the firm remaining the same. Therefore, *short-run* costs are treated as *variable costs*.

Long-run costs, on the other hand, are those that are incurred to increase the scale of production in the long-run. The costs that are incurred on the fixed factors like plant, building, machinery, etc., are known as long-run costs. It is important to note that the running cost and depreciation of the capital assets are included in the short-run or variable costs. Furthermore, ***long-run*** costs are by implication the costs that are incurred in the long-run. 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 variables in the utilization of fixed plant or other facilities whereas long-run costs are associated with the changes in the size and kind of plant.'[2]

4. *Incremental Costs and Sunk Costs.* Conceptually, ***incremental costs*** are closely related to the concept of marginal cost but with a relatively wider connotation. While marginal cost refers to the cost of the marginal unit (generally one unit) of output, incremental cost refers to the total additional cost associated with the decisions to expand the output or to add a new variety of product, etc. The concept of incremental cost is based on the fact that in the real world, it is not practicable (for lack of perfect divisibility of inputs) to employ factors for each unit of output separately. Besides, in the long run, when firms expand their production, they hire more of men, materials, machinery and equipments. The expenditures of this nature are incremental costs — not the marginal cost (as defined earlier). Incremental costs arise also owing to the change in product lines, addition or introduction of a new product, replacement of worn out plant and machinery, replacement of old technique of production, etc.

2. Dean, Joel, *Managerial Economics, op. cit.* , p. 262.

The ***sunk costs*** are those which are made once and for all and cannot be altered, increased or decreased, by varying the rate of output, nor can they be recovered. For example, once it is decided to make incremental investment expenditure and the funds are allocated and spent, all the preceding costs are considered to be the sunk costs. The reason is, such costs are based on the prior commitment and cannot be revised or reversed or recovered when there is a change in market conditions or change in business decisions.

5. *Historical and Replacement Costs.* *Historical cost* refers to the cost incurred in past on the acquisition of productive assets, e.g. land, building, machinery, etc., whereas *replacement cost* refers to the expenditure made for replacing an old asset. These concepts owe their significance to the unstable nature of input prices. Stable prices over time, other things given, keep historical and replacement costs on par with each other. Instability in asset prices makes the two costs differ from each other.

As regards their application, *historical cost* of assets is used for *accounting purposes*, in the assessment of the net worth of the firm whereas replacement cost figures in business decisions regarding the renovation of the plant.

6. *Private and Social Costs.* We have so far discussed the cost concepts that are related to the working of the firm and that are used in the cost-benefit analysis of business decisions. In simple words, all costs incurred by the business firms to run the business with the objective of making profile. All such costs fall in the category of *private costs*. There are, however, certain other costs that arise due to the functioning of the firm but do not figure normally in the business decisions nor are such costs explicitly borne by the firms. The costs of this category are borne by the society. Thus, the total cost generated by a firm's working may be divided into two categories: (*i*) costs paid out or provided for by the firms, and (*ii*) costs paid or borne by the society including the use of resources freely available plus the disutility created in the process of production. The costs of the category (*i*) are known as *private costs* and of category (*ii*) are known as *external* or *social costs*.

To mention a few examples of social cost, Mathura Oil Refinery discharging its wastage in the Yamuna river causes water pollution. Mills and factories located in a city cause air pollution, environment pollution and so on. Such costs are termed as *external costs* from the firm's point of view and *social costs* from the society's point of view.

The relevance of the social costs lies in the social cost-benefit analysis of the overall impact of a firm's operation on the society as a whole and in working out the social cost of private gains. A further distinction between private cost and social cost is, therefore, in order.

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 costs that are incorporated in the firm's total cost of production.

Social costs on the other hand, refer to the total cost borne by the society due to production of a commodity. Social costs include both private cost and the external cost. Social cost includes (*a*) the cost of resources for which the firm is not required to pay a price, i.e., atmosphere, rivers, lakes, etc., and also for the use of public utility services3 like roadways, drainage system, etc., and (*b*) the cost in the form of 'disutility' created through air, water, noise and environment pollution, etc. The costs of category (*b*) are generally assumed to equal the total private and public expenditure incurred to safeguard the individual and public

3. There may be some nominal payment for the use of public utilities in the form of tax, which may not cover the full cost thereof.

interest against the various kinds of health hazards and social tension created by the production system. The private and public expenditure, however, serve only as an indicator of 'public disutility'—they do not give the exact measure of the public disutility or the social costs.

11.3 THEORY OF SHORT-RUN COST: SHORT-RUN COST-OUTPUT RELATIONS

The theory of cost deals with how cost of production changes with change in output. In other words, the cost theory deals with cost-output relations. The basic principle of the cost-output relationship 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 important from a theoretical and managerial decision point of view is the rate of increase in total cost with increase in output and 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.

A cost function is a symbolic statement of the technological relationship between the cost and output. The general form of the cost function is written as

$$TC = f(Q), \text{ with } \Delta TC/\Delta Q > 0 \qquad \ldots(11.1)$$

The actual form of 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.

11.3.1 Short-Run Cost-Output Relations

In this section, we will analyze the cost-output relations in the short-run. The long-run cost output relations discussed in the following section. Before we discuss the short-run cost-output relations, let us first look at the cost concepts and the components used to analyze 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 are 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,

$TC = TFC + TVC$

...(11.2)

As mentioned earlier, *TFC* (i.e., the cost of plant, building, etc.) remains fixed in the short run, whereas *TVC* (the labour cost) varies with the variation in the output.

For a given quantity of output (*Q*), the average cost (*AC*), average fixed cost (*AFC*) and average variable cot (*VC*) can be defined as follws.

$T\ C\ T\ FC + TV\ C\ TFC\ TVC$

$AC =$

=

=

+

Q

Q

Q

$Q = AFC + AVC$

Thus,

$AFC = T\ FC$ and $AVC = TV\ C$

Q

Q

and

$AC = AFC + AVC$

...(11.3)

Marginal cost (*MC*) can be defined as the change in the total cost due to change in the total output by one unit, i.e.,

$MC = \Delta\ TC$

$TC - TC$

n

n 1

−

=

...(11.4)

$\Delta\ Q$

$Q - Q =$

n

$n-$

1

1

In case *TC* is expressed in functional form, *MC* is measured as the first derivative of cost function, i.e., $\partial\ TC$.

$\partial\ Q$

It may be added here 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 various cost curves.

11.3.2 Short-Run Cost Functions and Cost Curves

The cost-output relations are revealed by the cost function and are exhibited through cost curves4. 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. The nature of
estimated cost function depends on the
cost trend revealed by cost data. Given
the cost data, cost functions may take a
variety of forms, e.g., linear, quadratic
or cubic, yielding different kinds of
cost curves. The cost curves produced
by *linear, quadratic* and *cubic cost*
functions are illustrated below.

1. *Linear Cost Function.* When total
cost increases at a constant rate with
increase in production, it produces
a linear cost function. A linear cost
function takes the following form.

$TC = a + bQ$

...(11.5)

Fig. 11.1 *Linear Cost Functions*

where TC = total cost, Q = quantity
produced, $a = TFC$, and b = Change in TVC due to change in Q.

Given the cost function [Eq. (11.5)] AC and MC can be obtaied as follows.

$AC = TC\ a + bQ\ a$

=

=

$+\ b$

Q

Q

Q

$\partial\ TC$

and

$MC =$

$= b$

$\partial\ Q$

Note that since ' b' is a constant coefficient, MC remains constant, throughout in case
of a linear cost function.

4. Some examples of the trend in cost curves produced by estimated cost functions are given in Appendix of this
chapter.

To illustrate a linear cost function, let us suppose that an actual cost function is given as

$TC = 60 + 10\ Q$

...(11.6)

Given the cost function (11.6), one can easily work out TC, TFC, TVC, MC and AC for
different levels of output (Q) and can present them in the form of a table as shown in Table 11.1.

Table 11.1 *Tabular Cost Function*

Output Q

TFC = 60 TVC = 10Q TC = 60 + 10Q

MC = b = 10

AC = 60/Q + 10

1

60

10

70

–

70.0

2

60

20

80

10

40.0

3

60

30

90

10

30.0

4

60

40

100

10

25.0

5

60

50

110

10

22.0

6

60

60

120

10

20.0

7

60

70

130

10

18.6

8

60

80

140

10

17.5

9

60

90

150

10

16.6

10

60

100

160

10

16.0

Table 11.1 presents a series of *Q* and corresponding *TFC*, *TVC*, *TC*, *MC* and *AC* for output *Q* from 1 to 10. The figures in Table 11.1, graphed in Fig. 11.1, shows the relationship between total costs (*TC*, *TEC*, and *TVC*) and output.

Figure 11.1 shows the

behaviour of *TC*, *TVC* and
TFC. The horizontal line
shows *TFC* and the line *TVC*
= 10 *Q* shows the movement
in *TVC* with change in *Q*. The
total cost function is shown by

Fig. 11.2 *AC and MC Curves Derived from Linear Cost Function*

$TC = 60 + 10\,Q$.

More important is the
behaviour of *AC* and *MC* curves in Fig. 11.2. Given the cost function (11.6), *AC* and *MC* can be worked out as follows.

60

$AC =$

+10 and

$MC = 10$

Q

Note that, in case of a linear cost function, while *MC* remains constant, *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.*** When *TC* increases at increasing rate with constant increase in output (*Q*), the *TC* data produces a *quadratic cost function* expressed as

$TC = a + bQ + Q$ 2

....(11.7)

where *a* and *b* are constants and *TC* and *Q* are total cost and total output, respectively.

Given the cost function (11.7), *AC* an *MC* can be obtained as folows.

2

$AC = TC\ a + bQ + Q$

a

=

=

$+ b + Q$

...(11.8)

Q

Q

Q

$\partial\, TC$

$MC =$

$= b + 2\,Q$

...(11.9)

$\partial\, Q$

Let the actual (or estimated) cost function be given as

$TC = 50 + 5\,Q + Q$ 2

...(11.10)

Given the cot function (11.10),

$\partial\, C$

$AC = 50 + Q + 5$ and $MC =$

$= 5 + 2\,Q$

Q

$\partial\, Q$

The cost curves that emerge from the cost function (11.10) are graphed in Fig. 11.3 (*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* declines till output 8
and then begins to increase.

3. *Cubic Cost Function* When *TC* increases first at decreasing rate and then of increasing rate with increase in production, the *TC* data produces a cubic cost function. A cubic cost function is of the form

$TC = a + bQ - cQ^2 + Q^3$ ...(11.11)

where *a*, *b* and *c* are the parametric constants.

From the cost function (11.11), *AC* and *MC* can e derived as follows.

$$AC = \frac{TC}{Q} = \frac{a + bQ - cQ^2 + Q^3}{Q} = \frac{a}{Q} + b - cQ + Q^2$$

and $MC = \frac{\partial TC}{\partial Q} = b - 2cQ + 3Q^2$

Let us suppose that the cost function is empirically estimated as

Fig. 11.3 *Cost Curves Derived from a Quadratic Cost Function*

$TC = 10 + 6Q - 0.9Q^2 + 0.05Q^3$ ...(11.12)

Note that fixed cost equals 10. *TVC* can be obtained by subtracting 10—the fixed cost—from *TC*-function (11.12).

Thus,

$TVC = 6Q - 0.9Q^2 + 0.05Q^3$ ...(11.13)

The *TC* and *TVC*, based on Eqs. (11.12) and (11.13), respectively, have been calculated for *Q* = 1 to 16 and presented in Table 11.2. The *TFC*, *TVC* and *TC* have been graphically presented in Fig. 11.4. 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 two different trends with increase in output. 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 *TVC* curve. The two patterns of change in the *TVC* stems directly from the law of increasing and diminishing returns to the variable inputs. So long as the law of increasing returns is in operation, *TVC* increases at decreasing rate. And, when the law of diminishing returns comes into operation output increases at decreasing rate causing *TVC* to increase at increasing rate.

Table 11.2 *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

Given the equations (11.12) and (11.13), the behavioural equations for *AFC*, *AVC* and *AC* can be easily derived. Let us first derive the *AFC* curve.

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. (11.1) and *a* = 10 in Eq. (11.12). We know that

$AFC = TFC$

...(11.14)

Q

Fig. 11.4 *TC, TFC and TVC Curves*

Substituting 10 for *TFC* in Eq. 11.14, we get

10

$AFC = Q$

...(11.15)

Equation (11.15) expresses the behaviour of *AFC* in relation to change in *Q*. The behaviour of *AFC* with changing *Q* from 1 to 16 is given in Table 11.2 (Col. 5) and presented graphically by the *AFC* curve in Fig. 11.5. The *AFC* curve is a rectangular hyperbola.

Average Variable Cost (AVC). As defined above,

TVC

$AVC = Q$

Given the *TVC* function [Eq. (11.13], the *AVC* can be worked out as follows.

2

3

$6Q - 0.9Q + 0.05Q$

$AVC =$

Q

$= 6 - 0.9Q + 0.05Q$ 2

...(11.16)

Having derived the *AVC* function in Eq. (11.16), we may easily obtain the behaviour of *AVC* in response to change in *Q*. The behaviour of *AVC* for output from *Q* = 1 to 16 is given in Table 11.2 (Col. 6), and graphically presented in Fig. 11.5 by the *AVC* curve.

The Critical Value of AVC *The critical value of Q (in respect of AVC) is one that* *minimizes AVC*. From Eq. (11.16), we may compute the critical value of Q in respect of *AVC*. The *AVC* will be minimum when its (decreasing) rate of change equals zero. This can be accomplished by differentiating Eq. (11.16) and setting it equal to zero. hus, critical value of Q can be obtained as follows.

Fig. 11.5 *Short-run Cost Curves*

Critical value of

$Q = \partial AVC = -0.9 + 0.10 Q = 0$

∂Q

$0.10 Q = 0.9$

$Q = 9$

In our example, the critical value of $Q = 9$. This can be verified from Table 11.2.

The *AVC* is minimum (1.95) at output 9.

Aerage Cost (AC)

TC

The average cost (*AC*) is defined as *AC* =

.

Q

Substituting Eq (11.12 for *TC* in the above equation, we get

2

3

$10 + 6 Q - 0.9 Q + 0.05 Q$

AC =

Q

10

$= Q + 6 - 0.9 Q + 0.05 Q 2$

...(11.17)

The Eq. (11.17) gives the behaviour of *AC* in response to change in *Q*. The behaviour of *AC* for $Q = 1$ to 16 is given in Table 11.2 (Col. 7) and graphically presented in Fig. 11.5 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. The level of output that minimizes *AC* can be obtained by differentiating Eq. (11.17) and setting it equal to zero. Cost-minimizing *Q* can be obtained as follows.

∂AC 10

=

$-0.9 + 0.1 Q$

2

∂Q

Q

= 0

When we simplify this equation by multiplying it by Q 2, it takes the form of a quadratic equation as

$10 - 0.9 Q 2 + 0.1 Q 3 = 0$

When this equation is multiplied by 10, for further simplification, it takes the form,

$Q 3 - 9 Q 2 - 100 = 0$

...(11.18)

By solving5 equation (11.18), 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 11.2.

Marginal Cost (MC) The concept of marginal cost (*MC*) is particularly useful in cost analysis. *MC* is obtained technically by the first derivative of the *TC* function. Given the *TC* function in Eq. (11.12), the *MC* function can be obtained as

$MC = \partial TC = 6 - 1.8 Q + 0.15 Q 2$

...(11.19)

∂Q

Equation (11.19) represents the behaviour of *MC*. The behaviour of *MC* for $Q = 1$ to 16 computed as $MC = TCn - TCn - 1$ is given in Table 11.2 (Col. 8) and graphically presented by the *MC* curve in Fig. 11.5. The critical value of *Q* with respect to *MC* is 6 or 7. This can be seen from Table 11.2.

11.3.3 Law of Diminishing Returns and the Cost Curves

We now return to the law of diminishing returns and explain it through the cost curves. Figs. 11.4 and 11.5 present cost curves conforming to 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 the marginal returns decrease 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 11.5 presents the short-run laws of cost of production. As the figure shows, in the initial stage of production, both *AFC* and *AVC* are declining because of internal economies. Since *AC* = *AFC* + *AVC*, *AC* is also declining. This shows the operation of the law of increasing returns to the variable input in the initial stage of production. 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.

In Fig. 11.5, 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 economies resulting from increase in production.

5. One method of solving quadratic equation is to factorize it and find the solution. Thus,

$$Q^3 - 9Q^2 - 100 = 0$$

$$(Q - 10)(Q^2 + Q + 10) = 0$$

For this equation to hold, one of the terms must be equal to zero. Suppose

$$(Q^2 + Q + 10) = 0$$

Then,

$Q - 10 = 0$ and $Q = 10$.

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.

11.3.4 Some Important Relationships between Different Measures of Cost

Some important relationships between costs used in analyzing the short-run cost-behaviour may now be summed up as follows:

(*a*) Over the range of output *AFC* and *AVC* fall, *AC* also falls.

(*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* increases.

(*c*) *AC* and *MC* are related in following ways.

(*i*) When *MC* falls, *AC* follows, over a certain range of output. When *MC* is falling, the rate of fall in *MC* is greater than that of *AC*, because while *MC* is attributed to a single marginal unit, *AC* 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 units to 10 units (see Fig.11.5). Over this range of output, *MC* begins to increase while *AC* continues to decrease. The reason for this can be seen in Table 11.1: when *MC* starts increasing, it increases at a relatively lower rate that is sufficient only to reduce the rate of decrease in *AC*—not sufficient to push the *AC* up.

(*iii*) *MC curve intersects AC curve at its minimum*. The reason is, while *AC* continues to decrease, *MC* begins to rise at the same level of output. Therefore, they are bound to intersect. Also, 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.

11.3.5 Output Optimization in the Short-Run

The technique of output optimization has already been discussed in Chapter 4. The same optimization technique can be used to find the cost-minimising output. In this section, we show the application of the same technique to cost-minimising output.

Let us suppose that a short-run cost function is given as

$TC = 200 + 5Q + 2Q^2$

…(11.20)

As noted earlier, the level of output is optimized at the level of production at which $MC = AC$. In other words, at optimum level of output, $AC = MC$. Given the cost function in Eq. (11.20),

2

$200 + 5Q + 2Q$

$AC =$

Q

$= 200 + 5 + 2Q$

...(11.21)

Q

∂TC

and

$MC =$

$= 5 + 4Q$

...(11.22)

∂Q

By equating *AC* and *MC* equations, i.e., Eqs. (11.21) and (11.22), respectively, and solving them for *Q*, we get the optimum level of output. Since at equilibrium,

$AC = MC$

$200 + 5 + 2Q = 5 + 4Q = 2Q$

Q

$2Q^2 = 200$

$Q^2 = 100$

$Q = 10$

Thus, given the cost function (11.20), the optimum output is 10.

11.3.6 Derivation of Firm's Supply Curve

In the preceding section, we have discussed the conditions and method of output optimization by firm, given the cost conditions and the price of the product. The same technique can be used to derive the firm's supply cure. In this section, we explain how supply curve is derived.

Fig. 11.6 *Derivation of Firm's Supply Curve*

Recall the law of supply all other things remaining the same, when pure increases, supply increases. This law is illustrated through the supply curve. Derivation of firm's supply curve is illustrated in Fig. 11.6. As shown in the figure, given the cost curves, *AC* and *MC,* and price of its product at P_2, firm's $MC=MR_2$ at point E_2 – the point of equilibrium. Point E_2 determines the optimum output of the firm. It means that the firm optimizes its output at OQ_2. It implies that at price P_2, the firm will supply OQ_2 quantity of output. Given the firm's equilibrium, if price of the product decreases to P_1, firm's equilibrium point shifts to E_1 – the 'shut-down point'. At point E_1, the firm recovers only its variable

cost and makes loss to the extent of its fixed cost. If the firm is not able to sustain the loss, it will pull down the shutter. That is why the point at which $P = SMC$ is called 'shut-down point'. At shut-down point, firm's supply is limited to OQ_1. It implies that at price P_1, firm's supply is limited to OQ_1. That is, when price decreases, firm's supply decreases too. Now let the market conditions improve and price increases back to P_2. Not that with increase in price, the supply of product increases. Suppose price increases further to P_3. In that case equilibrium point – the output maximization point – shifts to point E_3 and firm's optimum output increases to OQ_3. Similarly, when price goes up to P_4, the optimum output – the firm's supply – increases to OQ_4. The relationship between price and supply is indicated by the emboldened part of the *MC* curve, given by the *SMC* curve. It may thus be concluded that as price of a product increases, supply of the product increases too. This conclusion gives the basis for the formation of the 'Law of Supply'.

11.4 THEORY OF LONG-RUN COST: LONG-RUN COST-OUTPUT RELATIONS

In the preceding section, we have discussed the *short-run theory of cost*. It tells us how cost changes when production is increased by increasing the variable input (labour), fixed input (capital) remaining constant. In this section, we discuss the ***long-run theory of cost.*** In the context of the production theory, *long run* refers to a period in which firms can use more of both the inputs – labour and capital – to increase their production. The long-run theory of cost deals with the long-run cost-output relationship. In other words, long-run theory of cost states the nature of relationship between output and cost with increase in scale of production.

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. With this perception of long-run-cost-out relationship, we may now show the derivation of 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. 11.7. Let us now suppose that the firm decides to add two more plants 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. 11.7 (*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*.

Long-run Average Cost Curve (LAC)

Like *LTC*, long-run average cost curve (*LAC*) is derived by combining the short-run average cost curves (*SACs*). Note that there is one *SAC* associated with each *STC*. The curve SAC_1 in panel (*b*) of Fig. 11.7 corresponds to STC_1 in panel (*a*). Similarly, SAC_2 and SAC_3 in panel (*b*) correspond to STC_2 and STC_3 in panel (*a*), respectively. Thus, given the STC_1, STC_2, and STC_3 curves in panel (*a*) of Fig. 11.6, there are three corresponding *SAC* curves as given by SAC_1, SAC_2, and SAC_3 curves in panel (*b*) of Fig. 11.7. 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 SAC_1, SAC_2 and SAC_3 as shown in Fig.11.7 (*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.

Fig. 11.7 *Long-run Total and Average Cost Curves*

The *SAC* curves can be derived from the data given in the *STC* schedule, from *STC* function or straightaway from the *LTC* curve.[6] Similarly, *LAC* curve can be derived from *LTC*-schedule, *LTC* function or from *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, *LAC*, i.e., unit cost of production, initially decreases, but ultimately increases as

6. The *SAC* curves can be obtained by measuring the slope of *STC* at different levels of output. For a simple exposition of the method, see Leftwich, R.H., *The Price System, and Resource Allocation* (The Dryden Press, Illinois), 4th edn., Appendix to Ch. 8.

shown in Fig. 11.7 (*b*). The decrease in unit cost is attributed to the internal and external economies of scale and the eventual increase in cost, to the internal and external diseconomies of scale. 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. 11.8 in which *SACs*, *SMCs* and *LAC* are the same as in Fig.11.7 (*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 minimum *LAC* at the different levels of production. Each of these outputs has an *SMC*. For example, if we draw a perpendicular from point *A*, it intersects *SMC* 1 at point *M* determining *SMC* at *MQ* 1 at output *Q* 1. The same process can be repeated for points *B* and *C* to find out *SMC* at outputs *Q* 2 and *Q* 3. Note that points *B* and *C* determine *SMC* at *BQ* 2 and *CQ* 3, respectively. 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. 11.7. 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 be noted is that *LMC* intersects *LAC* when the latter is at its minimum, i.e., point *B*. It indicates that there 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. 11.8 *Derivation of LMC*

Optimum Plant Size and Long-Run Cost Curves

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 that minimizes the *LAC*. Given the state of technology over time, theoretically there is technically a unique size of the firm and level of output associated with the least-cost concept. In Fig. 11.8, the

optimum size of the firm consists of two plants represented by *SAC* 1 and *SAC* 2. The two plants together 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 that gives minimum *LAC* where *SAC* = *SMC* = *LAC* = *LMC*. This size of plant assures the most efficient utilization of the resources. Any change in output level—increase or decrease—will make the firm enter the area of inoptimality.

11.5 ECONOMIES AND DISECONOMIES OF SCALE AND COST OF PRODUCTION

The economies and diseconomies of scale have already been discussed briefly in Chapter 10 in the context of laws returns to scale. The discussion there is confined to *internal economies* and *diseconomies* of scale of production. It is internal in the sense that production is an internal activity of the firm. While optimization of output in the long run is an important concern of business firms, *cost minimization* is an equally import decision area. Cost of production depends not only on *internal factor* – the productivity of inputs – but also on many *external factors* – the factors that arise out of the firm.

Since we are concerned in this chapter with the theory of cost, in this section we give a detailed analysis of internal and external economies and diseconomies of scale and how they determine the trend in cost of production. To begin with, let us have look at the trend of long-run average cost curve (*LAC)*.

As shown in Fig. 11.7(*b*), *LAC* decreases with the expansion of production scale up to *OQ* 2 and then it begins to rise. Decrease in *LAC* is caused by the economies of scale and increase in *LAC* is caused by diseconomies of scale. Economies of scale result in cost saving and diseconomies lead to rise in cost. *Economies and diseconomies of scale* determine also the returns to scale. Increasing returns to scale operate till economies of scale are greater than the diseconomies of scale, and returns to scale decrease when diseconomies are greater than the economies of scale. When economies and diseconomies are in balance, returns to scale are constant. In this section, we briefly discuss the various kinds of economies and diseconomies of scale and their effect on cost of production.

11.5.1 Economies of Scale

The economies of scale are classified as

(*a*) Internal or Real Economies, and

(*b*) External or Pecuniary Economies.

A. *Internal Economies*

Internal economies, also called 'real economies', are those that arise within the firm with addition of new production plants. This means that internal economies are available exclusively to the expanding firm. Internal economies may be classified under the following categories.7

7. For a more detailed discussion, see A. Koutsoyiannis, *Modern Microeconomics* (Macmillan, 1979), pp. 128–36.

(*i*) Economies in production;

(*ii*) Economies in marketing;

(*iii*) Managerial economies, and

(*iv*) Economies in transport and storage.

(*i*) Economies in Production Economies in production arise from two sources: (*a*) technological advantages, and (*b*) advantages of division of labour based on specialization and skill of labour.

Technological advantages. Large-scale production provides an opportunity to the expanding firms to avail the advantages of technological advances. Modern technology is highly specialized. The advanced technology makes it possible to conceive the whole process of production of a commodity in one composite unit of production. For example, production of cloth in a textile mill may comprise such plants as (*i*) spinning; (*ii*) weaving; (*iii*) printing and pressing; and (*iv*) packing, etc. Likewise, a composite dairy scheme may consist of plants like (*i*) chilling; (*ii*) milk processing; and (*iii*) bottling. Under small-scale production, the firm may not find it economical to have all the plants under one roof. It would, therefore, not be in a position to take the full advantage of a composite technology. But, when scale of production expands and firms hire more capital and labour, their total output increases more than proportionately till the optimum size of the firm is reached. It results in lower cost of production.

Advantages of division of labour and specialization. When a firm's scale of production expands, more and more workers of varying skills and qualifications are employed. With the employment of larger number of workers, it becomes increasingly possible to divide the labour according to their qualifications, knowledge, experience, expertise and skills and to assign them the function to which they are best suited. This is known as division of labour. *Division of labour* leads to a greater specialization of manpower. It increases productivity of labour and, thereby, reduces cost of production. Besides, specialized workers develop more efficient tools and techniques and gain speed of work. These advantages of division of labour improve productivity of labour per unit of labour cost and time. Increase in labour productivity decreases to per unit cost of production.

(*ii*) Economies in Purchase of Inputs Economies in input purchases arise from the large-scale purchase of raw materials and other material inputs and large-scale selling of the firm's own products. As to economies in the purchase of inputs, the large-size firms normally make bulk purchases of their inputs. The large scale purchase entitles the firm for certain discounts in input prices and other concessions that are not available on small purchases. As such, the growing firms gain economies on the cost of their material inputs. The internal economies arise also in marketing the firm's own product as (*a*) economies in advertisement cost; (*b*) economies in large-scale distribution through wholesalers, etc.; and (*c*) other large-sale economies. With the expansion of the firm, the total production increases. But the expenditure on advertising the product does not increase proportionately. Similarly, selling through the wholesale dealers reduces the cost of distribution of the firm's production. The firm also gains on large scale distribution through better utilization of 'sales force, distribution of sample, etc.'

(*iii*) Managerial Economies Managerial economies arise from (*a*) specialization in managerial activities, i.e., the use of specialized managerial personnel, and (*b*) systemization of managerial functions. For a large-size firm, it becomes possible to divide its management

into specialized departments under specialized personnel, such as production manager, sales manager, HR manager, financial manager, etc. The management of different departments by speciallized managers increases the efficiency of management at all the levels of management because of the decentralization of decision-making. It increases production, given the cost. Large-scale firms have the opportunity to use advanced techniques of communication, telephones and telex machines, computers, and their own means of transport. All these lead to quick decision-making, help in saving valuable time of the management and, thereby, improve the managerial efficiency. For these reasons, managerial cost increases less than proportionately with the increase in production scale upto a certain level, of course.

(*iv*) Economies in Transport and Storage Economies in transportation and storage costs arise from fuller utilization of transport and storage facilities. Transportation costs are incurred both on production and sales sides. Similarly, storage costs are incurred on both raw materials and finished products. The large-size firms may acquire their own means of transport and they can, thereby, reduce the unit cost of transportation, at least to the extent of profit margin of the transport companies. Besides, own transport facility prevents delays in transporting goods. Some large-scale firms have their own railway tracks from the nearest railway point to the factory, and thus they reduce the cost of transporting goods in and out. For example, Bombay Port Trust has its own railway tracks, oil companies have their own fleet of tankers. Similarly, large-scale firms can create their own godowns in various centres of product distribution and can save on cost of storage.

B. *External or Pecuniary Economies of Scale*

External economies are those that arise outside the firm and accrue to the expanding firms. External economies appear in the form of money saving on inputs, called *pecuniary economies*. Pecuniary economies accrue to the large-size firms in the form of discounts and concessions on (*i*) large scale purchase of raw material, (*ii*) large scale acquisition of external finance, particularly from the commercial banks; (*iii*) massive advertisement campaigns; (*iv*) large scale hiring of means of transport and warehouses, etc. These benefits are available to all the firms of an industry but large scale firms benefit more than small firms.

Besides, expansion of an industry encourages the growth of ancillary industries that supply inputs. In the initial stages, such industries also enjoy the increasing returns to scale. In a competitive market, therefore, input prices go down. The benefit of decreasing input prices accrues to the expanding firms in addition to discounts and concessions. For example, growth of the automobile industry helps the development of tyre industry and other motor parts manufacturing units. The economies of scale reaped by such industries flow also to automobile industry. If Maruti Udyog Limited starts producing tyres for its own cars and ancillaries, cost of Maruti cars may go up. Consider another example: growth of computer industry encourages growth of firms that manufacture and supply computer chips and other software. Competition between such firms and law of increasing returns reduces the cost of inputs. Reduction in input costs is an important aspect of external economies.

11.5.2 Diseconomies of Scale

The economics of scale have their own limits, i.e., scale economies exist only up to a certain level of production scale. The expansion of scale of production beyond that limit creates condition for *diseconomies of scale*. Diseconomies of scale are disadvantages that

arise due to the expansion of production scale beyond its optimum level and lead to rise in the cost of production. Like economies, diseconomies may be internal and external. Let us describe the nature of **internal** and **external diseconomies** in some detail.

1. *Internal Diseconomies.* Internal diseconomies are those that are exclusive and internal to a firm as they arise within the firm. Like everything else, economies of scale have a limit too. This limit is reached when the advantages of division of labour and managerial staff have been fully exploited; excess capacity of plant, warehouses, transport and communication systems, etc., is fully used; and economy in advertisement cost tapers off. Although some economies may still exist, diseconomies begin to outweigh the economies and the costs begin to rise.

Managerial Inefficiency. Diseconomies begin to appear first at the management level. Managerial inefficiencies arise, among other things, from the expansion of scale itself. With fast expansion of the production scale, personal contacts and communications between (*i*) owners and managers, (*ii*) managers and labour, and (*iii*) between the managers of different departments or sections get rapidly reduced. The lack of fast or quick communication causes delays in decision-making affecting production adversely.

Secondly, close control and supervision is replaced by remote control management.

With the increase in managerial personnel, decision-making becomes complex and delays in decision-making become inevitable.

Thirdly, implementation of whatever decisions are taken is delayed due to coordination problem in large scale organisations.

Finally, with the expansion of the scale of production, management is professionalized beyond a point. As a result, the owner's objective function of profit maximization is gradually replaced by managers' utility function, like job security and high salary, standard or reasonable profit target, satisfying functions. All these lead to laxity in management and, hence to a rise in the cost of production.

Labour Inefficiency Increasing number of labour leads to a loss of control over labour management. This affects labour productivity adversely. Besides, increase in the number of workers encourages labour union activities that cause loss of output per unit of time and hence, rise in the cost of production.

2 . *External Diseconomies.* External diseconomies are the disadvantages that arise outside the firm, especially in the input markets, due to natural constraints, specially in agriculture and extractive industries. With the expansion of the firm, particularly when all the firms of the industry are expanding, the discounts and concessions that are available on bulk purchases of inputs and concessional finance come to an end. More than that, increasing demand for inputs puts pressure on the input markets and input prices begin to rise causing a rise in the cost of production. These are *pecuniary diseconomies*.

On the production side, the law of diminishing returns to scale come into force due to excessive use of fixed factors, more so in agriculture and extractive industries. For example, excessive use of cultivable land turns it into barren land; pumping out water on a large scale for irrigation causes the water table to go down resulting in rise in cost of irrigation; extraction of minerals on a large scale exhausts the mineral deposits on upper levels and mining further deep causes rise in cost of production; extensive fishing reduces the availability of fish and the catch, even when fishing boats and nets are increased. These kinds of diseconomies make the *LAC* move upward.

11.6 THE LEARNING CURVE

In Section 11.4, we have discussed the traditional theory of long-run cost-output relations and illustrated the derivation of *LAC* (see Fig. l1.7). As Fig.11.7 shows, *LAC* declines as the scale of production increases to a certain level and, beyond this level of production, *LAC* begins to rise. In the preceding section (Section 11.5), we have discussed the *economies and diseconomies of scale*. The economies of scale provide the reasoning for decrease in *LAC* with increasing scale of production and diseconomies of scale provide the reasoning why *LAC* begins to increase beyond a certain scale of production.

Learning by Doing and the Learning Curve. The economists[8] and business analysts have however discovered another factor that causes a continuous decrease in *average cost of production* over a large scale of production. The factor is called ***learning by doing*** or ***learning by experience***. Firms engaged in the production of a commodity or service over a long period of time gain *experience*. They learn by performing the same activity repeatedly. The learning pertains to technology of production, organizational behaviour and management style. The acquired knowledge and experience helps firms in getting a work done in a shorter period of time and at a *lower cost*. This helps them in devising ways and means of reducing cost of production or increasing factor productivity or in achieving both these ends simultaneously, with increasing scale of production. In the ultimate analysis, their *average cost of production* continues to fall with increase in production, though the rate of fall in average cost goes on declining. The curve that represents the declining trend in long-run average cost of production is called the **learning curve** as shown in Fig. 11.9. The learning curve is widely used by business managers, economists and engineers to foresee the possible trend inlong-run average cost of production and plan production accordingly.

Fig. 11.9 *The Learning Curve*

8. K.J. Arrow, a renowned economist, wrote a classic paper "The Economic Implications of Learning by Doing" in *Review of Economic Studies,* June 1962. See also Linda Argote and Dennis Epple, "Learning Curve in Manufacturing", *Economic Science,* February 23, 1990, and Herald Gruber, "The Learning Curve in the Production of Semiconductor Memory Chips", *Applied Economics,* August 1992.

It is ***important*** to note here that the learning curve is different from the conventional *LAC* curve. While *LAC* give the average cost of plant-wise production, learning curve gives the average cost of ***cumulative output***, i.e., the total output of a commodity produced over a period of time.

The **learning curve** as illustrated in Fig. 11.9 shows the decline in average cost of production with *cumulative* increase in production over time. For example, suppose Maruti Udyog Limited (MUL) produces 100 cars, at some point of time, at the total cost of `20 million, i.e., at an average cost of `200,000. When MUL doubles its car production to 200, its total cost increases tò35 million. Note that the unit cost of car production decreases tò35 mn ÷ 200 = `175,000. Similarly, when MUL increases its total production to 400 cars, its total cost rises tò64 million and its per unit cost decreases tò64 mn ÷ 400 = `160,000. This kind of decrease in average cost of production due to cumulative increase in output is the result of the long-run *experience* of the MUL. This kind of decrease in average cost presented graphically, as shown in Fig.11.9, produces the *learning curve*.

Estimating Unit Cost by Learning Curve

The learning curve is expressed algebraically by a power cost function as given below.

$C = aQb$

...(11.23)

where *C* is the average cost at output *Q; a* is the cost of the first unit of output; and exponent *b* the rate of decrease in the average cost with cumulative increase in output. *The value of b is negative* because of decrease in cost with cumulative increase in output: the greater the value of *b,* the faster the decrease in the average cost.

The learning-curve function given in Eq. (11.23) can be expressed in logarithmic form as follows.

$\log C = \log a + b \log Q$

In this logarithmic form, *b* gives the slope of the learning curve. The same equation is used to estimate the learning curve.

For estimating the learning curve, historical data on cumulative production and average cost are used. Regression technique is used to estimate the numerical value of *a* and *b*. Suppose, given the historical data on output and average cost, the learning curve is estimated as follows.

$\log C = 5 – 0.4 \log 100$

Since log of 100 is 2, by substitution, we get

$\log C = 5 – 0.4 (2) = 5 – 0.8 = 4.2$

Since anti-log of 4.2 = 15,849, the unit cost of 100th unit of output is taken to bè15,849. One can similarly work out the unit cost for any quantity of output.

The **learning curve** is widely applied in many manufacturing and service industries like manufacturing of airplanes, shipbuilding, refined petroleum products, power plants,9 etc. The basic purpose behind the use of learning curve is to forecast the unit cost with cumulative increase in output. Besides, learning curves are also used to forecast manpower, machinery and material needs of the company, to determine and to quote future competitive price of the product and for planning production.

9. Salvatore, D., *Managerial Economics,* 4th Edn., Thompson, 2001, p. 304.

11.7 BREAK-EVEN ANALYSIS: PROFIT CONTRIBUTION ANALYSIS

We have been concerned in this chapter so far with the theory of cost, i.e., the nature of relationship between cost and output. The theory of cost guides both the analysts and the firms to find the nature of change in cost of production with change in output. Also, it provides conditions for determining cost minimizing output. However, despite its high analytical value, cost theory does not provide answer to certain basic questions pertaining to business decisions of new firms. Some such questions are: (i) given the product price and cost, what is the minimum output to recover the total cost? (ii) what is the output at which total cost (*TC)* and total revenue (*TR*) break-even? and (iii) if anticipated *TR* and *TC* functions produce non-linear curves, what output would maximize the profit? The economists have developed a new technique, in addition to cost analysis, that can be used to find answer to these questions. The technique is known as ***Break-Even analysis*** known also as **Profit Contribution Analysis** and **Cost-Volume Profit Analysis** . In this section, we discuss the *break-even analysis* and also its different variants.

11.7.1 Meaning of Break-even Analysis

The break-even analysis is an important analytical technique used to study the relationship between the total costs, total revenue and total profit and loss over the whole range of stipulated output. The break-even analysis is a technique of having a preview of profit prospects and a tool of profit planning. It integrates the cost and revenue estimates to ascertain the profits and losses associated with different levels of output.

The relationship between cost and output and between price and output may be linear or non-linear in nature. We shall discuss the break-even analysis under both linear and non-linear revenue conditions.

11.7.2 Linear Cost and Revenue Functions

To illustrate the break-even analysis under linear cost and revenue conditions, let us assume linear cost and linear revenue functions are given as follows.

Cost function: $TC = 100 + 10\,Q$

...(11.24)

Revenue function: $TR = 15\,Q$

...(11.25)

The cost function given in Eq. (11.24) implies that the firm's total fixed cost is given at `100 and its variable cost varies at a constant rate of `10 per unit in response to increase in output. The revenue function given in Eq. (11.25) implies that the price for the firm's product is given in the market at `15 per unit of sale.

What the firm needs to carry out the break-even analysis of its business operations is to make a chart of its total fixed cost (*TFC*), total variable cost (*TVC*), total cost (*TC*) and the total revenue (*TR*), and graph them to find the break-even point. The process of break-even analysis is illustrated graphically in Fig. 11.10. The line *TFC* shows the total fixed cost at `100 for a certain level of output, and the line *TVC* shows the variable cost rising with a slope DT $VC/\Delta\,Q = 10/1 = 10$. The line *TC* has been obtained by plotting the *TC* function. It can also be obtained by a vertical summation of *TFC* and *TVC* at various levels of output. The line *TR* shows the total revenue (*TR*) obtained as $Q \times P$. The *TR* and *TC* lines intersect at point *B*, where output is equal to 20 units. The point *B* shows that at $Q = 20$, firm's total cost equals its total revenue. That is, at $Q = 20$, *TC* breaks-even with

TR. Point *B* is, therefore, the *break-even point* and $Q = 20$ is the break-even output. Below this level of output, *TC* exceeds *TR*. The vertical difference between *TC* and *TR*, (i.e., *TC*–*TR*) is known as operating loss. Beyond $Q = 20$, $TR > TC$, and *TR*–*TC* is known as operating profit. It may thus be inferred that a firm producing a commodity under cost and revenue conditions given in Eq. 11.23 and Eq. 11.24, respectively, must produce at least 20 units to make its total cost and total revenue break-even.

The break-even output can also

Fig. 11.10 *Break-even Analysis: Linear Functions*

be calculated algebraically. We know that at break-even point, $TR = TC$

That is, in terms of *TR* and *TC* functions,

$$15\,Q = 100 + 10\,Q$$

$$5\,Q = 100$$

$$Q = 20$$

Thus, 20 is the break-even output. Given the *TR* and *TC* functions, production beyond 20 units will yield increasing profits, at least in the short-run.

Algebra of Break-Even Analysis. The break-even analysis can also be presented algebraically. At break-even volume,

$$TR = TC$$

where $TR = (P \times Q)$ and $TC = TFC + TVC$.

In break-even analysis *TVC* is defined as $TVC = AVC \times Q$. Thus,

$$TC = TFC + AVC \times Q$$

Now, break-even quantity (QB) can be obtained as follows.

$$TR = TC$$

$$QB \times P = TFC + AVC \times QB \qquad \ldots(11.26)$$

where QB = break-even volume.

Rearranging Eq. (11.26), we get

$$QB \times P - AVC \times QB = TFC$$

$$QB(P - AVC) = TFC$$

$$QB = \frac{TFC}{P - AVC} \qquad \ldots(11.27)$$

If firm's *TFC*, *AVC* and *P* are known, *QB* can be obtained straightaway from Eq. (11.27). For example, given the cost and revenue function in Eqs. (11.23) ad (11.24), respectively, $TFC = 100$; $P = 15$, and $TVC = 10$. Thus,

$$QB = \frac{TFC}{P - AVC} = \frac{100}{15 - 10} = \frac{100}{5} = 20.$$

Limitations The theory of break-even analysis, as presented above, is applicable only if cost and revenue functions are linear. Under the condition of linear cost and revenue functions, *TC* and *TR* are straight lines and they intersect at only one point (as shown in Fig. 11.10.) dividing the whole range of output into two parts—profitable and non-profitable. It may give the impression that the whole output beyond the break-even level is profitable. In real life, however, market conditions keep changing due to changing price and cost levels. In reality, the cost and revenue functions may be non-linear. Non-linearity arises because *AVC* and price vary with variation in the output. As a result, the total cost (*TC*) may increase at increasing rates while the total revenue (*TR*) increases at decreasing rates. Therefore, at some stage of output, *TC* may exceed *TR*. Thus, there might be two break-even points (as shown in Fig. 11.11) instead of one. This limits the profitable range of output and determines the lower and upper limits of profitable output. The analyst should, therefore, pre-test and verify the validity of cost and revenue functions rather than assuming straightaway the linearity conditions.

11.7.3 Non-linear Cost and Revenue Functions

Let us now describe the break-even analysis under non-linear cost and revenue functions. The non-linear functions are presented in Fig. 11.11. *TFC* line shows the fixed cost at *OF* and the vertical distance between *TC* and *TFC* measures the total variable cost (*TVC*). The curve *TR* shows the total sale proceeds or the total revenue (*TR*) at different levels of output and price. The vertical distance between the *TR* and *TC* measures the profit or loss for various levels of output.

As shown in Fig. 11.11, *TR* and *TC* curves intersect at two points, B_1 and B_2, where $TR = TC$. These are the lower and upper break-even points. For the whole range of output between OQ_1 and OQ_2, total revenue (*TR*) is greater than totoal cost (*TC*). This is profit-making range of output. It implies that a firm producing an output more than OQ_1 and less than OQ_2 will make profits. In other words, the profitable range of output lies between OQ_1

and *OQ* 2 units of output. Producing less or more than these limits will result in losses.

Fig. 11.11 *Break-even Analysis*: *Non-linear Functions*

11.7.4 Contribution Analysis

As discussed in Chapter 3, contribution analysis is the analysis of incremental revenue and incremental cost of a business decision or business activity. Break-even charts can also be used for measuring the contribution made by the business activity towards covering the fixed costs. For this purpose, variable costs are plotted below the fixed costs as shown in Fig. 11.12. Fixed costs are a constant addition to the variable costs. In that case, the total cost line will run parallel to the variable cost line.

Fig. 11.12 *Contribution Analysis*

The 'contribution is the difference between total revenue and variable costs' arising out of a business decision. At the break-even level of output *OQ* in Fig. 11.12, contribution equals fixed costs. Below the output *OQ*, the total contribution is less than the fixed cost. This amounts to loss to the firm. Beyond output *OQ*, contribution exceeds fixed cost. The difference between *TR* and *VC* is the contribution towards profits resulting from a business decision.

Fig. 11.13 *Profit Contribution Analysis*

Sometimes, contribution over a time period is plotted in order to indicate the commitment that the management has made for fixed expenditure, and to find the level of output from which it will be recovered and profit will begin to emerge. This kind of contribution analysis is graphically presented in Fig. 11.13. At output *OQ*, contribution equals fixed cost. Beyond output *OQ*, contribution includes net profit.

11.7.5 Profit: Volume Ratio

The profit volume (*PV*) ratio is another handy tool used to find the *BEP* for sales, specially for the multi-product firms. The formula for *PV* ratio is given below.

$$PV \text{ Ratio} = \frac{S - V}{S} \times 100$$

where *S* = Selling price, and *V* = Variable costs (average).

For example, if selling price (*S*) = `5 and variable cost ( *V*) = `4 per unit, then,

$$PV \text{ Ratio} = \frac{5 - 4}{5} \times 100 = 20 \text{ per cent}$$

The break-even point (*BEP*) of sales value is calculate by dividing the fixed expenses by *PV* ratio as follows.

$$BEP \text{ (Sale value)} = \frac{\text{Fixed Expenses}}{PV \text{ Ratio}}$$

For example, given the selling price at `5 per unit, average variable expenses at `3 per unit and fixed expenses (*F*) of `4,00 pr mnth, *BEP* (sale value) is calculated as follows.

$$BEP \text{ (Sale Value)} = \frac{F}{PV \text{ Ratio}} \text{ or } \frac{\text{Fixed Expenses}}{\dfrac{(S - V)}{S}} = \frac{4000}{\dfrac{(5 - 3)}{5}} = \text{`}10{,}000$$

The break-even sale volume can also be calculated by using the ontributin per unit of sale by the following formula.

Fixed Expenses

BEP (Sale Value) = Contribution per unit

4000

4000

BEP =

=

= 2,000 units

(5 – 3)

2

The *PV* ratio is not only helpful in finding the break-even point but it can also be used for making a choice of the product.

If there is no time constraint, the choice should always be for a product that assures a higher *PV* ratio. Otherwise, *PV* ratio per time unit is taken as the basis of choice. For example, suppose two products *A* and *B* involve the following variable cost and selling price.

Products

A

B

Selling price per unit (`)

2

2.5

Variable cost pe unit (`)

1

1.5

Machine hour per unit

2

1.0

Sel ing Price – Variable cost

PV ratio for *A* =

× 100

Sel ing price

2 – 1

=

× 100 = 50 per cent

2

Therefre, for each machine hour, *PV* Ratio = 50/2 = 25 per cent

PV Ratio for *B* 2.5 – 1.5 ×100 = 100 = 40 per cent

2.5

Therefore, for each machine hour, *PV* Ratio = 40 per cent. In this case, product *B* is preferable to product *A*.

11.7.6 Margin of Safety

The margin of safety is given by the difference between the sales at break-even point and the total actual sales. Three measures of the margin of safety are given below.

(*i*) Margin of safety = Profit × Sales

PV ratio

Profit

(*ii*) Margin of safety = *PV* ratio

(*iii*) Margin of safety = *S* − *S*

a

b × 100

Sa

where *Sa* = actual sales and *Sb* = Sales at *BEP*.

The safety margin can be worked out by using formula (*iii*) as follows. Suppose *TR* and *TC* functions are given respectively as

TR = 10 *Q*

TC = 50 + 5 *Q*

and

$Sa = 20$

Given the *TR* and *TC* functions, sales at *BEP*, i.e., *Sb*, can be obtained as shown below. At break-even point, *TR* = *TC*.

By substituting *Sb* for *Q* in *TR* and *TC* functions, we get

$TR = 10\ Sb$

and

$TC = 50 + 5\ Sb$

Thus, at break-even point,

$10\ Sb = 50 + 5\ Sb$

$10\ Sb – 5\ Sb = 50$

$5\ Sb = 50$

$Sb = 10$

By substituting *Sa* and *Sb* in formula (*iii*), we get

20 − 10

Margin of safety =

× 100 = 50 per cent

20

Margin of safety can be increased by increasing the selling price, provided the sales are not seriously affected. This can happen only when demand for the product is inelastic. The margin of safety can also be increased by increasing production and sales up to the capacity of the plant, and if necessary, even by reducing selling price provided the demand is elastic. The other methods include reduction in fixed expenses, reduction in variable expenses or having a product mix with greater share of the one that assures greater contribution per unit or which has a higher *PV* ratio.

Profit-Volume Analysis Charts. The general break-even and contribution break-even charts have been discussed above in Figs. 11.10 through 11.13. There can be a number of such charts or graphs showing existing and proposed situations with variation in sales price, fixed and variable cost and, consequently, variable contribution to fixed costs, profits, etc. One of such charts is the cash break-even chart.

Fig. 11.14 *Cash Break-Even Analysis*

A cash break-even chart can be prepared by taking cash inflow from sales and cash outlay on fixed and variable costs. The distribution of the totalcontribution may also be shown from the angle of incidence as shown in Fig. 11.14.

Fig. 11.15 *Profit Volume Analysis*

Another variation of the break-even chart is called profit-volume analysis chart or graph. In this chart, the horizontal axis represents the sales volume and the vertical axis shows profit or loss. The profit line is graphed by computing the profit or loss consisting of the difference between sales revenue and the total cost at each volume. The point

where the profit line intersects the horizontal axis is the break-even point. This has been shown in Fig. 11.15.

It may be noticed that break-even charts are good for displaying information. The same information is available from simple calculations.

Use of Break-Even Analysis

(*i*) Sales volume can be determined to earn a given amount of return on capital.

(*ii*) Profit can be forecast if estimates of revenue and cost are available.

(*iii*) Effect of change in the volume of sales, sale price, cost of production, can be appraised.

(*iv*) Choice of products can be made from the available alternatives. Product-mix can also be determined.

(*v*) Impact of increase or decrease in fixed and variable costs can be highlighted.

(*vi*) Effect of high fixed costs and low variable costs to the total cost can be studied.

(*vii*) Valid inter-firm comparisons of profitability can be made.

(*viii*) Cash break-even chart helps proper planning of cash requirements.

(*ix*) Break-even analysis emphasizes the importance of capacity utilization for achieving economies.

(*x*) Further help is provided by margin of safety and angle of incidence.

Limitations of Break-even Analysis

We have discussed above that the break-even analysis is based on linear assumptions. The linearity assumption can be removed by pre-testing the cost and revenue functions and by using, if necessary, the non-linearity conditions. Nevertheless, the break-even analysis as such has certain other limitations.

First, the break-even analysis can be applied only to a single product system. Under the condition of multiple products and joint operations, the break-even analysis can be applied only if product-wise cost can be ascertained which is, of course, extremely difficult.

Second, break-even analysis cannot be applied usefully where cost and price data cannot be ascertained beforehand and where historical data are not relevant for estimating future costs and prices. Despite these limitations, the break-even analysis may serve a useful purpose in production planning if relevant data can be easily obtained.

SUMMARY

• Chapter 11 deals with cost concepts used in cost analysis, theory of cost, cost minimization technique, and break-even analysis used in production decision.

• Cost concepts are categorized under two categories: (*i*) accounting cost concepts, and (*ii*) analytical cost concepts.

• Accounting cost concepts include (*a*) opportunity cost and actual cost, (*b*) business cost and full cost, (*c*) explicit costs and implicit costs, and (*d*) out-of-pocket cost and book costs.

• Analytical cost concepts include (*a*) fixed cost and variable cost, (*b*) total, average and marginal cost, (c) short-run cost and long-run cost, (*d*) incremental cost and sunk cost, (*e*) historical cost and replacement cost, and (*f*) private cost and social cost.

• Theory of cost reveals the relationship between output and cost of production. The theory of cost analyzes the cost-output relationship in short-run perspective and long-run perspective. Short run refers to a period during which some cost remain fixed, called fixed cost and long run refers to a period during which all kinds of costs are variable. Accordingly, there is short-run theory of cost and long-run theory of cost.

• In short-run theory of cost assumes that some costs (cost of capital) remains fixed, called fixed cost (*FC*) and some costs change with change in output, called variable cost (*VC*). Thus, the total short-run cost (TC) = $TFC + TVC$. Given *TC, TFC* and *TVC* for a certain level of output,

Average cost

(AC) = $TC/Q = TFC/Q + TVC/Q = AFC + AVC$.

And, Marginal cost (MC) = $\Delta TC/\Delta Q$.

• According to the short-run theory of cost, the nature of cost-output relationship depends on the nature of product and cost function. The cost-output relationship is estimated in the form of cost function. Cost function may be in the form of (*i*) linear function, (*ii*) quadratic function, or (*iii*) cubic function, producing different kinds of cost curve.

• In general, however, when output (*Q*) increases, *TC* increases initially at decreasing rate and ultimately at increasing rate. As a result, *AC* and *MC* decrease up to a certain level of output and then begin to increase. Thus, both *AC* and *MC* curve take a U-shape.

REVIEW QUESTIONS

1. Explain the following cost concepts with examples.

(*a*) Variable cost

(*b*) Fixed cost

(*c*) Explicit cost

(*d*) Implicit cost

2. Explain and illustrate the distinction between the following:

(*a*) fixed cost and variable cost

(*b*) actual cost and opportunity cost.

3. What is opportunity cost? Give some examples of opportunity cost. How is the concept of opportunity cost relevant for managerial decisions?

4. Which of the following statements is true?

(*a*) When $Q = O$, (*i*) $TC = TVC$; (*ii*) $TC > TVC$; or (*iii*) $TC < TVC$; (*b*) When $MC = O$, (*i*) *TC* is falling; (*ii*) *TC* is increasing; or (*iii*) *TC* is constant.

5. When the law of diminishing returns begins to operate, then

(*a*) *TVC* begins to fall at an increasing rate;

(*b*) *TVC* rises at a decreasing rate;

(*c*) *TVC* falls at a decreasing rate;

(*d*) *TVC* rises at an increasing rate; or

(*e*) *TVC* remains constant.

Which of the statements is true?

6. When *MC* changes, *AC* changes (*a*) at the same rate, (*b*) at a higher rate, or (*c*) at a lower rate? Illustrate your answer through a diagram.

7. Explain and illustrate the relationship between marginal cost, average cost, and total cost assuming a short-run non-linear cost function?

8. A manufacturing firm produces and sells 3,000 units of a product *X*, where its $AC = MC$ and makes only normal profit. The firm get an additional order of 500 units at the ruling price. Should the

firm, a profit maximizing one, accept or reject the order. Justify your answer by using imaginary cost curves.

9. (*a*) Why is short-run average cost curve U-shaped? Illustrate and explain the relationship between *SAC* and *SMC.*

(*b*) Suppose a cost function is given as $TC = 100 + 5Q + Q^2$. Find

(*i*) equation for *AC* and *MC*;

(*ii*) *AC* and *MC* for 10 units of output;

(*iii*) the output *Q* at which $AC = MC$.

10. What are the uses of cost function in business decisions? Suppose cost function of Yougesh Woollen Mills is given as follows.

$Q = 500 + 10Q + 5Q^2$

(*a*) Find the output that minimizes average cost.

(*b*) Suppose the firm is producing only 8 units of output. Should the firm increase or decrease the output?

11. Illustrate and explain the derivation of the long-run average and marginal cost curves. Explain the behaviour of the long-run average cost curve.

12. What is menat by learning curve? How is learning curve different from the theoretical long-run average cost curve?

13. (*a*) What is meant by the economies of scale? Distinguish between internal and external economies of scale.

(*b*) What are the sources of internal and external economies of scale of production?

14. (*a*) What is meant by diseconomies of scale? Why do diseconomies of scale arise?

(*b*) How do different diseconomies of scale affect the cost of production?

15. (*a*) What is meant by break-even analysis? What purpose does it serve in business decisions?

(*b*) Suppose cost and revenue functions of a firm are given as follows.

$TC = 500 + 20Q$ and $TR = 25Q$

Find the output that breaks *TC* and *TR* even.

16. Suppose sales and cost data of a company for a year are given as follows.

`

Net Sales	100,000
Cost of goods sold:	
Variable cost	40,000
Fixed cost	10,000
Gross profit	50,000
Selling costs:	
Variable cost	10,000
Fixed cost	5,000
Net Profit	35,000

(*a*) Compute the break-even point.

(*b*) Forecast the profit for the sale volume of `1,60,000 and `70,000.

(*c*) What would be sales volume to earn a net profit of `55,000?

17. (*a*) Discuss assumptions and limitations of break-even analysis.

(*b*) A firm has purchased a plant to manufacture a new product. Cost data for the plant is given

below:

Estimated annual sales

24,000 units

Estimated costs:

Material

`4.00 per unit

Direct Labour

`0.60 per unit

Overhead

`24,000 per year

Administrative Expenses

`28,000 per year

Selling cost of sales

`1,590 per year

(*i*) Calculate the selling price if profit per unit is `1.02 and

(*ii*) Find out the break-even point in terms of output.

18. Distinguish between the following:

(*a*) Marginal cost and incremental cost;

(*b*) Business cost and full cost;

(*c*) Actual cost and imputed cost;

(*d*) Private cost and social cost of private business.

19. From the following data find out: (*i*) PV ratio, and (*ii*) BEP.

Selling pricè50

Cost pricè40

Fixed cost

`5,000

[**Ans.** (*i*) 20; (*ii*) 250]

20. Which of the following statements are true?

(*a*) Economic rent is the same as economic profit.

(*b*) Imputed cost is the rent of hired building.

(*c*) When *AC* = *MC, AC* is minimum.

(*d*) When *MC* is rising *AC* > *MC.*

(*e*) Output is optimum when *AC* = *MC.*

[**Ans.** (*a*), (*c*) and (*e*)]

FURTHER READING

Baumol, William J., *Economics Theory and Operations Analysis,* (Prentice-Hall of India Pvt. Ltd., New Delhi, 1972), 3rd Edn., Chs. 9 and 10.

Bilas, RichaRd A., *Microeconomic Theory,* (McGraw-Hill, New York, 1971, 2nd Edn)., Chs. 6 and 7.

BRigham, EugEnE F. and JamEs L. PaPPas, *Managerial Economics,* (The Dryden Press, Hinsdale, Illinois., 1976), 2nd Edn., Chs. 4 and 5.

davis, R and S., chang, *Principles of Managerial Economics,* (Prentice-Hall, N.J., 1986), Ch. 12.

lEftWich, RichaRd H., *The Price System and Resource Allocation,* (The Dryden Press, Hinsdale, Illinois, 1973), Chs. 8 and 9.

mansfiEld, EdWin , allEn, W. BRucE, dohERty, a. nEil, and KEith, WEigElt, *Managerial Economics*, (W.W. Norton, New York, 5th Edn. 2002), Ch. 92.

nicholson, WaltER, *Intermediate Economics and Its Application,* (The Dryden Press, Hinsdale, Illinois., 1975), Chs. 3 and 5.

PEtERson, h. cRag and cRis, W. lEWis, *Managerial Economics*, 4th Edn., (Pearson Education, Indian Branch, 1999), Ch. 7.

salvatoRE, D., *Managerial Economics, In a Global Economy*, 4th Edn., (Thompson-South Western, 2001), Ch. 7.

sPEncER, M.H., *Managerial Economics*, (Richard D. Irwin, Homewood, Ill., 1968), Ch. 5.

Watson, donald S. and maRy A., hoiman, *Price Theory and Its Uses,* (Houghton Mifflin, Boston,
1978), Chs. 8–11.
WEBB, samuEl C., *Managerial Economics*, (Houghton Mifflin, Boston, 1976), Chs. 6–11.

APPENDIX

Some Empirical Cost Function Estimates

In previous sections, we have used some hypothetical cost functions and illustrated the behaviour of *AC* and *MC*. The question that arises now is: How do *AC* and *MC* actually behave in real life situations. The empirical studies that have investigated the actual behaviour of *AC* and *MC* in different kinds of industries, mainly in the US, UK and India, reveal that the behaviour of *AC* and *MC* in the short-run and long-run generally conforms to the theory of cost, though deviations are quite significant. Most long-run cost functions reveal the preponderance of L-shaped long-run average cost (*LRAC*). We present here the result of some empirically estimated cost functions to give an idea as to how *AC* and *MC* behave in actual practice.

Appendix 11.A *Results of Some Estimated Cost Functions*

Investigator

Year

Industry

*Type of Data**

Result

1. Short-Run Cost Functions

Johnston

1960

Electricity (US)

TS

AC falls at diminishing rate

tending to equal MC

Mansfield

1958

Railways (US)

—

MC constant

and Wein

Eiteman

1952

Manufacturing

Q

MC below AC prior

and Gutherie

to capacity

Lester

1946

Manufacturing

Q

Decreasing AVC

to capacity

Dean

1941

Hosiery

TS

MC constant

Dean

1941

Leather belts

TS

MC increasing

Ezekiel

1941

Steel

TS

MC declining

and Wylie

Yntema

1940

Steel

TS

MC constant

Dean

1936

Furniture

TS

MC constant

2. Long-Run Cost Function

Gupta

1968

Manufacturing

CS

AC L-shaped in

(India)

18 industries,

(29 industries)

U-shaped in 5 and

linear in 6

Nerlove

1961

Electricity

CS

LRAC declining then rising

Johnston

1960

Electricity (UK)

E

LRAC declining

Moore

1959

Manufacturing

E

Economies of scale:

declining LRAC

Albert

1959

Metal

E

Economies of scale

Bain

1956

Manufacturing

Q

Small economies

of scale

Gribbin

1953

Gas (UK)

CS

LRAC declining

Lomax

1951

Gas (UK)

CS

LRAC declining

Lomax

1952

Electricity (UK)

CS

LRAC declining

* *CS* = Cross section data; *TS* = Time series data; *E* = Engineering; *Q* = Questionnaire;

LRAC = long run average cost; *MC* = marginal cost

Source: A.A. Walters, "Production and Cost Function: An Econometric Survey: " *Econometrica*, 31, January-April 1963;

Vinod K. Gupta, Cost Functions, Concentration, and Barries to Entry in Twenty-nine Manufacturing Industries in India",

Journal of Industrial Economics, November 1968.

CHAPTER

12 Linear Programming

CHAPTER OBJECTIVES

In chapters 10 and 11, we have described theoretically the method and conditions of output optimization and cost minimization, respectively. In this chapter, we discuss linear programming – an important practicable technique of output optimization and cost minimization. By going through this chapter you will know the following aspects of linear programming.

• Basic concepts used in linear programming

• Formation of production, cost and profit functions to be estimated

• Application of linear programming technique for profit maximization

• Application of linear programming technique for cost minimization

• Application of Simplex Method – a more advanced technique to find solution to complex business problems

12.1 A MATHEMATICAL AID TO BUSINESS DECISIONS

We have noted in the preceding two chapters that all rational business decisions are oriented towards the optimization of resource allocation to achieve certain ends. For a firm, optimization essentially implies maximization of output or minimization of cost, given the resources and input prices. This eventually means maximization of profits. The conventional theory of production, as discussed in Chapters 10 and 11, offers optimum solution generally in abstract quantitative terms through diagrams and mathematical derivations, logic and symbols. The conventional theories provide only guidance to the decision-makers in their effort to optimize the resource allocation. They do not provide an exact solution to the practical problems of business. Besides, real life problems of maximization and minimization are much more complex than accounted for in the conventional theories of economics. These problems require real solutions in quantitative terms.

In actual business operations, decision-makers have to deal with a large number of variables with several constraints in their attempt to arrive at an optimum solution to the problem of resource allocation—machinery, men, material, finances, time and space. For example, suppose an automobile manufacturing company has a limited supply of steel ingot and a limited capacity to manufacture trucks, tankers, tractors, cars and jeeps, each having a different demand schedule. The company would be interested in knowing as to what should be

the optimum product-mix or allocation pattern of the limited resources so that output (or profit) is maximum. Consider another example. Food Corporation of India (*FCI*) procures foodgrains from different parts of the country; maintains stocks in godowns located at different centres and distributes foodgrains to the cities and towns and villages scattered all over the country. The *FCI* may be faced with the problem of optimizing the utilization of the available godown space, railway wagons; and also minimizing the transportation cost of procurement and distribution of foodgrains. For this purpose, the management needs to synchronize the movement of about 25 million tonnes of food grains with availability of railway wagons, godown facility, time frame of procurement and distribution of foodgrains. Obviously, the problem is complex. Often the choices are too many to be easily managed. The broad solutions offered by the conventional theories are not adequate and accurate enough to provide optimal choices. This problem is solved by applying a sophisticated mathematical technique, known as ***linear programming***.[1]

Linear programming is an important mathematical technique used in making business decisions.

It helps in measuring complex economic relations and, thereby, provides an optimum solution to the problem of resource allocation. Linear programming technique, thus, bridges the gap between abstract economic theories and managerial decision-making. The development and wide application of this technique to decision-making is attributed to the joint efforts of 'mathematicians, defence administrators, business executives, statisticians, and economists'. A comprehensive treatment of the subject is beyond the scope of this book.[2] In this chapter, we discuss briefly the technique of linear programming and illustrate its application to simple business problems. Our objective here is to present the concepts and technique of linear programming in a simple, non-technical manner with the aim of introducing the reader to the subject.

Definition and Nature

Linear programming, '*LP*' for short, is a mathematical technique for solving maximization and minimization problems involving variables having linear relationships with each other. Broadly speaking, linear programming is a technique to find an optimum solution to the problem of resource allocation to achieve certain ends under certain given conditions. It should be borne in mind that linear programming is a purely mathematical technique of getting an actual numerical solution to an optimization problem. Its economic content is nil in the sense that it provides little information regarding the working of the economy beyond what is provided by the conventional economic theories.[3]

12.2 SOME BASIC CONCEPTS USED IN LINEAR PROGRAMMING

Linear programming has its own language. It uses certain specific terminology and concepts with specific connotation in the formulation of the problem. Some of the basic concepts used in linear programming are discussed here briefly.

1. Linear programming was first formulated by a Russian mathematician, L.V. Kantorovich and, later, developed by G.B. Dantzig in 1947 as a technique of planning the diversified activities of the US Air Force. This technique is now widely used in finding solutions to constrained maximization and minimization problems in both economics and business decision-making.

2. A comprehensive treatment of 'Linear Programming' is supposed to be provided in 'Quantitative Techniques' or 'Quantitative Methods', a separate paper in B.B.A. and M.B.A. courses.

3. Baumol, W.J., *Economic Theory and Operations Analysis*, (Prentice-Hall of India (P) Ltd., New Delhi, 1978), 4th Edn., p. 72.

12.2.1 Objective Function

As mentioned above, all linear programming problems involve either minimization or maximization of a certain quantity. The quantity to be minimized or maximized is expressed through a function called 'objective function' or 'criterion function'. Objective functions are expressed in the form of equations. For example, for a firm producing and selling three goods X_1, X_2 and X_3, each yielding a unit profit of 3, 2 and 1, respectively, the profit maximization function (i.e., the objective function) may be expressed as

$$\text{Maximize } p = 3X_1 + 2X_2 + X_3$$

It may be noted here that minimization and maximization problems involving the same variables are 'dual' for one another because each minimization problem implies maximization of another quantity, and *vice versa*. For example, minimization of costs implies maximization of profits or revenues, and *vice versa*. Therefore, each minimization problem can be converted into a maximization problem, and vice versa.

12.2.2 Constrained Optimization

The objective functions of minimization or maximization problems are subject to certain constraints, which prevent the solutions being infinitely large or small. The constraints, also known as 'restraints', specify the limiting conditions that arise either out of limited resources or due to technological limitations. For example, maximization of production is constrained by the limited availability of inputs, e.g., number of machines, hours of work, quantity of raw materials, etc. Constraints are expressed in the form of equalities or inequalities. Supposing a firm has only 50 machine hours (M_1) and 100 man-hours (M_2) available to it, the constraints may be expressed as

$$M_1 \le 50 \text{ and } M_2 \le 100$$

12.2.3 Choice Variables

Choice variables are the variables that are chosen to minimize (or maximize) the objective function satisfying all the constraints. Each choice variable is an indicator of the level of an 'activity' or process. An activity or a process is physical operation, e.g., producing a commodity, buying or selling goods, transporting things, etc. A choice variable, however, need not necessarily be a physical activity. In many optimization problems, a choice variable may represent price which does not indicate a physical activity.

12.2.4 Non-negativity Condition

The non-negativity condition of the variables is an important requirement of linear programming. The significance of this condition lies in the fact that the variables with which businessmen deal, e.g., inputs like men, material, machines, space, output, etc., cannot be negative. These variables can be either equal to or greater than zero. The non-negativity condition with regard to a variable (v) is, therefore, expressed in the form of equality $v = 0$ or inequality, $v \geq 0$.

12.2.5 Feasible and Optimum Solutions

Feasible solutions are those that can be achieved with a given amount of resources under given constraints or limitations. For example, consider the 'budget line' in indifference

analysis of consumer behaviour or 'isocosts' in product analysis. All the possible combinations on or below the 'budget line' or 'isocosts' are the feasible solutions. Feasible solutions need not satisfy all the constraints. The solutions that satisfy all the constraints are known as 'optimum solutions' or 'optimum feasible solutions.'

12.3 ASSUMPTIONS OF LINEAR PROGRAMMING

Solution to an optimization problem through linear programming is based on the following basic assumptions.

12.3.1 Linearity

Linear programming problems assume a linear relationship between inputs and output. Linearity, i.e., linear input-output relationship is an assumption as well as a condition in linear programming solutions. In other words, linear programming solutions are based on the assumption of constant returns to a factor in the short-term. The linear relationship is given by a straight line equation. For example, suppose a car manufacturing company requires 50 man-hours (W), 25 machine hours (M) and 0.5 tonnes of steel (S) to produce one car (C). Then this input-output relationship is expressed as

$50\ W + 25\ M + 0.5\ S = 1\ C$

Linearity assumption means that application of linear programming technique is *limited* to the problems involving linear input-output relations. It implies that *LP* technique cannot be employed to solve problems that involve non-linear input-output relationships.

12.3.2 Continuity

Continuity of variables is another important assumption of linear programming. Continuity requires that all variables are quantifiable in terms of a numerical value. For, only numerical values provide continuity in measurements. Non-numerical or unquantifiable values are meaningless in an optimum solution.

12.3.3 Independence and Additivity

The variables and their quantitative specification are independent of other variables. It means that they should be capable of being arbitrarily chosen, given the constraints. Besides, they should also satisfy the *condition of additivity*, i.e., the property of being added together. Non-additive values cannot be used in linear programming solutions.

12.3.4 Proportionality

This condition is implied within the linearity condition. Proportionality simply implies that the relationship between the variables (e.g., between input and output) is proportional and does not change in the course of solution. In other words, proportionality means that the relationship between inputs and output is proportional and this relationship is fixed for all the levels of output. For example, if production of one unit of a commodity requires 2 units of an input, production of 10 units—a five times greater quantity of the commodity—would require 20 units (i.e., five times of input) and so on.

12.3.5 Constant Price

In linear programming problems, prices of inputs and outputs are assumed to be constant, irrespective of quantities bought and sold. It implies assuming a purely competitive price in both factor and commodity markets.

12.4 APPLICATION OF LINEAR PROGRAMMING TECHNIQUE

Having introduced 'linear programming' and described the basic conditions and assumptions of linear programming at an abstract level, we now proceed to show the application of linear programming. A better understanding and appreciation of the usefulness of this technique can be had by applying it to some practical business problems. In this section, we apply the linear programming technique to some simple profit maximization and cost minimization problems of business and illustrate how linear programming problems are formulated and solved. Let us first consider a simple profit maximization problem.

12.4.1 Profit Maximization Problem

To illustrate the application of linear programming technique to an elementary profit maximization problem, let us suppose that a firm produces two goods *X* and *Y*, with two inputs *A* and *B*. The total quantities of inputs *A* and *B* available per unit of time to the firm may be specified as *A* = 1600 units, and *B* = 2000 units. Let us also assume that producing one unit of *X* requires 4 units of *A* and 2 units of *B*, and one unit of *Y* requires 2 units of *A* and 5 units of *B*. Profits per unit of *X* and *Y* are estimated at `10 and `8, respectively. The overall production conditions may be summed up as presented in Table 12.1.

Table 12.1 *Production Conditions*

Inputs	*Total inputs available per time unit*	*Input requirement per unit of product* *X*	*Y*
A	1600	4	2
B	2000	2	5

Given the production conditions in Table 12.1, the firm intends to maximize its total profit (p). The problem confronted by the firm is to find an output-mix of *X* and *Y* that can yield maximum profit.

12.4.2 Formulation of Profit Maximization Problem in Linear Programming Mode

The solution to the above problem through linear programming technique requires restatement of the conditions of the problem in programming language mode. The following steps are involved in the formulation of the programming problem.

(i) Specification of the objective function. The first step in this regard is to formulate the objective function. The firm's objective function may be expressed as

Maximize p = 10 *X* + 8 *Y* …(12.1)

where *X* and *Y* denote quantities of goods *X* and *Y*. Their quantities multiplied by their corresponding constants (per unit profits) give the total profit. Linear programming technique helps in determining the values of *X* and *Y* that yield maximum profit.

(ii) Specification of constraint equations. The second step in formulating the linear programming solution mode is to specify the constraint equations. As Table 12.1 shows, 4 units of *A* are required to produce one unit of *X* and 2 unit of *A* are required to produce one unit of *Y*, and total supply of input *A* is given at 1600 units. Similarly, 2 units of *B* are required to produce one unit of *X* and 5 units of *B* to produce one unit of *Y* and total supply of input *B* is given at 2000 units. Given these conditions, the constraint equation with regard to input *A* may be specified as

$4X + 2Y \leq 1600\ A$

...(12.2)

and, the constraint equation with regard to input *B* may be specified as

$2X + 5Y \leq 2000\ B$

...(12.3)

These equations express the constraints regarding the inputs *A* and *B*, on the one hand, and specify the linearity assumption, on the other. Constraints 2, 4 and 5 in Eqs. (12.2) and (12.3) signify the constant returns to inputs *A* and *B*.

(iii) Specification of non-negativity conditions. Linear programming, being a mathematical tool for solving constrained optimization problem, may also turn out a negative quantity in the solution, if constraint equations involve inequality (£) signs. A negative quantity in an optimum solution is nonsensical. In order to eliminate 'negative' solutions, non-negativity conditions are finally imposed. In our example, the non-negativity condition may be expressed as

$X \geq 0$ and $Y \geq 0$

We are now in a position to present the whole linear programming problem in terms of equations. The entire *LP* problem may be expressed as follows.

The Problem

Maximize $\pi = 10X + 8Y$,

Subject to constraints,

(*i*) $4X + 2Y \leq 1600$

(*ii*) $2X + 5Y \leq 2000$

where $X, Y \geq 0$.

By solving these equations, we can get an optimum solution to our profit maximization problem. There are two methods of solving the problem: (*i*) graphical method, and (*ii*) simplex method. We illustrate first the use of graphical method. The use of simplex method will be illustrated in the subsequent section.

12.5 GRAPHICAL METHOD OF SOLVING LP PROBLEMS

12.5.1 Profit Maximization Problem

Graphical method is the simplest way of solving a simple *LP* problem. The first step in solving the problem of profit maximization by graphical method is to convert the constraint conditions from inequalities into equalities and to graph them. For example, constraint (12.2) above is changed from $4X + 2 \leq 1600$ to $4X + 2Y = 1600$. This equation can be easily graphed by locating any two or more points on a graph for the corresponding values of *X* and *Y*. A simple way to draw the constraint line is to locate the terminal points on the *Y* and *X* axes and

join them by a straight line. The terminal points can be located by assuming alternatively *X* and *Y* to be zero. If we assume $X = 0$, then $Y = 1600/2 = 800$ and if $Y = 0$, then $X = 1600/4 = 400$. These two points are marked by points *M* and *N*, respectively, in Fig.12.1. By joining these points by a straight line, we get the constraint line $4X + 2Y = 1600$, as shown in Fig. 12.1. All the points on and below the line *MN* satisfy the constraint $4X + 2Y \leq 1600$. The area *OMN* is the *feasibility area* or *feasibility*

Fig. 12.1 *Production Constraints and Feasibility Space*

space for single input *A*. It implies that any point within the feasibility space and on the border lines is a feasible point. The same procedure is adopted to graph the constraint $2X + 5Y \leq 2000$ for input *B*, and a constraint line *PQ* is obtained. The area *OPQ*, is the feasibility space for input *B*. All the points on the constraint line *PQ* and to the left of it satisfy the constraint $2X + 5Y \leq 2000$. Note that *PMR* area is relevant to only input *A* with no use of input *B*. Similarly, *NRQ* area is relevant to only input *B*, with no use of input *A*. But shaded area *OPRN* is relevant to both the inputs and both the outputs. Thus, the shaded area, *OPRN*, in Fig. 12.1 represents the feasible output choices. All these choices satisfy both the constraints and non-negativity conditions. That is, only those points, which fall on the boundary line of or inside the shaded area, satisfy all the feasibility conditions. Any point to the right of area marked by *MRQ*, represents a combination of *X* and *Y* that cannot be produced because both the inputs, *A* and *B*, would be inadequate. All the points in the area marked by *PMR* satisfy only constraint (12.2). Similarly, all the points in the area *NRQ* satisfy only constraint (12.3). It is only the feasibility space, i.e., the shaded area, that meets both the constraints and contains the point of solutions to the profit maximization problem.

The next step in solving the problem is to find the point within or on the boundary of the feasibility area, which represents a combination of *X* and *Y* yielding

Fig. 12.2 *Graphical Solution*

maximum profit. This can be done by graphing the objective function, that gives an **isoprofit line**, for different levels of output, and superimposing them over the feasibility region. To graph the objective function, it is necessary to find the slope of the objective function. For this purpose, the objective function $p = 10X + 8Y$ can be rewritten in terms of *Y* as

$$Y = p/8 - (10/8)X$$

Thus, where p = 0,

$$Y = -(10/8)X = -(5/4)X$$

The coefficient –5/4 gives the slope of the isoprofit line. It means that 1.25 units of *Y* yield the same profit as 1 unit of *X*. Given the slope, a series of isoprofit lines may be drawn and superimposed over the feasibility region, as shown in Fig. 12.2. Since profitability of the two goods is constant, isoprofit lines are parallel to each other. It is obvious from the figure that isoprofit marked p = `8000 cannot be achieved because it lies far beyond the feasibility space. The isoprofit lines, e.g., p = `2000 and `4000 passing through the feasibility space imply under-utilization of inputs. Therefore, they indicate a profit less than maximum. The shaded area to the right of these lines shows the scope for increasing profit. Note that the isoprofit line p = `4900 is tangent to the boundry of feasibility space at point *R*. The isoprofit line p = `4900 is, therefore, the only highest possible line representing the maximum possible profit given the resource constraints. The tangential point *R* represents a combination of *X* and *Y*, (i.e. , *X* = 250 and *Y* = 300) which yields maximum profit.

Maximum p = 10(250) + 8(300)

= `4900

The maximality of profit at `4900 can be checked in the following manner. Note that optimum solution lies on the corner points of the feasibility region, or on an upper segment of the feasibility boundary when the slope of objective function is equal to that of a constraint equation. All other points on the feasibility space yield a lower profit. This point can be proved by calculating the profit on any other point and comparing it with that at point *R*. Let us compare the combinations of two goods represented by each corner point and also the total profits therefrom. Combination of the two goods and the corresponding total profits are given in Table 12.2. Obviously, the profit at corner point *R* is maximum. ***Point R gives the optimum solution to the problem of profit-maximization.***

The optimality of the output-mix can also be checked by solving the constraint Eqs. (12.2) and (12.3) for two unknowns, X and Y.

Table 12.2 *Output-mix and Total Profits*

Corner Points	*Output-mix* $X + Y$	*Total Profits* $= 10(X) + 8(Y)$
O	0 + 0	10(0) + 8(0) = 0
P	0 + 400	10(0) + 8(400) = 3200
N	400 + 0	10(400) + 8(0) = 4000
R	250 + 300	10(250) + 8(300) = 4900

From constraint equation

$$4X + 2Y = 1600,$$

we get

$$X = \frac{1600 - 2Y}{4} = 400 - \frac{1}{2}Y$$

By substituting 400 – 1/2 Y for X in constraint $2X + 5Y = 2000$, we get

$2(400 – 1/2\ Y) + 5\ Y = 2000$

$4\ Y = 1200$ and $Y = 300$

By substituting 300 for Y in constraint equation

$4\ X + 2\ Y = 1600$

we get

$4\ X + 2(300) = 1600$

$4\ X = 1000$ and $X = 250$

Thus, we get the optimum output-mix as $X = 250$, and $Y = 300$. These quantities satisfy both input constraints and non-negative conditions and maximize the profit.

12.5.2 Cost Minimization Problem

In the preceding section, we have illustrated the application of linear programming technique to a simple profit maximization problem. Another typical problem frequently encountered in managerial decision-making is minimization of cost, given the constraints. We now illustrate the application of linear programming technique to a simple cost minimization problem through a simple hypothetical example.

Let us suppose that an automobile manufacturing firm produces only trucks and cars for which it uses only three inputs—labour, machine and steel. The firm gets the contractual supplies of inputs and, for compliance with agreement, it is required to make use of a minimum quantity of inputs, say, 160 man-hours, 36 machine hours and 48 tonnes of steel. The overall production conditions are given in Table 12.3.

Table 12.3 *Production Conditions*

Inputs	*Necessary minimum use of inputs*	*Input requirement per*	
		Car	*Truck*
Labour	160 (man-hours)	10	40
Machine	36 (machine-hours)	6	3
Steel	48 (tones)	4	8

Given the price of inputs, the production cost per truck has been worked out at `60,000 and the cost per car at `20,000. The task before the decision-makers is to work out the combination of trucks and cars, that can be produced at the minimum total cost. The problem may be reformulated in the linear programming mode as follows.

Minimize Cost $= 20{,}000\ C + 60{,}000\ T$ ...(12.4)

Subject to

$10\ C + 40\ T \geq 160$ (Labour) ...(12.5)

$6\ C + 3\ T \geq 36$ (Machine) ...(12.6)

$4\ C + 8\ T \geq 48$ (Steel) ...(12.7)

where $T \geq 0$ and $C \geq 0$

Graphical Solution

The graphical solution to this cost minimizing problem is illustrated in Fig. 12.3. The first step is to graph the input constraints following the procedure elaborated in the profit maximization problem. Consider, for example, the labour constraint equation, given as

$10C + 40T = 160$

If *T* is set to zero, *C* = 16, i.e., if the entire 160 man-hours are used to produce cars, 16 cars can be produced. This is indicated by point *M* in Fig. 12.3. Similarly, if *C* is set to zero, *T* = 4, i.e., 4 trucks can be produced as indicated by point *N* where *C* = 0. By joining points *M* and *N*, we get the labour-constraint line, *MN*. The line *MN* indicates the possible combinations of cars and trucks that can be produced by using 160 man-hours, other things given. The same procedure is used to graph the other constraint equations as shown by the lines, *JK* for steel and *TR* for machine. These lines are called **isocosts**. The area to the right of the isocosts, i.e., the shaded area, is the feasibility plane for larger amounts of inputs.

Fig. 12.3 *Cost-Minimisation: Graphical Solution*

The optimum solution to the problem of cost minimization lies either on the boundary line or in the shaded area. Thus, there are more than one solution. But only one of them is optimum. The optimum solution, i.e., the optimum combination of cars and trucks, can be obtained by drawing an isocost line *IC* having a slope of 3 (which is truck/car cost ratio = 60/30). The isocost (*IC*) which is tangent to the feasibility boundary at point *P*, offers the optimum solution, i.e., the combination of cars and trucks that minimize the total cost of production. The optimum solution is 8 cars + 2 trucks.

Table 12.4 *Output-Mix and Total Cost*

Corner Points	*Combination Cars + Trucks*	*Total cost 20,000(* C *) + 60,000(T) = C*
M	16 + 10	20,000 (16) + 60,000 (0) = 320,000
P	8 + 2	20,000 (8) + 60,000 (2) = 280,000
Q	4 + 4	20,000 (4) + 60,000 (4) = 320,000
R	0 + 12	20,000 (0) + 60,000 (12) = 720,000

The optimality of this solution can be checked by comparing the total cost for the combination of cars and trucks at each corner point *M*, *P*, *Q* and *R* because only one of these points offers the optimum solution. As shown in Table 12.4, the optimum solution to the problem of cost-minimization lies at point *P*. Incidentally, the solution cannot be checked algebraically because three simultaneous equations in our example involve two unknowns, *C* and *T*.

12.6 SIMPLEX METHOD

The graphical method of solving a linear programming problem discussed above has a serious limitation. It can be applied to only those problems that involve a small number of variables. It cannot be applied, with precision and accuracy, to the problems involving a large number of input and output variables. Business firms generally deal with a large number of variables. Taking into account a large number of variables makes business decisions much more complicated than visualized in our simple examples of profit maximization and cost minimization. The larger the number of outputs and inputs, the larger the number of feasible solutions. For instance, even a relatively small problem containing 'ten variables and five non-trivial constraints will have several thousand potential corners and hundreds of feasible corners'.4 It involves an enormous task of calculation, for which graphical method is inefficient.

Dantzig, a mathematician, developed a computational technique in 1947, known as ***simplex method*** to solve such problems. Simplex method is one of the most efficient and popular methods of solving linear programming problems. The simplex method is based on the premise that "an optimum solution can always be found among the corners of the feasible region", and hence "all other can be ignored in the calculation".5 It is a method for moving from one feasible corner (preferably the point of origin) to another in such a way that the objective function is always improved. Thus, the simplex method not only helps in finding the optimum solution automatically but also it indicates the existence of alternative optimum solutions.

The steps that are followed in using the simplex method6 to solve a *profit maximization problem* are listed below.

(*i*) First a corner point of the feasibility region is selected, preferably the 'origin', provided it is one of the feasible points.

(*ii*) The total profit at the selected point is then calculated and compared with the total profit computed for the adjacent points.

4. Strum, J.E., *Introduction to Linear Programming* (Holden Day Inc., San Francisco, 1972), p. 50.

5. Baumoli, W.J., *op. cit.*, p. 84.

6. For a simple and clear exposition of simplex method, students may refer to J.K. Thakural , *Mathematics for Business Students* (Mayur Paperback, New Delhi).

(*iii*) If profit at the adjacent points is higher, other points with still higher profits are considered. This process is repeated until the points of lower profit are eliminated and the one with maximum profit is obtained. In fact, it is the difference in profits that is calculated, not the total profits.

For an illustration, consider our objective function given in Eq. (12.1). It can be rewritten as

Maximize

$Z = 10X + 8Y$

and the constraints given in inequalities (12.2) and (12.3), i.e.,

$4X + 2Y \geq 1600$, $2X + 5Y \leq 2000$

where $X \geq 0$, $Y \geq 0$.

Let us now apply the simplex method to this problem of profit maximization.

Application of Simplex Method

The use of simplex method involves the introduction of an additional variable, known as *slack variable*. A slack variable is an additional variable added to the constraints. The basic function of the slack variable is to account for the amount of inputs that are unused at the point of solution. Slack variables are always non-negative because a negative slack variable will mean acquiring additional input, which will be a contradiction of the constraints.

The introduction of slack variables converts inequalities of constraint equations into equalities. For instance, after the introduction of slack variables, the constraints (12.2) and (12.3) can be rewritten in the form of equations as

$4X + 2Y + Sa = 1600$

$2X + 5Y + Sb = 2000$

where Sa, $Sb \geq 0$, and represent the unused quantities of inputs *A* and *B*, respectively.

It is important to note that introduction of slack variables converts the constraint inequalities into equalities and this makes it easier to test the feasibility of potential solutions by ascertaining the non-negativity of all the variables. Besides, it yields the following valuable information:

(*i*) The maximum values of slack variables Sa and Sb are 1600 and 2000, respectively, when X, $Y = 0$.

(*ii*) If Sa, $Sb = 0$ at the optimum solution, inputs are fully used and there is no excess capacity.

(*iii*) If Sa, $Sb > 0$, it indicates the existence of excess capacity.

12.6.1 Algebraic Solution

With the introduction of slack variables, Sa and Sb, the profit function,

$Z = 10\,X + 8\,Y$

can now be specified as follows:

Maximize

$Z = 10\,X + 8\,Y + (0)\,Sa + (0)\,Sb$

…(12.8)

Subject to constraints

(*i*) $4\,X + 2\,Y + Sa = 1600$

…(12.9)

(*ii*) $2\,X + 5\,Y + Sb = 2000$

…(12.10)

and $X \geq 0$, $Y \geq 0$, $Sa \geq 0$, $Sb \geq 0$

Now the problem is essentially to compute the set of values for X, Y, Sa and Sb that maximize Eq. (12.8) and satisfy the constraints (12.9) and (12.10). But, the problem, as stated above, cannot be solved as such since it involves only two equations with three unknowns. However, simplex method provides a solution to this problem. According to the simplex method, the optimum solution lies at a corner point of the feasibility region and at each corner the number of non-zero variables is exactly equal to the number of constraints. And the introduction of slack variables enables us to specify algebraically the potential corners of the optimum solution by setting any two of the three variables equal to zero and, thereby, reducing the number of variables to the number of equations. This enables us to rewrite the constraints as a system of two equations with two variables, which can be solved.

We may now determine algebraically the corner solutions of our example in the manner stated above. To begin with, let us consider first the constraint equations:

(*i*) $4\,X + 2\,Y + Sa = 1600$

and (*ii*) $2\,X + 5\,Y + Sb = 2000$

These equations contain all the four variables viz., X, Y, Sa and Sb. Although the value of any two variables can be set equal to zero, it is convenient to begin by setting the main variables X and Y equal to zero. This condition exists at the corner point of origin. When $X = 0 = Y$, then $Sa = 1600$ and $Sb = 2000$.

Substituting zero for X and Y in the objective function (12.8), we get

$Z = 10(0) + 8(0) = 0$

while

$Sa = 1600$, and $Sb = 2000$

Note that $Z = 0$, because $X = 0$ and $Y = 0$. If we increase the values of X and Y, profit will increase at the rate of 10 per unit of X and 8 per unit of Y. Here, a question arises: what are the upper limits of X and Y? The values of X and Y can be increased to the limit determined by the constraints. Suppose we decide to increase the value of Y.

The limit to which Y can be increased is given by the constraint Eqs. (12.9) and (12.10).

The constraint (12.9), i.e., $4\,X + 2\,Y + Sa = 1600$ allows the value of Y to increase to a maximum of 800, when X, $Sa = 0$. Similarly, constraint (12.10), i.e., $2\,X + 5\,Y + Sb = 2000$, allows us to increase Y only up to 400 when X, $Sb = 0$.

Note that $Y = 800$ satisfies constraint (12.9) only but $Y = 400$ satisfies both the constraints. Therefore, the upper limit of Y is 400, not 800. $Y = 400$ is an adjacent point ('to origin') where $X = 0$.

By substituting 400 for Y and 0 for X in the objective function, we get

$Z = 10(0) + 8(400) =$ `3200

Note that the objective is not to find the total profit at any of the corner points, but to know whether the total profit is maximum. Therefore, the question arises as to whether $Z =$ `3200 is maximum. There are two ways to find answer to this question.

First, to calculate the total profit at all the other corner points and compare them with `3200.

Second, to determine algebraically whether the total profit can be increased beyond `3200.

The simplex method follows the *second method*. Since Y has already been at a maximum permissible level under both the constraints, there is no scope for increasing profit by increasing Y. So the only possible way to increase the profit is to make $X > 0$. But

given the inputs constraints, X can be increased only by reducing Y. Hence, the change in profit will depend on the relations between X and Y and also on the slack variables. It is, therefore, required that the profit function be expressed in terms of X and slack variables, and also to find the limit to which X can be increased. The procedure of substituting X and Sb for Y in the profit function is illustrated below.

From Eq. (12.10), we get

$Y = 400 – 0.4\ X – 0.2\ Sb$

...(12.11)

Substituting Eq. (12.11) for Y in the objective function, profit function (12.8) may be rewritten in terms of X and Sb as $Z = 10\ X + 8(400 – 0.4\ X – 0.2\ Sb)$

$= 10\ X + 3200 – 3.2\ X – 1.6\ Sb$

$= 3200 + 6.8\ X – 1.6\ Sb$

...(12.12)

Equation (12.12) yields the following important information:

(*i*) at a point where X, $Sb = 0$, profit = `3200,

(*ii*) increasing X will increase profit since X is positive,

(*iii*) increasing Sb will decrease profit since Sb is negative.

This is the main logic of the simplex method. With this logic, we now proceed to solve the problem of profit maximization. As noted above, increasing X will increase the total profit, but up to a limit. This limit can be obtained through the two constraints. The constraint $4\ X + 2\ Y + Sa = 1600$ says that X can be increased to 400, and the constraint $2\ X + 5\ Y + Sb = 2000$ says that X can be increased upto 1000, when Y, $Sa = 0$. While the latter satisfies only one constraint, the former satisfies both the constraints (12.9) and (12.10). Therefore, 400 is the upper limit of X, when Y, *Sa* ¹ 0. But, since *Sb* ¹ 0, rather $Sb > 0$, the profit is not maximum, because some units of input B remain unused. What is, therefore, required is to compute the profit at the point where $Sa = 0 = SB$. For this, the profit function has to be expressed in terms of Sa and Sb, and Eq. (12.9) to be solved for X in terms of Sa and Sb as follows:

By substituting Eq. (12.11) for Y in constraint $4\ X + 2\ Y + Sa = 1600$, we get

$4\ X + 2\ (400 – 0.4\ X – 0.2\ Sb) + Sa = 1600$

or

$3.2\ X = 800 + 0.4\ Sb – Sa$

$X = 250 + 0.125\ Sb – 0.312\ Sa$

...(12.13)

By substituting Eq. (12.13) for X in the profit function (12.12), we get

$Z = 3200 + 6.8\ (250 + 0.125\ Sb – 0.312\ Sa) – 1.6\ Sb$

$= 3200 + 1700 – 0.75\ Sb – 2.12\ Sa$

$= 4900 – 0.75\ Sa – 2.12\ Sa$

...(12.14)

Equation (12.14) gives the total profit at the corner $Sa = Sb = 0$. To increase profit further, Sa or Sb must be increased. But these variables are negative. Hence making Sa, $Sb > 0$, will reduce the profit. Therefore, Sa and Sb must be set as Sa, $Sb = 0$. The maximum profit will thus bè4900 when Sa, $Sb = 0$.

12.6.2 Simplex Tableau Method

A simplex tableau is a systematic arrangement of coefficients of variables used in the linear programming problem. From a simplex tableau point of view, our problem with slack variables may be specified as follow.

Maximize:

$Z = 10\ X + 8\ Y + (0)\ Sa + (0)\ Sb$

...(12.15)

Constraints:

$4\ X + 2\ Y + 1\ Sa + (0)\ Sb = 1600$

...(12.16)

2 *X* + 5 *Y* + (0) *Sa* + 1 *Sb* = 2000

…(12.17)

The value of *X*, *Y*, *Sa* and *Sb* are assumed to be non-negative. The above information may now be arranged in a matrix form, known as *simplex tableau*. It provides a more convenient technique of solving a linear programming problem. The information contained in Eqs. (12.15) through (12.17) may be presented in a tabular from as given in Simplex Tableau 1, where *Cj* represents the coefficients of constraints.

Simplex Tableau 1

Cj →	*10*	*8*	*0*	*0*	*Solution*
↓	*X*	*Y*	*Sa*	*Sb*	*value*
0	4	2	1	0	1600
0	2	5	0	1	2000

The above tableau contains all the information contained in our problem. The body of the table contains the constraint-equation coefficients—each constraint in one row. The variables are shown as column headings. Once the required information is arranged in a tableau, a *set of rules* are followed to solve the problem.

Rules for Simplex Solution

The rule of basic solution. The first computational rule is to check whether there exists a basic solution to the problem. A basic feasible solution exists, if and only if, all the constant terms, i.e., coefficients in the constraint equations are non-negative. The basic solution can be obtained by setting all the ordinary variables (e.g., *X* and *Y*) equal to zero and then checking whether slack variables, viz., *Sa* and *Sb*, are positive. In our example, as shown in simplex tableau 1,

4 *X* + 2 *Y* + 1 *Sa* + (0) *Sb* = 1600 and 2 *X* + 5 *Y* + (0) *Sa* + 1 *Sb* = 2000

If we set *X* = 0 = *Y*, then *Sa* = 1600 and *Sb* = 2000.

This information is incorporated in the Simplex Tableau 2.

Simplex Tableau 2: *Basic Solution*

Cj	*Basic*	*10*	*8*	*0*	*0*	*Solution*
	variables	*X*	*Y*	*Sa*		

Sb

Value

0

Sa

4

2

1

0

1600

0

Sb

2

5

0

1

2000

Zj

0

0

0

0

0

Cj – *Zj*

10

8

0

0

The column *Cj* represents the coefficients of basic variables which appear in the solution. In addition to this, two other rows, viz., *Zj* and *Cj* – *Zj* are added to the bottom of the tableau as a matter of computational device. In order to find the value for *Z* for any given column *j*, the first rule is to multiply each constraint coefficient by corresponding coefficients of basic variables as given in column *Cj* and then to sum up the products. For example, $Z_1 = (4)(0) + (2)(0) = 0$ and so on. The *Cj* value is written at the top of each column. To obtain the value of *Cj* – *Zj*, simply subtract the value of *Zj* from *Cj*. These two rows play a central role in finding the optimal solution to the problem.

The initial solution shows zero profit since the value of objective function is

$Z = 10X + 8Y + (0)Sa + (0)Sb = 0$

Given the above solution, *X* = 0, *Y* = 0, *Sa* = 1600, *Sb* = 2000, and firm's profit is zero. But the firm's objective is not to make zero profits. The importance of this solution, however, lies in that it gives a starting point for the optimal solution. The following three distinct operations help us in arriving at an optimal solution:

(*i*) applying optimality rule to test the solution for optimality;

(*ii*) applying pivoting procedure to identify the incoming and outgoing variables;

(*iii*) revising the simplex tableau to develop a new solution.

1. *The Optimality Rule.* The optimality rule states that solution is optimal if $Cj - Zj \le 0$ for all *j*. That is, if every value in (*Cj* – *Zj*) row has a negative or zero value then the solution is optimum. By applying the optimality rule, we can test the optimality of the problem. For instance, in Simplex Tableau 2, there are two non-zero and non-negative values (10 and 8) in *Cj* – *Zj* row. Therefore, the row *Cj* – *Zj* does not provide an optimum solution to the problem. This condition of optimality implies that so long as the coefficient of a variable in the objective function is positive, an additional unit of the variable will increase the profit.

To arrive at the optimum solution, a standard computational procedure is followed to move from one basic solution to another, i.e., moving from one corner of the feasibility area to the other corner point. This means that with each change in solution, one of the basic variables must be removed (called *outgoing variable*) and a new variable brought in (called *incoming variable*).

2. ***Pivoting Procedure.*** In the pivoting procedure, the incoming variable is identified first. In case of maximization problems, the variables with the largest positive value in *Cj* – *Zj* row is taken as incoming variable. In our example, the largest positive value of (*Cj* – *Zj*) is 10 (See Simplex Tableau 3). Therefore, the variable *X* is incoming variable and the column under *X* is defined as 'pivot column'.

To identify the outgoing variable, first we divide the solution value (for each basic variable) by the corresponding row coefficient in the pivot column. The resultant values are given in Simplex Tableau 4.

Simplex Tableau 3

Cj

Basic

10

8

0

0

Solution values

variable

X

Y

Sa

Sb

1600

0

Sa

4*

2

1

0

600 4 = 400

Outgoing variable (pivot row)

0

Sb

2

5

0

1

2000 2000 = 1000

2

Zj

0

0

0

0

Cj – *Zj*

10

8

0

0

Incoming variable (pivot column)

The row corresponding to the smallest solution value (only positive values are considered) is defined as the pivot row. The basic variable in pivot row becomes the outgoing variable. In our example, 400 is the smallest positive value. Therefore, the row containing solution value 400 is the first pivot row, and *Sa* is the outgoing variable. The element where pivot row and pivot column intersect is defined as the pivot *element*. The pivot element in Tableau 3 is 4 marked be asterisk (*). The pivot elements have been marked by asterisk in the following tableaus also.

3. *Revising the Simplex Tableau.* Once the pivot element is identified, the next step in obtaining the optimal solution of the problem is to revise the non-optimal simplex tableau. The following procedure is adopted in revising the tableau.

(*a*) Divide all the elements of the pivot row by the pivot element (4*). This gives a revised row.

(*b*) Elements in other rows are revised by using the following formula:

$$\begin{bmatrix}\text{Elements in}\\ \text{revised row}\end{bmatrix} = \begin{bmatrix}\text{Elements}\\ \text{in}\\ \text{old row}\end{bmatrix} - \begin{bmatrix}\dfrac{\begin{bmatrix}\text{Corresponding}\\ \text{elements in}\\ \text{pivot row}\end{bmatrix} + \begin{bmatrix}\text{Corresponding}\\ \text{element in the}\\ \text{row in the}\\ \text{pivot col.}\end{bmatrix}}{\text{pivot element}}\end{bmatrix}$$

In Simplex Tableau 4, the first row has been obtained by dividing all the elements by 4, the pivot element. The elements in second row have been obtained by applying the same formula, i.e.,

$$0 = 2 - \left[\frac{4 \times 2}{4}\right]$$

$$4 = 5 - \left[\frac{2 \times 2}{4}\right]$$

and –1/2 = 0 – (1/4 × 2), and so on.

The revised tableau after applying the above procedure is given in Tableau 4.

Simplex Tableau 4

Cj		*10*	*8*	*0*	*0*	
Basic variable		*X*	*Y*	*Sa*		*Solution values*

Sb

10

X

1

1/2

1/4

0

1600 400 = 800

1 2

1200

0

Sb

0

4*

– 1/2

1

200 4 = 300

Zj

10

5

5/2

0

Cj – Zj

0

3

– 5/2

0

Now, again by inspecting the row (*Cj – Zj*) we can know whether we have obtained the optimal solution. For a solution to be optimal in our problem, all *Cj – Zj* must be less than zero. We notice that there is one element (3) greater than zero in *Cj – Zj* row and, therefore, the column above the element 3 is the pivot column. Applying the pivoting procedure, as described above, we find 4* as the pivot element.

4. *Revising the Simplex Tableau Again.* After the first revision, the solution obtained is not optimal. Therefore, having found the pivot element (4), we will further revise the simplex tableau following the same procedure. Given below is the further revised Tableau 5.

Simplex Tableau 5

Cj

Basic variable

10

8

0

0

Solution values

X

Y

Sa

Sb

10

X

1

0

5/12

– 1/8

250

8

Y

0

1

– 1/8

1/4

300

Zj

10

8

19/6

3/4

Cj – *Zj*

0

0

– 19/6

– 3/4

When we examine Simplex Tableau 5, we find that all the values in row *Cj* – *Zj* are less than zero. According to the optimality rule, therefore, we have found the optimal solution. In the optimal solution of the problem,

$X = 250$ and $Y = 300$

With these values of X and Y, the value of Z is maximum, i.e.,

$Z = 10(250) + 8(300) = 2500 + 2400 = 4900.$

12.7 DUALITY OF LINEAR PROGRAMMING

Economists and mathematicians[7] have developed the technique of converting every linear programming problem into its *dual*, i.e., the reverse form of the original problem. By using this technique, every maximization problem can be converted into its dual, and every minimization problem into its dual, and *vice versa*. The original problem is conventionally called the *primal* and its converted form as the *dual*.

7. The basic idea of duality was oriented by John Von Neumann and was later rigorously developed by David Gale, and H.W. Kuhn and A.W. Tucker.

12.7.1 Why Dual?

The concept of duality in linear programming has gained significance for the following reasons:

1. The dual linear programming problem is much easier to be solved than its primal problem. Once the dual problem is solved, it becomes much easier to obtain the solution of the primal problem.
2. Duality yields a number of powerful theorems which add substantially to our understanding of linear programming.
3. The dual problem has an ‘extremely illuminating economic interpretation,’ as it reflects the involvement of marginal analysis approach in seeking optimum solution to a linear programming problem.
4. The solution to a dual problem also reveals the impact of the relaxation or changes in the constraints.

12.7.2 Converting a Primal into a Dual

As noted above, a *primal* can be converted into its dual—a primal maximization into its *dual* of minimization problem and *vice-versa*. We show here how a maximization problem can be converted into dual by minimization problem. Let us suppose that in a maximization problem, the primal is given as below:

The Primal Problem

Maximize $p = 50X_1 + 20X_2$

subject to

$5X_1 + X_2 \leq 100$

(Land = V_1)

$X_1 + 6X_2 \leq 60$

(Capital = V_2)

$3X_1 + 4X_2 \leq 120$ (Labour = V_3)

This problem can now be converted into a dual problem following the steps outlined below:

1. The ‘maximization’ in the objective function of the *primal* is converted into minimization in its dual.
2. The choice variables in the objective function of the dual are defined in terms of

capacity variables, say V_1, V_2 and V_3 for land, capital and labour, respectively, so that the number of choice variables in the dual objective function is equal to the number of constraints in the primal.

3. The profit constants (50 and 20) are replaced by the capacity constraints (100, 60 and 120 to become coefficients of V_1, V_2, and V_3, respectively).

4. In the constraint inequalities for the dual, the row-wise coefficients of the primal are written as the column-wise coefficients, number of constraints in the dual being equal to the number of choice variables in the primal.

5. The 'less than or equal to' sign ($\leq$) in the constraint inequalities of the (maximization) primal are replaced by 'more than or equal to' sign ($\geq$) in the (minimization) dual.

6. The profit constraints (or coefficient) in the objective function of the primal become *R* (Rent), *H* (Hours) and *W* (Wages) of the constraints in the dual.

The primal converted into its dual following the above rules can be expressed as follows.

The Dual of the Primal Problem

Minimize $Z = 100 V_1 + 60 V_2 + 120 V_3$

subject to

$5 V_1 + V_2 + 3 V_3 \geq 50$

$V_1 + 6 V_2 + 4 V_3 \geq 20$ and $V_1 \geq 0$, $V_2 \geq 0$, $V_3 \geq 0$

The points of distinction between the primal and its dual are:

(*i*) the objective function of the primal involves only two choice variables whereas dual objective function involves three variables, and

(*ii*) the primal involves three structural constraints containing two variables and the dual involves two constraints, containing three variables.

Following the conversion rules, one may consider the dual as the primal problem and construct its dual which will be exactly the same as the original problem.

Furthermore, by adding *slack variables*, the structural constraints may be rewritten as equations in order to make the problem soluble. The primal with slack variable can be constructed into its dual following the procedure illustrated above. It should be noted that inequalities in primal constraints are expressed by $\leq$ sign and hence the slack variables have a positive sign (+). But in dual problems, constraint inequalities are expressed by $\geq$ sign, and hence, slack variables in dual will take a negative sign. The dual problem can then be solved through the simplex method.

12.7.3 Duality Theorems

Duality theorems are useful in understanding the nature of dual linear programming. Besides, the theorems also work, in a way, as the tests of optimality of the solution. The two basic theorems are following.

Theorem I: *In the solved primal and dual problems the maximum value of* p *must be equal to the minimum value of Z, and any pair of feasible solutions for which Z =* p *must be optimal.*

This theorem implies that when the optimum solutions of primal and dual problems are obtained, the total value of *Z* can be assigned to the scarce inputs used, which is exactly equal to the total value of p. In other words, this theorem states that, in an accounting sense, the firm's total profit will never exceed the accounting value of its scarce inputs.

Theorem II: *Optimum solution requires the firm to produce those commodities whose accounting loss figures are zero and the scarce inputs having non-zero shadow-cost must be used to capacity*. This theorem may be symbolically written as a set of equations as $Q_j L_j = 0$ for each commodity, *j*, and $U_i V_i = 0$ for each product, i.e., [where, Q_j = quantity of commodity *j* in the primal; L_j = accounting loss per unit of output, *j* is the dual slack variable; U_i = the unused capacity of inputs—primal slack variables, and V_i = the value (accounting price) of input.].

The significance of theorem II lies in its recommendatory value. Note that $Q_j L_j = 0$, for each *j* th commodity, either *Q* or *L*, or both must be zero.

It means that if $Q_2 > 0$, then we must have $L_1 = 0$, i.e., the accounting loss must be zero. Similarly, if $V_1 > 0$, in equation $U_1 V_1 = 0$, then U_1 must have zero value, or else there will be unused input capacity. Since in an accounting sence, excess capacity (unused capacity of an input) is imputed with a zero value (i.e., excess capacity is treated as loss), for $L = 0$, there should not be any excess capacity of inputs. It also implies that if a commodity has non-zero accounting loss, it should not be produced from the optimality point of view.

SUMMARY

- Linear programming is a mathematical technique for solving maximization and minimization problems involving interrelated and interdependent variables.
- Some prerequisites of using linear programming techniques are (*i*) determining the objective function, (*ii*) specifying the constraints, if any, (*iii*) making choice of relevant variables, and (*iv*) fixing nonnegative conditions of variables.
- The following assumptions are made to apply the linear programming: (*i*) variables have linear relationship, (*ii*) variables are quantifiable, (*iii*) variables are interdependent and additive, and (*iv*) prices remain constant.
- Linear programming can be applied to such business decision-making problems as profit maximization under the constraints of limited inputs, (*ii*) cost minimization problem with input constraints, and (*iii*) output maximization problems under input constrains.
- There are three methods of linear programming: (*i*) graphical method, (*ii*) algebraic method using equations, and (*ii*) simplex method. Graphical and algebraic methods are simple techniques used to solve the simple minimization and maximization problems. Simplex method is a highly technical and most widely used technique of linear programming.

REVIEW QUESTIONS

1. What is linear programming technique? Discuss usefulness of its application to practical business problems. Illustrate your answer by using a linear programming problem of production planning.

2. Maximize

$\pi = 3X + 6Y$

4. Maximize

$Z = 3X_1 + 5X_2$

subject to $3X + Y \leq 48$

subject to

$X_1 + 3X_2 \geq 9$

$X + 3Y \leq 48$

$X_1 + X_2 \geq 5$

where

$X \geq 0, Y \geq 0$

$2X_1 + X_2 \geq 6$

[**Ans.** p = 108]

where

$X_1 \geq 0, X_2 \geq 0$ [**Ans.** p = 108]

3. Maximize

$\pi = 5X_1 + 6X_2 + 7X_3$

subject to

X

5. Maximize

$p = 12X$

$1 + X_2 \leq 10$

$1 + 9X_2$

$2X$

subject to

$4X$

$1 + X_3 \leq 18$

$1 + 2X_2 \leq 32$

where

X

X

$1 + X_2 + X_3 \geq 0$

$1 + X_2 \leq 10$

[**Ans.** p = 186]

$3X_2$ [illegible] 21

where

$X_1 \geq 0, X_2 \geq 0$

6. Construct dual problems from Exercises 2 to 5, solve them, and compare your results.

7. Show the solution zone of the following inequalities on a graph paper.

$5X + y \geq 10;$

$X + y \geq 6$

$X + 4y \geq 12;$

$X \geq 0, y \geq 0$

Find X and y for which $3X + 2y$ is minimum, subject to these inequalities. Use the graphical method.

8. A firm produces three item, *A*, *B* and *C* at two plants, *X* and *Y*. The number of items produced at an operating cost per hour are given as follows.

Plants	*Item Produced Per hour*			*Operating cost*
	A	*B*	*C*	*per hour*
X	30	60	40	`1200
Y	30	20	100	`800

It is necessary to produce at least 400 items of type *A*, 800 of type *B* and 880 of type *C* per day. Find by simplex method using the concept of dual, the number of hours each plant must run on a day to minimize the costs.

9. What is a 'dual' of a linear programming problem?

10. (*a*) State the utility of linear programming in business decision-making.

(*b*) A toy company manufactures two types of dolls; a popular doll *A* and a deluxe type doll *B*. Each doll of type *B* takes twice as long to produce one doll of type *A*, and the company would have time to make a maximum of 2000 dolls per day if it produced only the popular type. The supply of plastic is sufficient to produce 15,000 dolls per day (both *A* and *B* combined). The deluxe version requires a fancy dress of which there are only 600 per day available. The company makes a profit of `3 and `5 per doll, respectively, on dolls *A* and *B*. Formulate the linear programming problem to determine the most profitable combination of dolls *A* and *B*.

Product	*Machine A*	*Machine B*
X	4 hours	5 hours
Y	5 hours	2 hours

11. (*a*) Discuss the various ways in which linear programming might be used in aid of business decisions.

(*b*) A manufacturer produces two products *X* and *Y* in two steps on machines *A* and *B*. The processing times for the two products on the two machines are:

Machine *A* has 80 hours available and *B* has 120 hours. Product *X* has profit of `10 per

100 pieces and *Y* `5 per 10 units. There is no restriction on sales. Formulate the linear

programming model and find by simplex method how much of *X* and *Y* must the manufacturer

produce so as to maximize his profits.

12. A producer produces three commodities *A*, *B* and *C*

and has limited inputs—storage and petrol.

Input needs

Products

A

B

C

Storage space = 8 units and petrol = 3 units.

Storage

2

2

2

Profit per commodity (per unit)

Petrol

1

1

0

A: `3, *B*: `7, and *C*: `6.

Find the product mix that maximizes profit.

FURTHER READING

Baumol, W.J., *Economic Theory and Operations Analysis*, 3rd Edn., (Prentice-Hall of India Pvt. Ltd., New Delhi, 1972), Chs. 5, 6.

Dantzig george, B., *Linear Programming and Extensions*, (Princeton University Press, Princeton, N.Y., 1963), Chs. 1–6, 11, 27.

Dorefman, r., P.a. SamuleSon, and r.m., SoloW, *Linear Programming and Economic Analysis*, (McGraw-Hill, N.Y., 1958), Chs. 2, 4, 7.

gaSS Saul, i., *Linear Programming*, 3rd Edn., (McGraw-Hill, N.Y., 1969), Chs. 3–5, 11.

loomBa n. Paul, *Linear Programming— An Introductory Analysis*, (Tata-McGraw-Hill, Bombay, 1964).

PeterSon, H. Crag and CriS leWiS, *Managerial Economies* (Pearson Education; 4th Edn. 1999), Ch. 8.

Salvatore, DominiCk, *Managerial Economies in a Global Economy* (Thompson South Western, 4th Edn., 2001), Ch. 8.

Strum Jay, e., *Introduction to Linear Programming*, (Holden-Day, Inc., San Francisco, 1972), Chs. 1–3, 5, 12.

Part IV

MARKET STRUCTURE, PRICING THEORY AND PRACTICES

CHAPTER

13 Market Structure and Pricing Decisions

CHAPTER OBJECTIVES

The objective of this chapter is to have an overview of market system and the kinds of markets in which firms play their role and find ways and means to determine the price of their product. This chapter presents a brief discussion on the follows aspects of the market system.

• Concept of ‘market’ in analytical sense

• Market system — how market works by demand and supply

• Demand side of the market

• Supply side of the market

• Market mechanism — how market forces determine the market equilibrium

• Market structure — the different kinds of markets and degree of competition

13.1 INTRODUCTION

We have discussed the theory of production—the input-output relations—in Chapter 11 and theory of cost—the cost-output relations—in Chapter 12. Knowing input-output relations and output-cost relations is not the objective of business firms. Profit maximization is the objective of business firms, as assumed in traditional theory of firm. Profit is maximized when $TR – TC$ is maximized. Since $TR = P \times Q$, where P = price, and Q = quantity sold, given the cost-output relation, the main task for profit maximization is to find the profit maximizing price. This takes us to the *theory of price determination*. The objective of this Part of the book is to discuss the theory of price determination under different market conditions. The theory or model of price determination developed by economic pundits is based on the market conditions. Market conditions are determined by the number of firms in an industry and the nature and degree of competition among the firms. The number of firms and degree of inter-firm competition determine the *market structure* of the industry. The economists have proposed different kinds of theoretical models for price determination on the basis of the nature and the characteristics of the market. Market conditions provide the *playing field* of the business firms. Therefore, before we discuss the theory of price determination, let us have a clear and comprehensive view of the market mechanism, market structure and firm’s power of pricing decisions. Let us begin by knowing economic ‘concept of market’, i.e., ‘what is market’ in economic sense.

13.2 THE CONCEPT OF MARKET

In general usage, the term ‘market’ means a place or locality where goods and services are bought and sold, e.g., Connaught Place market, Chandani Chowk market, Karol Bagh Market, etc. The term ‘market’ is also used with reference to a product, e.g., wheat market, vegetable market (Sabji mandi), gold market, stock market, etc.

In economics, the word ‘market’ is used in a rather abstract sense. The market means a system by which sellers and buyers of a commodity interact to settle its price and the quantity to be bought and sold. According to Samuelson and Nordhaus, “A market is a mechanism by which buyers and sellers interact to determine the price and quantity of a good or service”. Market for a commodity consists of the buyers and sellers who interact to settle its price and quantity to be transacted. The sellers and buyers may be individuals, firms, factories, dealers and agents.

Some important aspects of the market concept are the following ones:

(*i*) A market need not be situated in a particular place or locality. The geographical area of a market depends on how far and wide are the buyers and sellers scattered. It may be as small as a fish market in a corner of a city or as large as the entire world, e.g., the global markets for arms, cars, electronic goods, aeroplanes, computers, oil, medicines, etc.

(*ii*) Buyers and sellers need not come into personal contact with each other. The transactions can be carried out through postal services or telecommunication system—telephone, fax, agents, or e-mail etc. People do buy many goods and services without ever meeting the supplier.

(*iii*) The word ‘market’ may refer to a commodity or service (e.g., fruit market, car market, share market, money market, paper market, labour market, etc.) or to a geographical area, Bombay market, Indian market, Asian market, or the world market.

(*iv*) The economists distinguish between markets also on the basis of (*a*) nature of goods and services, e.g., factor market, input market and output market; (*b*) number of firms and degree of competition, e.g., perfectly competitive market (very large number of firms), monopolistic market (many firms with differentiated products); oligopolistic market, and so on.

13.3 THE MARKET SYSTEM: AN OVERVIEW

Market system refers to the process by which buyers and sellers of a product interact to settle the price of the product and carry out the sale-purchase transactions. The market system works on the basis of a basic ***market principle.*** The market principle is based on the ***fundamental laws of demand and supply***. Buyers create the demand and sellers create the supply of the product. Demand and supply work as two opposite ***market forces.*** The working of the market system is governed by the fundamental ***laws of demand and supply***.

The ***laws of demand and supply*** play a crucial role in determining the price of a product and the size of the m arket—total demand and total supply. A clear understanding of how markets work is essential for business decisions on production and future planning. As noted above, market system works on the basis of two ***market forces—demand and supply.*** The demand and supply forces represent two sides of the markets, viz., (*i*) *Demand side,* and (*ii*) ***Supply side***. The demand and supply sides work on the laws of demand and supply, respectively.

The law of demand and the derivation of the demand curve have already been discussed in detail in Ch. 7 and the law of supply and derivation of supply curve in Ch. 8. However, before we proceed to discuss price determination under different market setting, it will be helpful to have a brief review of market system, market mechanism and market structure, i.e., the different kinds of market settings.

13.4 DEMAND SIDE OF THE MARKET

The demand side of the market for a product refers to all its consumers and the price that they are willing to pay for buying a certain quantity of the product during a period of time. The quantity that consumers buy at a given price determines the market size. It is the size of the market that determines the business prospects of a firm and an industry. The demand side of the market is governed by the ***Law of Demand***. The law of demand governs the market in the sense that when prices go up, demand goes down and size of the market in reduced, all other things remaining the same. Similarly, when prices decrease, demand increases causing a rise in sales and market size tends to increase. Let us now look at the law of demand in some detail and see how it governs the demand side of the market.

13.4.1 Law of Demand

The law of demand states the nature of relationship between the quantity demanded of a product and the price of the product. Although quantity demanded of a commodity depends also on many other factors, e.g., consumer's income, price of the substitutes and complementary goods, consumer's taste and preferences, advertisement, etc., the current price is the most important and the only determinant of demand in the short run. But the law of demand states the relationship between the demand for a product and its price only. The ***law of demand*** can be stated as *all other things remaining constant, the quantity demanded of a commodity increases when its price decreases and decreases when its price increases*. This law implies that demand and price are inversely related. Marshall, the originator of the law, has stated the law of demand as "the amount demanded increases with a fall in price and diminishes with a rise in price". This law holds under *ceteris paribus* assumption, that is, *all other things remain unchanged*. The law of demand can be illustrated through a *demand schedule* and a *demand curve* as shown below .

13.4.2 Demand Schedule

A *demand schedule* is a tabular **Table 13.1** *Monthly Demand Schedule for Shirts*

presentation of different prices of

a commodity and its corresponding

Pc (Price in `)

Qc (Delhi in '000)

quantity demanded per unit of time.

A hypothetical annual market demand

800

8

schedule for shirts is given in Table

600

15

13.1. This table presents price of shirts

400

30

(*P*

300

40

c) and the corresponding number of

shirts demanded (*Q*

200

55

c) in Delhi per week.

Table 13.1 illustrates the law

100

80

of demand. As data given in the

table shows, the demand for shirts (*Qc*) increases as its price (*Pc*) decreases. For instance, at pricè800 per shirt, only 8 thousand shirts are demanded in Delhi per week. When price decreases tò400, the demand for shirts increases to 30 thousand and when price falls further tò100, demand rises to 80 thousand. This relationship between price and quantity demanded gives the law of demand.

13.4.3 Demand Curve

A *demand curve* is a graphical presentation of the demand schedule. A demand curve is obtained by plotting a demand

Fig. 13.1 *The Demand Curve for Shirts*

schedule. For example, when the data given in the demand schedule (Table 13.1) is presented graphically as in Fig. 13.1, the resulting curve *DD′* is the demand curve. The curve *DD′* in Fig. 13.1 depicts the law of demand. It slopes downward to the right. It has a negative slope. The negative slope of the demand curve *DD′* shows the *inverse relationship* between the price of shirts and their quantity demanded.

13.5 SUPPLY SIDE OF THE MARKET

In a market economy, buyers' decision on 'what to buy' and 'how much to buy' constitute the demand side of the market, and sellers' decision to 'what to sell' and 'how much to sell' make the supply side of the market. In this section, we discuss the supply side of the market beginning with the law of supply.

13.5.1 Market Supply

Supply means the quantity of a commodity that its producers or sellers offer for sale at a given price, per unit of time. Market supply, like market demand, is the sum of supplies of a commodity made by all individual firms or their supply agencies. The market supply of a product is governed by the law of supply.

13.5.2 The Law of Supply

The supply of a commodity depends on its price and cost of its production. In other words, supply is the function of price and production cost.1 The law of supply is, however, expressed generally in terms of price-quantity relationship. The *law of supply* can be stated as follows:

1. Cost of production determines, in fact represents, the minimum price of a commodity.

The supply of a product increases with the increase in its price and decreases with decrease in its price, other things remaining constant. It implies that the supply of a commodity and its price are positively related. This relationship holds under the assumption that "other things remain the same". "Other things" include technology, price of related goods (substitute and complements), and weather and climatic conditions in case of agricultural products.

13.5.3 The Supply Schedule

A supply schedule is a tabular presentation of

Table 13.2 *Supply Schedule of Shirts*

the law of supply. A supply schedule is a table

Price (in `)

Supply (Shirts in '000)

showing different prices of a commodity and the

100

10

corresponding quantity that suppliers are willing

200

35

to offer for sale. Table 13.2 presents a hypothetical

300

50

supply schedule of shirts, i.e., number of shirts

400

60

supplied per week at different prices.

600

75

800

80

13.5.4 The Supply Curve

A supply curve is a graphical presentation of the supply schedule. The supply curve *SS*' given in Fig. 13.2 has been obtained by plotting the data in Table 13.2. The points *S*, *P, Q, R*, *T* and *S*' show the price-quantity combinations on the supply curve *SS*'. The supply curve, *SS*', depicts *the law of supply.* The upward slope of the supply curve indicates the rise in the supply of shirts with the rise in its price and fall in the supply with fall in prices. For example, at pricè200, only 35 thousand shirts are supplied per week. When price rises tò400, supply increases to 60 thousand shirts.

As shown in Fig. 13.2, *a supply curve has a positive slope*. The positive slope or upward movement of the supply curve is caused by the rise in cost of production and

Fig. 13.2 *Supply Curve of Shirt*

seller's effort to make a larger profit. The rise in cost of production results from the law of diminishing returns. In fact, supply curve is derived from the marginal cost curve.

13.5.5 Shift in the Supply Curve

We have shown above that a change in the price of a commodity causes a change in its quantity supplied along a given supply curve. Although price of a commodity is the most important determinant of its supply, it is not the only determinant. Many other factors influence the supply of a commodity. Given the supply curve of a commodity, when there

is change in its other determinants, the supply curve shifts rightward or leftward depending on the effect of such changes. Let us now explain how other determinants of supply cause shift in the supply curve.

(i) Change in Input Prices. When input prices decrease, the use of inputs increase. As a result, product supply increases and the supply curve *SS* shifts to the right to *SS*", as shown in Fig. 13.3. Similarly, when input prices increase, product supply curve shifts leftward from *SS* to *SS*′.

(ii) Technological Progress. Technological changes that reduce cost of production or increase efficiency in production cause increase in product supply. For instance, introduction

Fig 13.3 *Shift in the Supply Curve*

of high yielding variety of paddy and new techniques of cultivation increased per acre yield of rice in India in the 1970s. Such changes make the supply curve shift to the right.

(iii) Price of Product Substitutes. Given its technology and production capacity, a firm can produce more than one good which require a similar technology. For example, a refrigerator company can also produce ACs; Tatas, famous for truck production can also produce cars; Maruti Udyog can produce trucks, and so on. Fall in the price of one of the product substitutes may lead to the rise in the supply of other due to capacity utilization for profit maximization. This may cause shift in the supply curve.

(iv) Nature and Size of the Industry. The supply of a commodity depends also on whether an industry is monopolized or competitive. Under monopoly, supply is fixed. When a monopolized industry is made competitive, the total supply increases. Besides, if size of an industry increases due to new firms joining the industry, the total supply increases and industry supply curve shifts rightward.

(v) Government Policy. When government imposes restrictions on production, e.g., import quota on inputs, rationing of or quota imposed on input supply, etc., production tends to fall. Such restrictions make supply curve shift leftward.

(vi) Non-economic Factors. Factors like labour strikes and lock-outs, war, drought, flood, communal riots, epidemics, etc. also adversely affect the supply of commodities and make the supply curve shift leftward.

13.6 MARKET EQUILIBRIUM: EQUILIBRIUM OF DEMAND AND SUPPLY

Determination of Price in a Free Market. In sections 13.4 and 13.5, we have explained the laws of demand and supply respectively, and how demand and supply behave in response to the change in price and other determinants. In this section, we explain how demand and supply interact to strike a balance, how market attains *equilibrium*, and how *equilibrium price* is determined in a *free market*. A free market is one in which market forces of demand and supply are free to take their own course and there is no outside control on price, demand and supply.

13.6.1 The Concept of Market Equilibrium

In physical sense, the term equilibrium means the ‘state of rest’. In general sense, it means that forces working in opposite directions are in balance. In the context of market analysis, *equilibrium refers to a state of market in which quantity demanded of a commodity equals the quantity supplied of the commodity.* The equality of demand and supply produces an *equilibrium price.* The equilibrium price is the price at which quantity demanded of a commodity equals its quantity supplied. That is, at equilibrium price, demand and supply are in equilibrium. Equilibrium price is also called *market-clearing price*. Market is cleared in the sense that there is no unsold stock and no unsupplied demand.

13.6.2 Determination of Market Price

Equilibrium price of a commodity in a free market is determined by the market forces of demand and supply. In order to analyze how equilibrium price is determined, we need to integrate the demand and supply curves. For this purpose, let us use our earlier example of shirts. Let us suppose that the weekly market demand and supply schedules for shirts in Delhi are given as shown in Table 13.3.

Table 13.3 *Weekly Demand and Supply Schedules for Shirts*

Price per

Demand

Supply

Market

Effect on

Shirt (`)

(’ 000 shirts)

('000 shirts)			*Position*	*Price*
100	80	10	Shortage	Rise
200	55	28	Shortage	Rise
300	40	40	Equilibrium	Stable
400	28	50	Surplus	Fall
500	20	55	Surplus	Fall
600	15	60	Surplus	Fall

As the table shows, there is only one price of shirts (`300) at which quantity demanded per week equals the quantity supplied at 40 thousand shirts. It means that the shirt market in Delhi is in equilibrium at pricè300. At all other prices, the shirt market is in *disequilibrium*—the state of imbalance between supply and demand. When market is in the state of disequilibrium, either demand exceeds supply or supply exceeds demand. As the table shows, at all prices below `300, demand exceeds supply showing *shortage* of shirts in the market. Likewise, at all prices abovè300, supply exceeds demand showing *excess supply.*

Under the conditions of a free market, disequilibrium itself creates the condition for equilibrium. When there is excess supply, it forces downward adjustments in the price and quantity supplied. When there is excess demand, it forces upward adjustments in the price and quantity demanded. The process of downward and upward adjustments in price and quantity continues till the price reaches `300 and quantities supplied and demanded are in balance at 40 thousand shirts. This process is automatic. Let us now look into the process of price and quantity adjustments called '**market mechanism**'.

13.6.3 Market Mechanism: How the Market Brings about Balance

Market mechanism is a process of interaction between the market forces of demand and supply to determine the equilibrium price. To understand how it works, let the price of

shirts be initially set at `100. At this price, the number of shirts demanded (80,000) exceeds the quantity supplied (10,000) by 70 thousand shirts. The shortage gives sellers an opportunity to raise the price and it prepares buyers to accept and pay a higher price. As a result, price goes up. Increase in price enhances the profit margin. This induces firms to produce and sell more in order to maximize their profits. This trend continues till price rises tò300. As Table 13.3 shows, at pricè300, the buyers are willing to buy 40 thousand shirts. This is exactly the number of shirts that sellers would like to sell at this price. At this price, there is neither shortage nor excess supply of shirts in the market. Therefore, `300 is the equilibrium price. The market is, therefore, in equilibrium.

Similarly, at all prices abovè300, supply exceeds demand showing excess supply of shirts in the market. The excess supply forces the competing sellers to cut down the price in order to clear their unsold stock. Some firms find low price unprofitable and go out of the market and some cut down their production. Therefore, supply of shirts goes down. On the other hand, fall in price invites more customers. This process continues until price of shirts falls tò300. At this price, demand and supply are in balance and **market is in equilibrium**. Therefore, price at `300 per shirt is equilibrium price.

Graphical Illustration of Price Determination. The determination of equilibrium price is illustrated graphically in Fig. 13.4. The demand curve *DD′* and the supply curve *SS'* have been obtained by plotting the demand and supply schedules, (given in Table 13.3) on the price and quantity axes.

As Fig. 13.4 shows, demand and supply curves intersect at point *E* determining the equilibrium price at `300. At this price, the quantity demanded (40 thousand shirts) equals the quantity supplied. Thus, the equilibrium price is `300 and equilibrium quantity is 40 thousand shirts. The equilibrium condition is not fulfilled at any other point on the demand and supply curves. Therefore, if price is set at any point other than `300, there would be either excess supply or shortage of shirts in the market.

Fig. 13.4 *Equilibrium of Demand and Supply: Price Determination*

Let us now suppose that the market is in disequilibrium and see how market forces work to bring about balance in demand for and supply of shirts. Let the price be initially set at `600. At this price, suppliers offer to sell 60 thousand shirts whereas buyers are willing to buy only 15 thousand shirts. The supply, obviously, far exceeds the demand. The excess supply equals 45 thousand shirts. The suppliers would, therefore, lower down the price gradually in order to get rid of the unsold stock and cut down the supply simultaneously. On the other hand, when price falls, the law of demand comes into operation and demand for shirts increases too. In this process, the supply-demand gap is reduced. This process continues until price reaches `300 at point *E*, the point of equilibrium where demand and supply are equal at 40 thousand shirts. At this price, the market is in equilibrium and there is no inherent force at work which can disturb it.

Likewise, if price is initially set at `100, the buyers would be willing to buy 80 thousand shirts whereas suppliers would be willing to supply only 10 thousand shirts. Thus, there would be a shortage of 70 thousand shirts in the market as shown by the distance *JK* in Fig. 13.4. Due to shortage the buyers would be willing to pay a higher price. This will lead to increase in price which will encourage the suppliers to increase their supply. This process of adjustment will continue as long as demand exceeds supply. When price rises tò300, the market reaches its equilibrium.

Price Determination by Demand and Supply Functions

In the previous section, we have illustrated graphically how equilibrium of demand and supply is determined at the point of intersection of the demand and supply curves. If demand and supply functions are known, the equilibrium quantity and equilibrium price can also be determined numerically. Suppose demand function for a commodity *X* is given as follows.

$Qd = 150 – 5\,Px$

and the supply function as

$Qs = 10\,Px$

We know that market equilibrium takes place when quantity supplied equals quantity demanded, i.e., at equilibrium, $Qs = Qd$. By substituting supply and demand functions for Qs and Qd, respectively, we get equilibrium price as follows.

$Qs = Qd$

$10\,Px = 150 – 5\,Px$

$Px = 10$

At equilibrium price $Px = 10$, the quantities supplied and demanded are in equilibrium. Equilibrium demand and supply can be obtained by substituting 10 for demand and supply functions as shown below.

$10\,Px = 150 – 5\,Px$

10(10) = 150 – 5(10) = 100.

Fig. 13.5 *Determination of Equilibrium Price and Quantity*

The algebraic determination of equilibrium price and quantity is illustrated graphically in Fig. 13.5. The demand curve *DD′* has been drawn by using the demand function *Qd* = 150 – 5 *Px* and the supply curve *SS′* by using the supply function *Qs* = 10 *Px*. As the figure shows, demand and supply curves intersect at point *P*. A perpendicular drawn from point *P* to the quantity axis determines the equilibrium quantity at 100 units and a line drawn perpendicular from point *P* to the price axis determines the equilibrium price at `10. At this price, the quantity demanded equals the quantity supplied and hence the market is in equilibrium.

13.7 SHIFT IN DEMAND AND SUPPLY CURVES AND MARKET EQUILIBRIUM

When demand increases or decreases price remaining constant, the demand curve shifts upwards or downwards, respectively, it is called *shift in demand curve*. Similarly, price remaining constant when supply increases or decreases, supply curve shifts rightward or leftward, respectively, it is called *shift in supply curve.* In this section, we show how market equilibrium is effected by the shift in demand and supply curves.

13.7.1 Shift in Demand Curve

Whenever there is a shift in the demand and/or supply curve, there is also a shift in the equilibrium point. The effect of shift in the demand curve on the equilibrium is shown in Fig. 13.6. Suppose that the initial demand curve is given by the curve *DD′* and supply curve by *SS′*. The demand and supply curves intersect at point *P*. The equilibrium price is determined at *PQ* and equilibrium quantity at *OQ*. Let the demand curve now shift

from its position *DD′* to *DD″*, supply curve remaining the same. The demand curve *DD″* intersects the supply curve *SS′* at point *M*. Thus, shift in the demand curve causes a shift in the equilibrium from point *P* to point *M*. At the new market equilibrium, quantity demanded and supplied increases from *OQ* to *ON* and price increases from *PQ* to *MN.* Note that, the supply curve remaining the same, a rightward shift in the demand curve results in a higher equilibrium price and quantity.

Fig. 13.6 *Shift in Demand Curve*

Fig. 13.7 *Shift in Supply Curve and*

and Equilibrium

Equilibrium

13.7.2 Shift in Supply Curve

Figure 13.7 shows the effect of shift in the supply curve on the equilibrium. Suppose that the demand curve is given as *DD′* and the initial supply curve as *SS′*. The curves *DD′* and *SS′* intersect at point *P,* determining equilibrium price at *PQ* and equilibrium quantity at *OQ*. Let the supply curve now shift from its position *SS′* to *SS″*, demand curve remaining unchanged. The new supply curve *SS″* intersects the demand curve at point *M*. Thus, a new equilibrium is struck at point *M* where equilibrium price is *MN* and equilibrium quantity is *ON.* Note that a rightward shift in the supply curve, demand curve remaining the same, causes equilibrium price to fall and output to increase.

13.7.3 Simultaneous Shift in Demand and Supply Curves

We have seen above how a rightward shift in the demand curve causes a *rise* in market price (Fig. 13.6), and how a rightward shift in the supply curve (Fig. 13.7) causes a *fall* in the market price. Let us now look at the effect of simultaneous and parallel shifts in demand and supply curves on the equilibrium price and output. The effect of a simultaneous and parallel rightward shift in demand and supply curves on the equilibrium price and output depends on how big or small is the relative shift in demand and supply curves. The *simultaneous and parallel* shift in demand and supply curves in different measures and its effect on equilibrium price and output are illustrated in parts (a) and (b) of Fig. 13.8.

Part (a) of Fig. 13.8 shows that if the shift in the supply curve is bigger than that in the demand curve, then equilibrium price decreases and output increases. For example, suppose that initial demand and supply curves are given by *DD* 1 and *SS* 1, respectively, intersecting at point *E* 1 and determining equilibrium price at *P* 1 and output at *Q* 1. Let the demand curve shift to *DD* 2 and supply curve from *SS* 1 to *SS* 3 intersecting at point *E* 3.

Note that shift in the supply curve is bigger than that in the demand curve. As a result, equilibrium price falls to P_0 and output increases to Q_3. But, if demand and supply curves shift in equal measure, as shown by DD_2 and SS_2, equilibrium price remains unchanged though output increases to Q_2.

Fig. 13.8 *Parallel Shift in Demand Supply Curves and Its Effect on the Equilibrium Price and Output*

Similarly, part (b) of the figure shows the effect of a bigger shift in the demand curve on the equilibrium price and output. It can be seen in the figure that the shift in demand curve from DD_1 to DD_2 is bigger than the shift in the supply curve from SS_1 to SS_2. In this case, both equilibrium price and output increase.

13.8 MARKET STRUCTURE AND DEGREE OF COMPETITION

In Section 13.6, we have discussed and illustrated determination of price assuming simple market conditions and simple demand and supply functions. In reality, however, market structure is extremely complex and so is the system of price determination. This has lead to formation of different theories of price determination under different kinds of market conditions. In this section, we discuss briefly the ***market structure*** – the market morphology.

Market structure refers to the number of firms in an industry and the degree of competition among the firms. The categorization of market structure is based on the following factors.

- Number of firms—the sellers
- Degree and nature of competition
- Level of product differentiation
- Possibility of entry and exit of the firms

The number of sellers of a product in a market determines the *nature and degree of competition* in the market. The nature and degree of competition make the *structure of the market*. Depending on the number of sellers and the degree of competition, the market structure is broadly classified as given in Table 13.4.

Table 13.4 *Types of Market Structures*

Market structure
No. of firms and
Nature of
Control
Method of
degree of
industry
over price
marketing
production
where
differentiation
prevalent

1. **Perfect**
Large no. of
Financial mar-
None
Market
Competition
firms with
kets and some
exchange
homogenous
farm products
or auction
products

2. **Imperfect Competition:**

(*a*) **Monopol-**

Many firms with

Manufacturing:

Some

Competitive

istic com-

real or perceived

tea, toothpastes,

advertising,

petition

product differen-

TV sets, shoes,

quality rivalry

tiation

refrigerators, etc.

(*b*) **Oligopoly**

Little or no pro-

Aluminium, steel, Some

Competitive,

duct differentia-

cigarettes, cars,

advertising,

tion

passenger cars,

quality

etc.

rivalry

(*c*) **Monopoly**

A single prod-

Public utilities:

Considera-

Promotional

ucer, without

Telephones,

ble but

advertising if

close substitute

Electricity, etc.

usually

supply

regulated

is large

Source: Samuelson, P.A. and W.D. Nordhaus, *Economics*, McGraw-Hill, 15th Edn., 1995, p. 152.

13.9 MARKET STRUCTURE AND PRICING DECISIONS

The market structure determines a firm's power to fix the price of its product a great deal. The degree of competition determines a firm's degree of freedom in determining the price of its product. The degree of freedom implies the extent to which a firm is free or independent of the rival firms in taking its own pricing decisions. Depending on the market structure, the degree of competition varies between zero and one. And, a firm's discretion or the degree of freedom in setting the price for its product varies between one and none in the reverse order of the degree of competition. As a matter of rule, the *higher the degree of competition, the lower the firm's degree of freedom in pricing decision and control over the price of its own product and vice versa.* Let us now see how the degree of competition affects pricing decisions in different kinds of market structures.

Under ***perfect competition,*** a large number of firms compete against each other for selling their product. Therefore, the degree of competition under perfect competition is close to one, i.e., the market is highly competitive. Consequently, firm's discretion in determining the price of its product is close to none. In fact, in perfectly competitive market, price is determined by the market forces of demand and supply and a firm has to accept the price determined by the market forces. If a firm uses its discretion to fix the price of its product above or below its market level, it loses its revenue and profit in either case. For, if it fixes the price of its product above the ruling price, it will not be able to sell its product, and if it cuts the price down below its market level, it will not

be able to cover its average cost. In a perfectly competitive market, therefore, firms have little or no choice in respect to price determination.

As the number of firms decreases, the degree of competition decreases. And, as a result, firm's control over the price and its discretion in pricing decision increases. For example, under ***monopolistic competition,*** where degree of competition is high but less than one, the firms have some discretion in setting the price of their products. Under monopolistic competition, the degree of freedom depends largely on the number of firms and the level of product differentiation. Where product differentiation is real, firm's discretion and control over the price is fairly high and where product differentiation is nominal or only notional, firm's pricing decision is highly constrained by the prices of the rival products.

When the number of firms is few, the market takes the form of an ***oligopoly***. Under ***oligopoly***, the degree of competition is quite low, lower than that under monopolistic competition. The firms, therefore, have a good deal of control over the price of their products and can exercise their discretion in pricing decisions, especially where product differentiation is prominent. However, the fewness of the firms gives them an opportunity to form a cartel or to make some settlement among themselves for fixation of price and non-price competition.

In case of a ***monopoly***, the degree of competition is close to nil. An uncontrolled monopoly firm has full freedom to determine the price of its product. A monopoly, in the true sense of the term, is free to fix any price for its product, of course, under certain constraints, viz., (*i*) the objective of the firm, and (*ii*) demand conditions.

The theory of pricing provides different models of how price is determined in different kinds of market structures. The characteristics of different kinds of market structures and price determination in each type of market has been discussed in a theoretical framework in the subsequent chapter since most markets in modern times are either digopolistic or monopolistically competitive. It implies that most firms have some monopoly power. So let us look at how monopoly power is measured.

13.10 MEASURES OF MONOPOLY POWER

While describing the market structure reference has made to monopoly, monopolistic competition and oligopoly. The firms under these kinds of markets have some monopoly powers in price determination. Therefore, before we proceed to discuss the price determination under different kinds of markets, let us have an idea of measures of monopoly power. Like perfect competition, pure private monopolies are rare phenomena. The real business world is, in fact, characterized largely by monopolistic competition and oligopoly. In these kinds of markets, firms hold some monopoly power in the industry which they exercise in determining their price and output. Some economists have suggested methods of measuring monopoly power of a firm in different kinds of markets in their own ways. In this section, we discuss the various measures of monopoly power of monopolistic and oligopoly firms, suggested by the economists.

It may be noted at the outset that measuring monopoly power has been a difficult proposition. The efforts to devise a measure of monopoly power have not yielded a universal or non-controversial measure. As Alex 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".2 If not for any other reason, for 'sheer intellectual curiosity' economic theorists realised the necessity for working on this problem, for they could not with good conscience go on talking about 'great' or 'little' monopoly power or about various degrees of monopoly power without trying to ascertain the meaning of these words.3 Therefore, devising at least a 'conceivable' measure of monopoly, even if 'practical' measurement is impossible, continues to interest economists, for at least two reasons.

First, apart from intellectual curiosity, people would like to know about the economy in which they live, about the industrial structure, and about the industries from which they get their supplies and how their prices are determined.

Second, growth of private monopolies has often led to economic inefficiency and exploitation of consumers. Therefore, the governments of many countries have found it necessary to formulate policies and to 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 power and monopolistic trade practices.

The Methods of Measuring Monopoly Power

In spite of problems in measuring the power of monopoly, economists have devised a number of measures of monopoly power though none of these measures is free from flaws. Yet the various measures do provide an insight into 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 and to prevent their growth. We discuss here briefly the various measures of monopoly power suggested by the economists.

1. Number-of-Firms Criterion

One of the simplest measures of degree of monopoly power of firms 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, the 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 *absolute monopoly power*. On the contrary, in an industry characterized by perfect competition, the number of firms is so large that each firm supplies an insignificant proportion of the market and no firm has any control on the price, and, hence, no monopoly power whatsoever.

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 'number-of-firms' criterion of measuring monopoly power is of little practical use.

2. Concentration Ratio

The *concentration ratio* is one of the widely used criteria for measuring monopoly power. The concentration ratio is obtained by calculating the percentage share of a group of large firms in the total output of the industry. According to Hunter, 'The number of firms chosen for calculating the ratio usually depends on some fortuitous element—normally the census of production

2. Hunter, Alex, "Measurement of Monopoly Power" in ' *Monopoly and Competition*', ed. By Alex Hunter, (Penguin Book, 1970), p. 92.

3. Maclup, Fritz, *The Political Economy of Monopoly*, (The Johns Hopkins Press, Baltimore, 1952), p. 470.

arrangement of the country concerned'.4 In Britain, for example, the share of the three largest firms of a census industry and in the USA, the share of the four largest firms is the basis of calculating concentration ratio.5 However, the number of firms chosen may be as large as 20 depending on the market size and purpose of enquiry. Apart from the share of the largest firms in the industry output, "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 other characteristics."6

Shortcomings. Concentration ratio, although a very widely used measure of monopoly power, has its own ***shortcomings***.

First, the measures of concentration ratio involve statistical and conceptual problems. For example, production capacity may not be used straightaway as it may include 'unused, obsolete or excess capacity' and the value of assets involves valuation problem as accounting method of valuation and market valuation of assets may differ. Employment figure may not be relevant in case of capital-intensive industries and their use 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.7

Second, the measures of concentration ratio do not take into account the size of the market. The 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. For, it is quite likely that the national market is divided among a thousand sellers, each seller being a monopolist in his own area.

Third, 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 that may be different under different classification of industries. Therefore, an industry that has concentration ratio under one 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.

3. Excess Profit Criterion

J.S. Bain and, following him, many other economists have used *excess profit*, i.e., profit in excess of the opportunity cost, as a measure of monopoly power. If profit rate of a firm continues to remain significantly 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 excess profit, the opportunity cost of owner's capital and a 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 as the ratio of the divergence between the opportunity costs (O) and the actual profit (R), to the latter. Thus degree of monopoly power may be expressed as

$$\text{Monopoly Power} = \frac{R - O}{R}$$

If (R – O)/ R = 0, there exists no monopoly, and if it is greater then zero, there is monopoly. The higher the value of (R – O)/ R, the greater the degree of monopoly.

4. Hunter, A., *op. cit.,* p. 101.

5. Hunter, A. *Ibid*.

6. Maclup, Fritz, *op. cit.* , p. 477.

7. Hunter, A., *op. cit.* , p. 102.

Another measure of degree of monopoly based on *excess profit* has been devised by A.P. Lerner.8 According to him, the degree of monopoly power may be measured by the following formula.

$$\text{Monopoly Power} = \frac{P - MC}{P}$$

where P = price, MC = marginal cost.

Since for a profit maximizing firm, $MR = MC$, Lerner's measure of monopoly power may be expressed also as

$$\text{Monopoly Power} = \frac{P - MR}{P}$$

We have discussed earlier (see Chapter 8, Section 8.4.5) that

$P/(P - MR) = e$

and that $(P - MR)/P = 1/e$ (where $P = AR$)

Thus, Lerner's measure of monopoly power may be expressed also in terms of $1/e$.

It may thus be inferred that lower the elasticity, the greater the degree of monopoly, and *vice versa*. Therefore, monopoly power may exist even if the firm's $AR = AC$ and it earns only normal profit.

Drawbacks 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, any formula devised to measure the degree of monopoly power should bring out the difference between the monopoly output and competitive output or the 'ideal' output under optimum allocation of resources. The divergence between P and MC used in Lerner's formula does not indicate the divergence between the actual monopoly output and 'ideal' output. Lerner has possibly used the divergence between P and MC as the substitute for the divergence between actual monopoly output and 'ideal' output. "This substitution of a price-cost discrepancy for a difference between actual and 'ideal' output is probably the greatest weakness of a formula which is supposed to measure deviation from the optimum allocation of resources."9

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, Lerner's formula is of little practical use for measuring monopoly power.

4. Triffin's Cross-Elasticity Criterion

According to Robert Triffin, cross-elasticity of demand for the product of a monopoly firm can be used as a measure of its monopoly power. Triffin's criterion seems to have been derived from the definition of monopoly itself. According to his 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. But, this criterion indicates only the relative power of each firm. It does not provide a single index of monopoly power.

8. "The Concept of Monopoly and the Measure of Monopoly Power," *Review of Economic Studies*, June 1934.

9. Chamberlin, E.H., *The Theory of Monopolistic Competition*, Harvard University Press, Cambridge, Mass., 1933.

SUMMARY

• In economic sense, the term 'market' refers to a mechanism by which buyers and sellers interact to settle the price and quantity bought and sold.

• From analytical point of view, markets have two sides – demand side and supply side. Both the sides of the market operate under certain laws. Demand side of the market works on the law of demand and supply side on the law of supply.

• Market mechanism works on the laws of demand and supply. Market price is determined by the market forces where quantity demanded is equal to quantity sold. This is called market price or the equilibrium price.

• Another feature of the market system is market structure. Market structure refers to the number of firms in an industry and the degree of competition among the firms. In fact, the number of firms determines the degree of competition.

• On the basis of number of firms and degree of competition, market structure is classified under two categories: (*i*) perfect competition or perfectly competitive market, and (*iii*) imperfect competition or imperfectly competitive market.

• Under perfect competition, there is large number of firms selling homogenous products and price is determined by the market forces of demand and supply which has to be accepted by all the firms.

• Imperfect market structure is classified under three categories: (*i*) monopolistic competition, (*ii*) oligopoly, and (*iii*) monopoly.

• Under monopolistic competition, there are many firms selling differentiated or homogeneous products. Firms have the power to determine the price of their product taking in view the market conditions, i.e., the action and reaction of competing firms.

• Under oligopoly, there are a few firms producing and selling differentiated goods. There is a high degree of competition. Price determination under oligopolistic market is highly competitive. There are different methods of price determination under oligopoly. To avoid the disadvantages of competition, oligopoly firms may form cartels to determine price.

• In case of monopoly, there is a single firm producing and selling a product which does not have a close substitute. The monopoly firm has the absolute power to determine the price of its own product. In reality, however, monopolies exist only under the permission of the government under certain conditions regarding price and production with purpose providing gains to the society as a whole.

REVIEW QUESTIONS

1. 'Demand' is a word of common usage. What is the meaning of demand in economics? How is demand different from desire want and need?

2. Define market demand. What are the determinants of market demand in the short run and in the long run? How do increase in consumers' income and price of the substitute goods affect the demand for a commodity?

3. What are the factors that are held constant while deriving an individual demand curve? What will happen when such factors are not held constant?

4. Explain demand schedule, demand curve and demand function. Derive a demand curve from the function $Q = 50 - 10P$.

5. Explain the law of supply through a supply schedule and a supply curve. Why does a supply curve slope upward to the right? What factors cause a rightward shift in the supply curve?

6. Explain why market equilibrium is determined at the intersection of the demand and supply curves. How is market equilibrium affected when consumers' income changes, all other factors remaining the same?

7. Find out equilibrium quantity from demand function $Qd = 25 – 10\ P$ and supply function $Q = 25\ P$. What is the change in price if demand function changes to $Qd = 30 – 10P$, supply function remaining the same?

8. Suppose annual sales data of a book publishing company produces a demand function as $Q = 5000 – 50\ P$. From this demand function, find out:

(*a*) demand schedule and demand curve,

(*b*) number of books sold at pricè25,

(*c*) price for selling 2500 copies,

(*d*) price for zero sales,

(*e*) sales at zero price.

9. What is meant by equilibrium price and quantity? What factors cause an upward right and upward left shift of the equilibrium point from its original position?

10. From a demand function $Qd = 2000 – 30\ P$ and a supply function $Qs = 20\ P$, find out

(*a*) equilibrium price,

(*b*) equilibrium quantity, and

(*c*) gap between demand and supply at P = `20 and $P$ = `50.

11. Which of the following statements are True or False?

(*a*) The demand for a commodity is inversely related to the price of its substitutes.

(*b*) When income increases, the demand for essential goods increases more than proportionately.

(*c*) Decrease in input prices causes a leftward shift in the supply curve.

(*d*) There cannot be a market without a place.

(*e*) The desire for a commodity backed by ability and willingness to pay is demand.

(*f*) The law of demand states the relationship between the quantity demanded and price of a commodity, consumers’ income, price of the related goods and advertisement.

(*g*) An individual’s demand curve marks the upper limits of his/her intentions to buy a commodity at different prices.

(*h*) A market demand curve represents the maximum quantity that an individual would be willing to buy at different prices.

(*i*) Demand for car and price of petrol are inversely related.

(*j*) Most demand functions are of the form $D = a + bP$.

[(**Ans.** True: (*e*), (*g*), (*h*), False: (*a*), (*b*), (*c*), (*d*), (*f*), (*i*), (*j*)]

12. Which of the following conditions makes an approximate definition of ‘market’?

(*a*) Market is a meeting place for buyers and sellers.

(*b*) The buyers and sellers meet to transact business.

(*c*) The buyers and sellers must transact business by or without meeting in a place.

13. What is meant by market structure? What factors determine the market structure?

14. What are the different types of market structure. What are main features of different kinds of markets?

15. Distinguish between (*i*) Perfect competition and monopolistic competition, and (*ii*) monopolistic competition and oligopoly.

16. What is meant by monopoly power? What are the measures of monopoly power? What are their limitations?

CHAPTER

14 Price and Output Determination under Perfect Competition

CHAPTER OBJECTIVES

The objective of this Chapter is to discuss price determination under perfect competition. The content of this Chapter includes:

• Definition of perfect competition

• Characteristics of perfect competition

• Price determination in short run

• Price determination in long run

• Relevance of perfect competition analysis

14.1 INTRODUCTION

In this Chapter, we begin to discuss the theory of price determination in different kinds of markets. As noted in Chapter 13, market structure is classified under (*i*) perfect competition, (*ii*) monopoly, (*iii*) monopolistic competition, and (*iv*) oligopoly. The different kinds of markets and their basic characteristics have already been discussed in the preceding

Chapter. In this chapter, we explain how price is determined under perfect competition and how firms determined their profit maximizing output.

14.2 DEFINITION AND CHARACTERISTICS OF PERFECT COMPETITION

Perfect competition *may be defined as a kind of market in which there are a large number of buyers and sellers each one buying and selling a homogeneous product and no buyer or seller has control over the price of the product.* The number of buyers and sellers is so large that the quantity bought and sold by each buyer and seller is insignificant and no buyer or seller has the power to alter the market price. In a perfectly competitive market, there is complete absence of rivalry among the firms. In fact, in a perfectly competitive market, competition among the individual firms is so widely dispersed that it amounts to no competition.

It may be added at the outset that the concept of perfectly competitive market is an abstract concept. Perfect competition as conceived by the economists does not exist in reality, except in sporadic markets like auctions. However, analysis of price and output determination under perfect competition 'lays the foundation of pricing theory'. That is

why it has become a common practice to begin the analysis of price determination with price determination under perfect completion.

14.3 CHARACTERISTICS OF PERFECT COMPETITION

A perfectly competitive market is supposed to have the following characteristics.

1. Large number of buyers and sellers. Under perfect competition, the number of buyers and sellers is assumed to be very large. The number of buyers and sellers is so large that the share of each buyer in total demand and the share of each seller in total supply is so insignificant that neither a buyer nor a seller can influence the market price. Therefore, firms are deemed to be *price-takers,* not price makers.

2. ***Homogeneous products.*** Products supplied by all firms are almost homogeneous. Homogeneity of products means that products supplied by various firms are so identical in appearance and use that buyers do not distinguish between them nor do they prefer the product of one firm to that of another. Product of each firm is regarded as a perfect substitute for the product of other firms. Hence, no firm can gain any competitive advantage over the other firms. Nor do the firms distinguish between the buyers. For example, wheat and vegetables produced by all the farmers, other things given, are treated as homogeneous.

3. ***Free entry and free exit of firms.*** There is no legal or market-related barrier on the entry of new firms into or exit of existing ones from the industry. Firms are free to enter the industry and quit it at their free will.

4. ***Perfect mobility of factors of production.*** For a market to be perfectly competitive, there should be perfect mobility of resources. This means that the factors of production can move freely into or out of an industry and from one firm to another. This is however a purely theoretical assumption.

5. ***Perfect knowledge.*** There is perfect dissemination of the information about the market conditions. Both buyers and sellers are fully aware of the nature of the product, its availability or saleability and of the price prevailing in the market. In simple words, *perfect knowledge* means that each buyer knows what can be bought at what price and where. Similarly, each seller knows what can be sold at what price and where.

6. ***Absence of collusion or artificial restraint.*** There is no sellers' union or other kinds of collusions between the sellers such as cartels or guilds, nor is there any kind of collusion between the buyers, e.g. , consumers' associations or consumer forum. Each seller and buyer acts independently. The firms enjoy the freedom of independent decisions.

7. ***No government intervention.*** In a perfectly competitive market, there is no government intervention with the working of the market system. There is no licencing system regulating the entry of firms to the industry, no regulation of market prices, i.e., fixation of lower or upper limits of prices, no control over the supply of inputs, no fixation of quota on production, and no rationing of consumer demand, no subsidy to producers or to consumers, etc.

Perfect competition, as characterized above, is ***an uncommon phenomenon*** in the real business world. However, the actual markets that approximate to the conditions of perfectly competitive model include the share markets, securities and bond markets, and agricultural product markets, e.g., local vegetable markets. Although perfectly competitive markets are uncommon phenomena, perfect competition model has been the most popular model used in economic theories due to its analytical value as it provides a starting point and analytical framework for pricing theory.

Perfect Competition and Pure Competition

Sometimes a distinction is made between *perfect competition* and *pure competition*. The difference between the two is only a matter of degree. While *perfect competition* has all the features mentioned above, *pure* competition does not have *perfect factor* mobility and perfect knowledge. In simple words, *perfect competition less perfect mobility of factors and perfect knowledge is regarded as pure competition*. In this book, however, we will use the two terms interchangeably.

14.4 PRICE DETERMINATION UNDER PERFECT COMPETITION

In a perfectly competitive market, price of a product is determined by the market forces – market demand and market supply – not by the individual firms. Price determined by the market forces is called ***market price.*** Once market price is determined, the only option open to individual firms is to determine their equilibrium output—the output that maximizes their profit. How market price is determined in a perfectly competitive market has already been explained in detail in Chapter 13 (Sec. 13.6). The system of market price determination is recalled here briefly with the purpose of showing the derivation of demand curve—demand line, in fact—faced by the individual firms under perfect competition.

The price determination under perfect competition is illustrated in Fig. 14.1. As shown in Panel (*a*) of Fig. 14.1, market demand is shown by the demand curve, *DD′* and market supply curve by *SS′* . Demand and supply curves intersect at point *E* determining the market equilibrium. At the equilibrium point *E,* market price is determined at *OP* = *EQ* and market size at total demand and supply at *OQ.*

Fig. 14.1 *Determination of Market Price and Demand for Individual Firms*

The analysis of market equilibrium provides the basis for ***deriving demand curve for individual firms.*** Once market price is determined, the price is determined for all individual firms. Individual firms have no power to change the price. The firms have the only option to sell their product at the given price. At market price *OP,* the firms can sell any quantity of their product. This condition implies that the demand curve for an individual firm is given by a ***straight line*** as shown by line *PD* in Panel (*b*) of

Fig. 14.1. The line *PD* represents firm's ***demand line,*** *P* = *MR* line, i.e., price *(P)* is equal to marginal revenue *(MR).* The same demand line will be used to show how firms determine their equilibrium output, i.e., the profit maximizing output.

14.5 OUTPUT DETERMINATION IN SHORT RUN: SHORT-RUN EQUILIBRIUM OF THE FIRM

In the preceding section, we discussed the *determination of price* under perfect competition. Once market price is determined, the firms are left with option to find the output that can maximize their profit. In this section, we proceed to discuss how firms determine their profit-maximizing output, i.e., how they find their equilibrium given the market and cost conditions. In the long-run, however, market and cost conditions tend to change. Therefore, firm's decision on output determinations is analyzed under (*i*) short-run conditions, and (*ii*) long-run conditions. In this section, firm's short-run equilibrium is discussed under short-run conditions. Firm's equilibrium under long-run conditions will be discussed in the next section.

Short-run Equilibrium. A short-run is, by definition, a period in which firm's cost curves are given; price of the product is given; and the number of firms is also given. The determination of the short-run equilibrium of the firm is illustrated in Fig. 14.2. Market price determination in short-run is shown in Panel (*a*) and firm's equilibrium in Panel (*b*). Given the original demand and supply curves as *DD* and *SS,* respectively, market price is determined at the point of intersection of demand and supply curves at point *E*. At market equilibrium point *E*, market price is determined at *OP* 1 = *EQ*.

Fig. 14.2 *Pricing under Perfect Competition in the Short-run*

Given the price *EQ* (= *OP* 1), an individual firm can produce and sell any quantity at this price. But any quantity will not yield maximum profit. Given their cost curves, the firms are required to adjust their output to the price *EQ* so that they maximize their profit.

The process of firm's output determination and its equilibrium are shown in Fig. 14.2(b). As noted earlier (See Chapter 2, Section 2.5.1), profit is maximum at the level of output where *MR* = *MC*. Since price is fixed at *EQ*, firm's *AR* = *EQ*. If *AR* is constant, *MR* = *AR*. The firm's *MR* is shown by *AR* = *MR* line. Firm's upward sloping *MC* curve intersects

AR = *MR* at point *E*. At point *E*, *MR* = *MC*. Point *E* is, therefore, the firm's equilibrium point. An ordinate drawn from point *E* to the output axis, as shown by the line *EM*, determines the profit-maxmizing output at *OM*. At this output the firm's *MR* = *MC*. This satisfies the necessary condition for profit maximization. The total maximum profit has been shown by the area P_1TNE in Fig. 14.2(b).

The total profit is calculated as

Profit = (*AR* – *AC*) *Q*.

In Fig. 14.2(*b*), *AR* = *EM*; *AC* = *NM*; and *Q* = *OM*.

Substituting these values into the profit equation, we get Profit = (*EM* – *NM*) *OM*.

Since *EM* – *NM* = *EN*, Profit = *EN* × *OM* = P_1TNE. This is the maximum supernormal profit, given the price and cost curves, in the short run.

Firms may make losses in the short-run While firms may make supernormal profit, there may be conditions under which firms make losses in the short-run. For instance, suppose demand curve *DD* shifts downward for some reason to *DD*′, as shown in Fig. 14.2(a). As a result, market price decreases from *EQ* to $E'Q'$. This will force a process of output adjustments till firms reach a new equilibrium at point E'. At equilibrium point E', firm's $AR' = MR' = MC$. But, as Fig. 14.2(b) shows, *AR* < *AC*. Therefore, the firms incur a loss. But, since in the short-run, it may not be desirable to close down the production, the firms try to minimize their loss, by adjusting their output downward to OM' where it covers only its *MC*, i.e. , $E'M'$. The firms survive in the short-run so long as they cover their *MC,* especially the variable cost of labour.

It may thus be concluded that, in a perfectly competitive market, a firm may make supernormal profit, normal profit or losses as well, in the short run depending on market conditions.

14.6 DETERMINATION OF OUTPUT IN LONG RUN

In the preceding section, we have discussed the short-run equilibrium of the firm, i.e., output determination by an individual firm in short run under perfect competition. We proceed now to discuss firm's equilibrium in long run. By definition, ***long run*** is a period in which (*i*) firm's cost and revenue curves are subject to change, (*ii*) firms can increase their scale of production by increasing their capital and output, (*iii*) existing firms can exit and new firms can enter the industry, and (*iv*) market conditions can change with change in market demand and supply conditions. Besides, production technology may change affecting the demand for labour and capital. In this section, we will analyse the determination of long-run equilibrium of the firm under perfect competition.

Equilibrium of the Firm

As noted above, market conditions change in the long run. The determination of long-run equilibrium of the firm is illustrated graphically in Fig. 14.3. To analyse the long-run equilibrium of the firm, let us begin by looking at short-run equilibrium of the firm. Suppose (*i*) short-run market price is given at OP_1 and firm's demand line as *AR* = *MR,* and (*ii*) firm's short-run average and marginal cost curves are given as SAC_1 and SMC_1, respectively as shown in Panel (*a*) of Fig. 14.3. Given the firm's cost and revenue curves, firm's short-run equilibrium is determined at point E_1 in panel (*b*) of Fig. 14.3 and firm's

short-run equilibrium output is OQ_1. At this output, the firm makes **abnormal profit** equal to $E_1M = E_1Q_1 - MQ_1$, which is the same as *AR* – *AC* per unit. Abnormal profit brings about ***three major changes*** in market conditions.

First*,* abnormal profit motivates firms to increase the scale of their production. As a result, they gain from economies of scale and their cost curves shift downward as shown by SAC_2 and SMC_2. Consequently, firm's long-run cost curves, *LAC* and *LMC,* tend to decline as shown in Fig. 14.3(b).

Second*,* attracted by abnormal profit new firms enter the industry. Consequently, market supply of the product increases as shown by the shift in supply curve from SS_1 to SS_2 in Fig. 14.3(a).

Third*,* as a result of the first two changes, the price of the product decreases from

E_1N_1 to E_2N_2 or from OP_1 to OP_2 , as shown in Fig. 14.3(a).

Fig. 14.3 *Long-run Equilibrium of the Firm and Industry*

Given the new market price, OP_2, firms attain their equilibrium in the long run at point E_2 where $AR = MR = LMC = LAC = SMC = SAC$ as shown in Fig. 14.3(b). As the figure shows, the firms of industry reach their equilibrium in the long run where both short- and long-run equilibrium conditions are satisfied simultaneously. In a perfectly competitive market, the cost and revenue conditions are given for the firms by the market conditions. Therefore, when price goes down to OP_2, what firms are required to do is to adjust their output to the given revenue and cost conditions in order to maximize their profit. Through this process of adjustment for output, the firms reach the equilibrium in the long run at point E_2 in Panel (*b*). Point E_2 is the point of equilibrium for all the firms in the long run.

In case market price falls below OP_2, say, to OP_3, all the firms make losses. This brings in a reverse process of adjustment. While some firms quit the industry, some firms cut down the size of the firm. As a result, total supply decreases, demand remaining the same. Consequently, price tends to rise. This process of output adjustment continues until industry reaches back to its equilibrium at point E_2, where *LAC* is tangent

to $P = AR = MR$ for each firm in the industry. At point E_2, 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 restore the equilibrium.

SUMMARY

• Perfect competition is a market structure in which there are large number of buyers and sellers of a homogeneous products and no buyer or seller has any control over the market price of the product.

• The features of perfectly competitive market are (*i*) there are large number of buyers and sellers; (*ii*) product is homogenous, (*iii*) price is determined by market demand and supply, not buy individual sellers, (*iv*) there is free entry and exit of firms, (*v*) factors of production are freely mobile between firms, (*vi*) both buyers and sellers have perfect knowledge about the market conditions, (*vii*) firms have uniform cost curves, and (*viii*) there is no government control over the market.

• Under perfect competition, firms have no option to determine the price of their product. Price is determined by the market forces. Once price is determined, firms have to accept it and they have the only option to determine their profit maximizing output.

• In short run, firms may make abnormal profit, normal profit or losses, depending on price and cost conditions.

• In the long run, market conditions tend to change due to (*i*) existing firms increasing their scale of production, (*ii*) decrease in cost of production due to economies of scale resulting from increase in scale of production, (*iii*) new firms may enter the industry leading to increase in market supply and consequently decrease in the market price.

• Under the long-term market conditions, as noted above, all firms make only normal profit in long run.

REVIEW QUESTIONS

1. What are the characteristics of perfect competition? Distinguish between perfect and pure competition.
2. What is the relative position of a firm in a perfectly competitive industry? How does it choose its price and output?
3. Under what market conditions a firm is a price-taker? What would happen to a firm if it becomes price-maker?
4. Analyse the equilibrium of a firm under the conditions of perfect competition in the short run?
5. Discuss the importance of *AR, AC, MR* and *MC* in determining firm's equilibrium under perfect competition.
6. Explain the short run equilibrium of a competitive firm. Under what conditions would a competitive firm close down its business in the short run?
7. Do you agree that perfect competition leads to optimumization of the size of the firm? Illustrate graphically.

8. Under perfect competition average revenue equals average cost in the long run. Why do firms produce under such a condition?

9. Show how under the condition of perfect competition in the long run, the price of a commodity equal to its average and marginal cost.

10. Distinguish between short-run and long-run equilibrium of a firm operating under perfect competition. What is difference, if any, in conditions of equilibrium in the two cases?

11. Bring out the essential difference in the nature of equilibrium of a firm under perfect competition in the short run and in the long run.

12. How is short-run supply curve of a firm derived under perfect competition? Explain and illustrate.

13. What is basis of the short-run supply curve of a competitive firm? What is its likely shape?

14. Write a short note on the relationship between firm's short-run cost curves and supply curve.

15. Illustrate the derivation of the long-run supply curve of an industry under perfect competition and decreasing cost.

16. If all the firms in a perfectly competitive industry have U-shaped cost curves, can then supply curve of the industry be downward sloping?

17. Suppose that a competitive firm is in long-run equilibrium. What will happen to price in the long run if there is a rise in demand for the product of the industry?

18. Which of the following statements are correct?

(*a*) Perfect competition less perfect knowledge and perfect mobility is pure competition.

(*b*) Under perfect competition, a firm fixes its price where its $AR = MR$.

(*c*) A firm is a price-taker under perfect competition.

(*d*) In a perfectly competitive industry, a firm is in equilibrium in the short run only when its $AC = AR = MR = MC$.

(*e*) The short run supply curve has a negative slope.

(*f*) A firm reaches its shut-down point when price goes below its AC.

(*g*) Industry supply curve is a horizontal summation of its firm's supply curves.

(*h*) An industry is in equilibrium in the short run when market is cleared.

(*i*) Change in the industry equilibrium changes firm's equilibrium.

(*j*) Industry supply curve has a positive slope under decreasing cost conditions.

(*k*) In the long run, a firm is in equilibrium when its $AR = MR = LAC = LMC$.

[**Ans.** (*a*), (*c*), (*g*), (*h*), (*i*) and (*k*)]

19. Which of the following features are absent in pure competition?

(*a*) Large number of buyers and sellers

(*b*) Free entry and free exit

(*c*) Perfect knowledge.

(*d*) Perfect mobility

(*e*) Absence of collusion.

20. For a firm, the 'shut-down' point falls

(*a*) anywhere below SAC

(*b*) where $SMC = SAVC = P$

(*c*) where $SMC = SAV$

(*d*) where $SAC = SAVC$?

21. Which of the following is relevant for a perfectly competitive industry?

(*a*) Industry equilibrium is affected by the change in firm's equilibrium.

(*b*) Change in industry's equilibrium affects firm's equilibrium.

(*c*) Change in industry's equilibrium does not affect firm's equilibrium.

(*d*) Change in firm's equilibrium does not affect industry's equilibrium.

22. Under perfect competition, firms are in equilibrium in the long run, when

(*a*) $P = SMC = SAC$

(*b*) $SMC = SAC = AR = MR$,

(*c*) $LAC = LMC = AR = MR$

(*d*) $AR = MR$ but $LMC > LAC$?

Write the correct statement.

[**Ans.** 19 (*c*) and (*d*), 20 (*b*), 21 (*b*), 22 (*c*)]

FURTHER READING

Browning, E.K. and Browing, J.M., *Macroeconomic Theory and Applications* (New Delhi Hall, 1998):

Kalyani Publishers), 2nd Edn., Chapters 8 and 9.
Clark. J.M., 'Towards a Concept of Workable Competition'. *American Economic Review*, 1940, 30 (2): 241–256.
gould, J.P. and Lazear, E.P., *Microeconomic Theory* (Homewood, II: Richard D. Irwin, 1993), 6th Edn., Chapter 9.
koutsoyiannis, A., *Modern Microeconomics* (London: Macmillan, 1978), 2nd Edn., Chapter 5.
Maddala, G.S. and Miller, E., *Microeconomics: Theory and Applications* (New York, NY: McGraw-Hill Book Com., 1989), Chapter 10.
Marshall, A., *Principles of Economics*, Book VI (London: Macmillan, 1920).
PindyCk, R.S. and Rubinfeld, D.L., *Microeconomics* (New York, NY: Prentice Hall, 2007), 5th Edn., Chapters 8 and 9.

CHAPTER

15 Price and Output Determination under Monopoly

CHAPTER OBJECTIVES

The objective of this Chapter is to analyse price and output determination under monopoly, another kind of rare market. The content of this chapter includes:

- Meaning and features of monopoly
- Factors behind the emergence of monopoly
- Price and output determination in short run
- Price and output determination in long run
- Price discrimination by monopoly firms
- Equilibrium of multi-plant monopoly
- Measures of monopoly power

15.1 INTRODUCTION

In the preceding chapter, we have discussed price and output determination under perfect competition – a rarest of rare kind of markets. In this chapter, we discuss the theory of price and output determination in another kind of rare market – ***monopoly.*** In modern times, most monopolies are established or governed by the government. Price in government monopolies and government governed monopolies are determined under social considerations, not under market conditions. In this chapter, however, we discuss price and output determination under the conditions of private monopolies. Let us first look at definition of monopoly and factors that lead to emergence of private monopolies.

15.2 DEFINITION AND SOURCES OF MONOPOLY

15.2.1 Definition of Monopoly

The term *pure monopoly* means an absolute power of a firm to produce and sell a product that has no close substitute. In other words, a monopolized market is one in which there is only one seller of a product having no close substitute. The cross elasticity of demand for a monopoly product is either zero or negative. *A monopolized industry is a single-firm industry*. Firm and industry are identical in a monopoly setting. In a monopolized industry, equilibrium of the monopoly firm signifies the equilibrium of the industry.

However, the precise definition of monopoly has been a matter of opinion and purpose. For instance, in the opinion of Joel Deal,1 a noted authority on managerial economics, a monopoly market is one in which 'a product of lasting distinctiveness is sold. The monopolized 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/users to distinguish between them and to accept or reject a commodity as a substitute. Another concept of pure monopoly has been offered by E.H. Chamberlin2 who envisages monopoly as the control of all goods and services by the monopolist. But such a monopoly has hardly ever existed, hence his definition is questionable. In the opinion of some authors, any firm facing a sloping demand curve is a monopolist. This definition, however, includes all kinds of firms except those under perfect competition.3 For our purpose here, we use the general definition of pure monopoly, i.e., a firm that produces and sells a commodity which has no close substitute.

15.2.2 Causes and Kinds of Monopolies

The emergence and survival of a monopoly firm is attributed to the factors which prevent the entry of other firms into the industry and eliminate the existing competitors. The barriers to entry are, therefore, the major sources of monopoly power. The main barriers to entry are:

(*i*) legal restrictions or barriers to entry of new firms

(*ii*) sole control over the supply of scarce and key raw materials

(*iii*) efficiency in production, and

(*iv*) economies of scale.

(i) Legal Restrictions. Some monopolies are created by law in the public interest. Most of the erstwhile monopolies in the public utility sector in India, e.g., postal, telegraph and telephone services, telecommunication services, generation and distribution of electricity, Indian Railways, Indian Airlines and State Roadways, etc., were public monopolies. Entry to these industries was prevented by law. Now most of these industries are being gradually opened too the private sector. Also, the state may create monopolies in the private sector also, through licence or patent, provided they show the potential of and opportunity for reducing cost of production to the minimum by enlarging scale of production and investing in technological innovations. Such monopolies are known as *franchise monopolies*.

(ii) Control over Key Raw Materials. Some firms acquire monopoly power because they have traditional control over certain scarce and key raw materials which are essential for the production of certain goods, e.g., bauxite, graphite, diamond, etc. For instance, Aluminium Company of America had monopolized the aluminium industry before World War II because it had acquired control over almost all sources of bauxite supply[4]. Such monopolies are often called ‘raw material monopolies’. The monopolies of this kind emerge also because of monopoly over certain specific knowledge of technique of production.

(iii) Efficiency in Production. In an open economy, some firms attain monopoly status because of their superior efficiency in production. Efficiency in production, especially under

1. *Managerial Economics,* (Prentice-Hall, Englewood Cliffs, N.J., 1951).

2. *The Theory of Monopolistic Competition,* (Harvard University Press, Mass, 1933).

3. Watson, D.S., *Price Theory and its Uses,* (Scientific Book Agency, Calcutta, 1967), p. 294.

4. Ferguson, C.E., *Macroeconomic Theory,* (Richard D. Irwin, Illinois, 1972), p. 28.

imperfect market conditions, may be t he result of long experience, innovative ability, financial strength, availability of finance at lower cost, low marketing cost, managerial efficiency, etc. Efficiency in production reduces cost of production below the cost level of competing firms. As a result, a firm gains higher competitive strength and can eliminate rival firms and gain the status of a monopoly. Such firms are able to gain governments’ favour and protection.

(iv) Economies of Scale. The economies of scale are a primary and technical reason for the emergence and existence of monopolies in an unregulated market. If a firm’s long-run minimum cost of production or its most efficient scale of production coincides almost with the size of the market, then the large-size firm finds it profitable in the long-run to eliminate competition through price cutting in the short-run. Once its monopoly is established, it becomes almost impossible for the new firms to enter the industry and survive. Monopolies created on account of this factor are known as ***natural monopolies***. A natural monopoly may emerge out of the technical conditions of efficiency or may be created by law on efficiency grounds.

15.3 MONOPOLY PRICING AND OUTPUT DECISION:

SHORT-RUN ANALYSIS

As under perfect competition, the theory of pricing and output determination under monopoly is based on profit maximization hypothesis, given the revenue and cost conditions. Although cost conditions, i.e., *AC* and *MC* curves, in a competitive and monopoly market are generally identical, revenue conditions, i.e., *AR* and *MR* curves, are different under monopoly. Unlike a competitive firm, a monopoly firm faces a downward sloping demand curve. The reason is a monopolist has the option and power to reduce the price and sell more and to raise the price and sell less. Therefore, given the price-demand relationship, demand curve under monopoly is a typical downward sloping demand curve.

In case of a demand curve is sloping

downward, marginal revenue (*MR*) curve

lies below the *AR* curve and, technically,

Fig. 15.1 *Price Determination under*

Monopoly: Short-run

the slope of the *MR* curve is twice that

of *AR* curve5.

5. *Proof*. Let us suppose a linear demand function is given as $Q = \propto - \beta P$. From this demand function, a price

function is derived as $P = a - bQ$, where b gives slope of the demand curve . Price function is, in fact, the same as *AR* function.

Given the price function, total revenue equation (*TR*) can be worked out as

$TR = Q.\ P = Q\ (a - bQ) = aQ - bQ\ 2$

Since *MR* equals the first derivative of the *TR* equation,

$\partial\ TR$

$MR =$

$= a - 2\ bQ$

$\partial\ Q$

Obviously, the slope of *MR* curve is – 2 *b* whereas the slope of *AR* = – *b*. Thus, the slope of *MR* curve is twice that of *AR* curve.

The short-run revenue and cost conditions faced by a monopoly firm are presented

in Fig. 15.1. Firm's average and marginal revenue curves are shown by the *AR* and *MR*

curves, respectively, and its short-run average and marginal cost curves are shown by *SAC*

and *SMC* curves, respectively. The price and output decision rule for profit maximizing

monopoly is the same as for a firm in the competitive industry.

As noted earlier, profit is maximized at the level of output at which *MC* = *MR*. Given

the profit maximization condition, a profit maximizing monopoly firm chooses a price-

output combination at which *MR* = *SMC*. Given the monopoly firm's cost and revenue

curves in Fig. 15.1, its *MR* and *SMC* intersect at point *N*. An ordinate drawn from point

N to *X*-axis, determines the profit maximizing output at *OQ*. The ordinate *NQ* extended

to the demand curve (*AR* = *D*) gives the profit maximizing price *PQ*. It means that given

the demand curve, the output *OQ* can be sold per time unit at only one price, i.e., *PQ*

(= *OP* 1). Thus, the determination of equilibrium output simultaneously determines the

equilibrium price for the monopoly firm. Once price is fixed, the unit and total profits are

also simultaneously determined. Hence, the monopoly firm is in a state of equilibrium.

At output *OQ* and price *PQ*, the monopoly firm maximizes its unit and total profits.

Its per unit monopoly or economic profit (i.e., *AR* – *SAC*) equals *PQ* – *MQ* = *PM*. Its

total profit, p = *OQ* × *PM*. Since *OQ* = *P* 2 *M*, p = *P* 2 *M* × *PM* = area *P* 1 *PMP* 2 as shown by the shaded rectangle. Since cost and revenue conditions are not expected to change in

the short-run, the equilibrium of the monopoly firm will remain stable.

Determination of Monopoly Price and Output: Algebraic Solution

The determination of price and output by a monopoly firm in the short-run is illustrated

above graphically (see Fig. 15.1). Here, we present an algebraic solution to the problem

of determination of equilibrium price and output under monopoly.

Suppose demand and total cost functions for a monopoly firm are given as follows.

Demand function :

$Q = 100 - 0.2\ P$

…(15.1.1)

Price function :

$P = 500 - 5\ Q$

...(15.1.2)

Cost function

:

$TC = 50 + 20\ Q + Q\ 2$

...(15.2)

Given the demand and cost functions, the problem before the monopoly firm is to find

the profit maximizing output and price. The problem can be solved as follows.

We know that profit is maximum at output at which *MR* = *MC*. So the first step is to

find *MR* and *MC* from the demand function and cost function, respectively. We have noted earlier that *MR* and *MC* are the first derivation of *TR* and *TC* functions, respectively. *TC* function is given, but *TR* function is not. So, let us find *TR* function first. We know that

$TR = P.\ Q$

Since $P = 500 - 5\ Q$, by substitution, we get

$TR = (500 - 5\ Q)\ Q$

$TR = 500\ Q - 5\ Q\ 2$

...(15.3)

Given the *TR* function (15.3), *MR* can be obtained by differentiating the *TR* function.

$\partial\ TR$

$MR =$

$= 500 - 10\ Q$

$\partial\ Q$

Likewise, *MC* can be obtained by differentiating the *TC* function (15.2).

$\partial\ TR$

$MC = \partial\ Q = 20 + 2\ Q$

Now that *MR* and *MC* function are known, profit maximizing output can be easily obtained. Recall that profit is maximum where *MR* = *MC*. As given above,

$MR = 500 - 10\ Q$

and

$MC = 20 + 2\ Q$

By equalizing the *MR* and *MC* functions, we get profit maximizing output (*Q*) as follows.

$MR = MC$

$500 - 10\ Q = 20 + 2\ Q$

$480 = 12\ Q$

$Q = 40$

The output *Q* = 40 is the profit maximizing output.

Now profit maximizing price can be obtained by substituting 40 for *Q* in the price function (15.1.2). Thus, *P* = 500 – 5 (40) = 300. Profit maximizing price is `300.

Given the *TR* and *TC* functions, the total profit (p) can be obtained as follows.

p = $TR - TC$

By substitution, we get p = $500\ Q - 5\ Q\ 2 - (50 + 20\ Q + Q\ 2)$

$= 500\ Q - 5\ Q\ 2 - 50 - 20\ Q - Q\ 2$

By substituting profit maximizing output (40) for *Q*, we get

p = $500(40) - 5(40)(40) - 50 - 20(40) - (40 \times 40)$

$= 20{,}000 - 8{,}000 - 50 - 800 - 1600 = 9{,}550$

Thus, total maximum profit comes tò9,550.

Does a Monopoly Firm Always Earn Abnormal Profit?

There is no certainty that a monopoly firm will always earn an economic or abnormal profit. Whether a monopoly firm earns abnormal profit or normal profit or incurs loss depends on

(*i*) its cost and revenue conditions;

(*ii*) threat from potential competitors; and

(*iii*) government policy in respect of monopoly.

If a monopoly firm operates at the level of output where *MR* = *MC,* its profit depends on the relative levels of *AR* and *AC*. Given the level of output, there are three possibilities.

(*i*) if $AR > AC$, there is abnormal profit for the firm,

(*ii*) if $AR = AC$, the firm earns only normal profit, and

(*iii*) if $AR < AC$, though only a theoretical possibility, the firm makes losses.

15.4 MONOPOLY PRICING AND OUTPUT DECISION:

LONG-RUN ANALYSIS

The decision rules regarding optimal output and pricing in the long-run are the same as in the short-run. In the long-run, however, a monopolist gets an opportunity to expand

the scale of its production with a view to enhancing 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 profit, and (*c*) risk of inviting legal restrictions. Let us assume, for the time being, that none of these conditions limits the expansion of a monopoly firm and discuss the price and output determination in the long-run.

The equilibrium of monopoly firm and its price and output determination in the long-run is shown in Fig. 15.2. The *AR* and *MR* curves show the market demand and marginal revenue conditions faced by the monopoly firm. The *LAC* and *LMC* show the long-run cost conditions. It can be seen in Fig. 15.2, that monopoly's *LMC* and *MR* intersect at point *P* determining profit maximizing output at OQ_2. Given the *AR* curve, the price at which the total output OQ_2 can be sold is P_2Q_2. Thus, in the long-run, equilibrium output will be OQ_2 and price P_2Q_2. This output-price combination maximizes monopolist's long-run profit. The total long-run monopoly profit is shown by the rectangle $LMSP_2$.

Comparison of Short-run and Long-run Equlibrium

It can be seen in Fig 15.2 that compared to short-run equilibrium, the monopolist produces a larger output and charges a lower price and makes a larger monopoly profit in the long-run. In the short-run, monopoly's equlibirium is determined at point *A*, the point at which SMC_1 intersects the *MR* curve. Thus, monopoly's short-run equilibrium output is OQ_1 which is less than long-run output OQ_2. But the short-run equilibrium price P_1Q_1 is higher than the long-run equilibrium price P_2Q_2. The total short-run monopoly profit is shown by the rectangle JP_1TK which is much smaller than the total long-run profit LP_2SM. This, however, is not necessary: it all depends on the short-run and long-run cost and revenue conditions. It may be noted at the end that if there are barriers to entry, the monopoly firm may not reach the optimal scale of production (OQ_2) in the long-run, nor can it make full utilization of its existing capacity. The firm's decision regarding plant

Fig. 15.2 *Equilibrium of Monopoly in the Long-run*

expansion and full utilization of its capacity depends solely on the market conditions.

15.5 PRICE DISCRIMINATION UNDER MONOPOLY

The theory of price determination under monopoly, as discussed above, gives the impression that under monopoly only one price is fixed and the same price is charged from all consumers of the product. This is, however, not the general case. Monopoly firms have been found to charge different prices from different class of consumers. This is called ***price discrimination.*** Price discrimination means selling the same or slightly differentiated product to different sections of consumers at different prices, not commensurate with

the cost of differentiation. Consumers are discriminated on the basis of their income or purchasing power, geographical location, age, sex, colour, marital status, quantity purchased, time of purchase, etc. There is another kind of price discrimination. The same price is charged from the consumers of different areas while cost of production in two different plants located in different areas is not the same. Some common examples of price discrimination, not necessarily by a monopolist, are given below:

(*i*) physicians and hospitals,6 lawyers, consultants, etc., charge their customers at different rates mostly on the basis of the customer's ability to pay;

(*ii*) merchandise sellers sell goods to relatives, friends, old customers, etc., at lower prices than to others and offer off-season discounts to the same set of customers;

(*iii*) railways and airlines charge lower fares from the children and students, and for different class of travellers;

(*iv*) cinema houses and auditoria charge differential rates for cinema shows, musical concerts, etc., from different class of audians;

(*iv*) some multinationals charge higher prices in domestic market and lower prices in foreign markets, called 'dumping', and

(*v*) lower rates for the first few telephone calls, lower rates for the evening and night trunk-calls; higher electricity rates for commercial use and lower for domestic consumption, etc. are some other examples of price discrimination.

15.5.1 Necessary Conditions for Price Discrimination

First, *different markets must be separable* for a seller to be able to practice discriminatory pricing. The markets of different classes of consumers must be so separated that buyers of one market are not in a position to resell the commodity in the other. Markets are separated by (*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; (*iii*) lack of distribution channels, e.g., transfer of electricity from domestic use (lower rate) to industrial use (higher rate).

Second, *the elasticity of demand for the product must be different in different markets.* The purpose of price discrimination is to maximize the profit by exploiting the markets with different price elasticities. It is the difference in the price elasticity which provides monopoly firm with an opportunity for price discrimination. If price elasticities of demand in different markets are the same, price discrimination would reduce the profit by reducing demand in the high price markets.

Third, *there should be imperfect competition in the market.* The firm must have monopoly over the supply of its product to be able to discriminate between different classes of consumers, and charge different prices.

Fourth, profit maximizing output must be much larger than the quantity demanded in a single market or by a section of consumers.

6. A study of medical doctors in the US by Ruben A. Kessel, "Price Discrimination in Medicine," *JL of Law and Eco.* , October 1959, reveals that charity, not profit maximization, is the objective of price discrimination. The idea behind charging a higher fee from the rich patients is to finance the treatment of poor. This is however a rare phenomenon. One can find such medical practioners and hospitals in India also.

15.5.2 Degrees of Price Discrimination

The degree of price discrimination refers to the extent to which a seller can divide the market or the consumers and can take advantage of it to extract the consumer's surplus. The economic literature presents three degrees of price discrimination.

First degree 7 The *first degree price discrimination* is the limit of discriminatory pricing. First degree or perfect price discrimination is feasible when the market size of the product is very small and the monopolist is in a position to know the price each consumer or each group of consumers is willing to pay, i.e., he knows his buyer's demand curve for his product. Under this condition, the monopolist sets the price accordingly and tries to extract the entire *consumer surplus.*8 What the seller does is that he sets the price at its highest level—the level at which all those who are willing to buy the commodity buy at least one unit each. After extracting the consumer surplus of this segment of consumers for the first unit of commodity, the monopolist gradually lowers down the price, so that the consumer surplus of the users of the second unit is extracted. This procedure is continued until the entire consumers' surplus available at the equilibrium price, i.e., at the price at which $MC = MR$, is extracted. Consider, for example, the case of medical services of exclusive use. 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 patient.

Second degree Where market size is fairly large, perfect discrimination is neither feasible nor desirable. In that case, a monopolist uses second degree discrimination or the 'block pricing method'. A monopolist adopting the *second degree price discrimination* intends to siphon off only the major part of the consumer's surplus, rather than the whole of it. The monopolist divides the potential buyers into blocks, e.g., rich, middle class and poor, and sells the commodity in blocks. The monopolist sells its product first to the rich customers at the highest possible price. Once this part of the market is supplied, the firm lowers down the price for middle class buyers. Finally, bottom price is used for the poor class of buyers. The second degree price discrimination is illustrated in Fig. 15.3.

The second degree price discrimination is
feasible where (*i*) the number of consumers is
large and price rationing can be done, as in

Fig. 15.3 *Second Degree Price*
Discrimination

case of utility services like telephones, supply of water, etc.; (*ii*) demand curve for all the consumers is identical; (*iii*) a single rate is applicable for a large number of buyers. As shown in Fig. 15.3, a monopolist practicing second degree price discrimination, charges the highest price OP_1 for OQ_1 units and a lower price OP_2 for the next Q_1Q_2 units, and the lowest price OP_3 for the next Q_2Q_3 units. Thus, by adopting a block pricing system, the monopolist maximizes his total revenue (*TR*) at

$$TR = (OQ_1 . AQ_1) + (Q_1Q_2 . BQ_2) + (Q_2Q_3 . CQ_3)$$

7. Joan Robinson calls it 'perfect discrimination' from a monopolist's point of view.

8. Consumer surplus is the difference between the price a consumer is willing to pay and the price he actually pays.

Third degree When a monopoly firm faces two or more well segregated markets, *it uses third degree price discrimation*. In other words, when a profit maximizing monopoly firm sets different prices in different markets having demand curves with different elasticities, it is practising the *third degree price discrimination*. It happens quite often that a monopolist has to sell its goods in two or more markets, completely separated from one another, 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 find different price-quantity combinations that can maximize profit in each market. For this purpose, the firm divides its total output between the market segments so that its $MC = MR$ in each market, and fixes price accordingly. For example, suppose that a monopoly firm has only two markets, *A* and *B*. The *ARa* and *ARb* curves represent demand curves of markets *A* and *B*, respectively. The *ARa* and *MRa* curves given in Fig. 15.4(*a*), represent the *AR* and *MR* curves in market *A*. Similarly, *Db* and *MRb* curves in Fig. 15.4(*b*) represent the *AR* and *MR* curves in market *B*. In order to determine the optimum output, the firm has to find out the aggregate *MR* curve. The horizontal summation of *ARa* and *ARb* curves gives the total demand curve for the two markets, as shown by $AR = D$ in Fig. 15.4(*c*) and the horizontal summation of *MRa* and *MRb* gives the aggregated *MR* [(Fig. 15.4(*c*)]. The firm's marginal cost is shown by *MC* that intersects *MR* at point *T*. Thus, the optimum level of output for the firm is determined at *OQ* at which $MR = MC$.

Fig. 15.4 *Third Degree Price Discrimination*

Given the size of the two markets, the whole of its output *OQ* cannot be sold in any one of the markets at a profit maximizing price. Therefore, the monopolist has to allocate output *OQ* between the two markets in such proportions that the necessary condition of profit maximization is satisfied in both the markets, i.e., $MC = MR$ in both the markets. This is accomplished by drawing a line from point *T* parallel to *X*-axis, through *MRb* and *MRa*. The points of intersection, *S* and *R* on curves *MRa* and *MRb*, respectively, determine the optimum share for markets *A* and *B*. As shown in the Fig. 15.4, the monopoly firm maximizes its profit in market *A* by selling *OQa* units at price *AQa* and in market *B*, by selling *OQb* units at price *BQb*. Note that $OQa + OQb = OQ$.

The third degree price discrimination may be suitably practised between any two or more markets separated from each other by geographical distance, transport barriers, cost of transportation and legal restrictions on the inter-regional transportation of commodities.

15.5.3 An Algebraic Solution

Price and output determination under third degree price discrimination has been illustrated graphically in Fig. 15.7. Here, we present an algebraic solution to the problem of price and output determination by a discriminating monopoly.

Let us suppose that a monopoly firm is faced with two markets, *A* and *B*, with two different demand functions given as $Qa = 16 - 0.5\,Pa$ and $Qb = 22 - Pb$.

The demand functions yield two different price functions given below.

$$Pa = 32 - 2\,Qa \qquad \text{...(15.4)}$$

and

$$Pb = 22 - Qb \qquad \ldots(15.5)$$

Suppose also that the firm's total cost function (*TC*) is given as

$$TC = 10 + 2Q + Q^2 \qquad \ldots(15.6)$$

Given the price and cost functions, the problem is how to determine the most profitable output and how to allocate this output between the two markets that profit in each market is maximised. Profit (p) is maximum where

$$p = TR - TC \text{ is maximum} \qquad \ldots(15.7)$$

In our example, *TC* function is known, but *TR* function is unknown. So we need to find *TR* function. For a price discriminating monopoly, total revenue (*TR*) equals the sum of revenue from the two markets. That is,

$$TR = Pa \cdot Qa + Pb \cdot Qb \qquad \ldots(15.8)$$

By substituting Eqs. (15.4) and (15.5) for *Pa* and *Pb*, respectively, in Eq. (15.8), we get

$$TR = (32 - 2Qa)\,Qa + (22 - Qb)\,Qb$$

$$= 32Qa - 2Qa^2 + 22Qb - Qb^2 \qquad \ldots(15.9)$$

Now total profit (p) can be obtained by substituting Eqs. (15.6) and (15.9) for *TC* and *TR*, respectively, in Eq. (15.7). Thus, we get the profit function as

$$p = 32Qa - 2Qa^2 + 22Qb - Qb^2 - (10 + 2Q + Q^2)$$

$$= 32Qa - 2Qa^2 + 22Qb - Qb^2 - 10 - 2Q - Q^2 \qquad \ldots(15.10)$$

For profit to be maximum, *Q* in Eq. (15.10) must be equal to profit maximizing sales in markets *A* and *B*. That is,

$$Q = Qa + Qb$$

By substituting, *Qa* + *Qb* for *Q* in Eq. (15.10), we can rewrite it as

$$p = 32Qa - 2Qa^2 + 22Qb - Qb^2 - 10 - 2(Qa + Qb) - (Qa + Qb)^2$$

$$= 32Qa - 2Qa^2 + 22Qb - Qb^2 - 10 - 2Qa - 2Qb - Qa^2 - 2QaQb - Qb^2$$

$$= 30Qa + 20Qb - 3Qa^2 - 2Qb^2 - 2QaQb - 10 \qquad \ldots(15.11)$$

Equation (15.11) represents the total profit function. A necessary condition for p to be maximum is that marginal change in profit must be equal to zero. Total profit is composed of profits in markets *A* and *B*. It implies, therefore, that for total profit to be maximum, marginal change in profit in both the markets must be equal to zero. The marginal change in profits in markets *A* and *B* can be expressed in terms of first derivative of the total profit-function with respect to Q_a and Q_b. Thus, marginal profit in market *A* can be expressed as

$$\frac{\partial \pi}{\partial Q_a} = 30 - 6Q_a - 2Q_b \quad ...(15.12)$$

and for market *B*, as

$$\frac{\partial \pi}{\partial Q_b} = 20 - 4Q_b - 2Q_a \quad ...(15.13)$$

The profit maximizing condition may be restated by setting the marginal profit functions (15.12) and (15.13) equal to zero. Thus, for profit to be maximum in market *A*,

$$30 - 6Q_a - 2Q_b = 0 \quad ...(15.14)$$

and in market *B*,

$$20 - 4Q_b - 2Q_a = 0 \quad ...(15.15)$$

We have now two simultaneous equations—Eqs. (15.14) and (15.15) —with two unknowns (Q_a and Q_b), which can be solved for Q_a and Q_b as follows.

$$30 - 6Q_a - 2Q_b = 0 \quad (1)$$

$$20 - 2Q_a - 4Q_b = 0 \quad (2)$$

In order to solve for Q_b, Eq. (2) is multiplied by 3 and subtract from Eq. (1).

$$30 - 6Q_a - 2Q_b = 0$$

$$60 - 6Q_a - 12Q_b = 0$$

$$+ \qquad +$$

$$-30 + 10Q_b = 0$$

$$10Q_b = 30$$

$$Q_b = 3$$

The value of Q_a can now be obtained by substituting 3 for Q_b in Eq. (1) or (2). Thus,

$$30 - 6Q_a - 2(3) = 0$$

$$-6Q_a = -24, \; Q_a = 4$$

To conclude, the monopoly firm maximizes its total profit by selling 4 units in market *A* and 3 units in market *B*.

Price Determination

The profit maximizing prices can now be obtained by substituting Q_a and Q_b with their estimated values (4 and 3, respectively) in price functions (15.4) and (15.5), respectively. The price for market *A* can be obtained as

$$P_a = 32 - 2Q_a = 32 - 2(4) = 24$$

and price for market *B* as

$$P_b = 22 - Q_b = 22 - 3 = 19$$

Thus, in market *A*, price = `24 and in market *B*, price = `19.

Profit Determination

Now that prices and sales for the two markets are known, total profit can be obtained by substituting numerical values for Q_a and Q_b in profit function (15.11) reproduced below.

$$\pi = 30Q_a + 20Q_b - 3Q_a^2 - 2Q_b^2 - 2Q_aQ_b - 10$$

By substituting 4 for Q_a and 3 for Q_b, we get

$$\pi = 30(4) + 20(3) - 3(4)(4) - 2(3)(3) - 2(4)(3) - 10$$

$$= 120 + 60 - 48 - 18 - 24 - 10 = 80$$

The total profit is `80. This profit satisfies the conditions of the maximum profit. It is, therefore, maximum.

15.6 EQUILIBRIUM OF A MULTI-PLANT MONOPOLY

We have so far analysed the equilibrium of a single-plant monopoly firm. In this section, we explain the case of a multi-plant monopoly—a monopoly firm producing a homogeneous product in more than one plant. For the sake of simplicity, we make the following assumptions.

1. A monopoly firm has two plants *A* and *B;*
2. Its cost conditions differ from one plant to another, and
3. The firm is aware of its *AR* and *MR* curves.

Let us now see, how a profit maximizing monopoly firm determines its total output and how it allocates the total output between the two plants, so that output of each unit is optimum. The cost conditions of the two plants are given as shown in Fig. 15.5(a) and (b). The combined marginal cost, *MC*, can be obtained by horizontally summing the *MCA* and *MCB*, so that $MC = MCA + MCB$. The combined *MC* curve is shown in panel (c) of Fig. 15.5. When *MR* and *MC* are known, the monopolist can easily find the profit maximizing output and price. Going by the profit maximizing rule ($MC = MR$), the total equilibrium output would be *OQ* as shown in panel (c) of Fig. 15.5. Now. the problem is how to allocate *OQ* between the two plants, so that output of each plant is optimum. The optimum output for each plant can be obtained by applying the profit maximization rule, i.e. in each plant $MR = MC$ so that

$MR = MCA = MCB$

Determination of Plant Output. The profit maximizing output of each plan that satisfies the profit maximising rule can be obtained by drawing a horizontal line $MR = MC$ (parallel to X-axis) from point *E* through *MCB* and *MCA*. As shown in Fig. 15.5(c) the points of intersection E_1 and E_2 , determine the optimum output for plants *A* and *B*, respectively, Thus, the optimum output for plant *A* is determined at *OQA* and *OQB* for plant *B.*

Figure 15.5 *Multi-Plant-Monopoly Output Allocation*

Determination of Price. Turning to the question of price determination, the profit maximization price is determined at *OP* (= *QT)* in panel (c) of Figure 15.5. Once profit-maximizing price is determined at *OP*, the total maximum profit of each plant can he worked out. For example, total profit from plant *A* can he worked out as follows.

Total profit of Plant $A = (AR – AC)\,Q$. In plant *A,*

$AR = OP = aO$; $AC = dO = cQA$; and $Q = OQA = dc$

Thus, total profit of plant can be worked out as follows

Total profit of plant

$A = (aO – dO)\, dc$

$= a\,b\,c\,d$ at price *OP*

Thus, at price *OP*, total profit from plant *A* is *abcd* (= $ad \times OQA$). The same method can be used find the total profit of plant *B*. The total profit from plant *B* is *efgh* (= $eh \times OQB$). Although this analysis is based on only two plants, it can be extended to a large number of plants.

15.7 MONOPOLY *VS*. PERFECT COMPETITION: COMPARISON OF PRICE AND OUTPUT

In this section, we answer a policy related question: Is perfect competition or monopoly is more desirable from social welfare point of view? This question can be answered (*i*) by comparing the price and output in the two kinds of markets, and (*ii*) by measuring the deadweight loss under the two kinds of market conditions.

15.7.1 Comparison of Long-run Price and Output

Comparison of long-run price and output under monopoly and perfect competition gives the final answer. Therefore, we compare here long-run price and output under monopoly and perfect competition.

A comparative analysis of equilibrium price and output under perfect competition and monopoly in the long run can be had from Fig. 15.6. Let us assume that *LMC* and *LAC* curves are identical in both competitive industry and monopoly. The long-run equilibrium of a competitive industry requires that all its firms are in equilibrium, i.e., all the firms have their

Figure 15.6 *Comparison of Price and Output: Monopoly vs Perfce Competition*

$AR = MR = LAC = LMC$. This condition is satisfied at point E_1 in Fig. 15.6. Thus, in a competitive industry, equilibrium price is OP_1 and equilibrium output OQ_2. Now if this industry is monopolized, the revenue conditions (*AR* and *MR)* will be different. The *AR* and *MR* curves for the monopolized industry are given by the *AR* and *MR* curves, respectively. Given its *AR* and *MR* curves, the monopoly firm will reach its equilibrium and maximize its total profits at the level of output at which its $MR = MC$. The equilibrium level of output is shown by point *B*—the point of intersection of *MR* and *LMC* curves. Thus, the equilibrium output under monopoly will be OQ_1, and the equilibrium price will be OP_2.

Conclusion. As the foregoing analysis shows, output under monopoly (OQ_1) is lower than that under perfect competition (OQ_2), and price under monopoly (OP_2) is higher than that (OP_1) under perfect competition. Under these conditions, therefore, perfect competition more desirable than monopoly.

15.7.2 Deadweight Loss under Monopoly

A general complaint against monopolies is that monopoly causes loss of social welfare and distortions in resource allocation. The suboptimal allocation of resources and loss of social welfare are illustrated in Fig. 15.7, assuming a constant-cost industry winch has $LAC = LMC$. The revenue conditions are shown by the *AR* and *MR* curves. Given the cost and revenue conditions, a perfectly competitive industry will produce OQ_2, at winch $LAC = LMC = AR$. Its price will be OP_1.

Under the same cost and revenue conditions, a monopoly firm finds its equilibrium at point *K* at which the monopoly firm produces OQ_1, and charges price OP_2. The comparison of prices and outputs under monopoly and perfect competition gives the measure of the loss of social welfare.

Figure 15.7 *Deadweight Loss Under Monopoly*

Loss of Social Welfare under Monopoly The loss of social welfare is measured in terms of loss of consumer surplus. The total consumer surplus equals the difference between the total price which a society is willing to pay for the consumption of a commodity and the total price that it pays for that commodity. If an industry is perfectly competitive, the

total output available to the society will be OQ_2 at price OP_1 (see Fig. 15.7). The total price which the society pays for OQ_2, is given by the area $OP_1LQ2 = OP_1 \times OQ_2$. The total price which it is willing to pay for the output $OQ2$ is given by the area $OALQ_2$ which, in Marshallian terminology, is the value which society would be willing to pay for output OQ_2. Thus,

Consumer's surplus = $OALQ2 - OL_1LQ2 = ALP_1$

If the industry is monopolized, the equilibrium output is set at OQ_1 and price at OP_2. This leads to a loss of a part of consumer surplus.

Loss of consumer surplus under monopoly = $ALP_1 - AMP_2 = P_2MLP_1$.

Of this total loss of consumer surplus (P_2MLP_1) , P_2MKP_1 goes to the monopolist as monopoly profit. The remainder $MKL = P2MLP$ l $= P_2MKP_1$ goes to none. Therefore, it is termed as *dead-weight loss* to the society caused by monopoly.

Empirical Evidence of Deadweight Loss Theoretically, there is a strong possibility of a deadweight loss under monopoly in comparison to perfect competition. However, empirical evidence of the deadweight loss attributed to monopoly is inconclusive. Harberger[9] was the first to estimate the loss of welfare due to monopoly in the United States. To estimate the loss of welfare, he assumed constant cost and unitary elastic demand curve for all industries and used 1924-1928 data of 73 manufacturing industries. He regarded the loss of welfare as the difference between the monopoly price and the competitive price. According to his estimates, the US economy had suffered a loss of social welfare to the extent of 0.1 per cent of its national income, which was obviously negligible. An implication of this finding was that efforts to curb monopolies were useless. Interestingly, 'One economist quipped that economists might make a larger social contribution fighting fires and eradicating termites than attempting to curb monopolies.[10]

Although many other studies confirmed Harberger's finding, some economists were highly critical of his data and approach. For example, Stigler[11] criticized Harberger's results on the ground of statistics used in his study. Kamerschen[12] used 1956–1961 data in a similar study and produced a considerably different result. According to his findings, the welfare loss due to monopolies in the US was about 6 per cent of national income. Scherer and Ross find welfare loss due to monopolistic mal-allocation of resources somewhere between 0.5 and 2 per cent of the US GDP. Some other critics of Harberger have estimated the social loss to be of the order of 4–8 per cent of national income. However, the methodology and statistics used by Harberger's critics have been criticized by others. In the mean time, some other aspects of monopoly have been added to the debate, e.g. the dynamic efficiency that arises due to technological advanced made by monopolistic competition, the poor quality of goods supplied by them, advantage of the resources unused by the monopolies, and the use and abuse of monopoly profits. Finally, it may be concluded that the issue of deadweight loss attributed to monopolies remains a *controversial issue*.

9. Harberger, A.C., "Monopoly and Allocation", *Am. Eco., Rev*., 44, May 1954.

10. Samuelson, P.A. and Nordhaus, W.D., *Economics*, 15th Edn. (McGraw-Hill, Inc. New York, 1995), p. 175.

11. Stigler, G.J. (1956), 'The Statistics of Monopoly and Merger', *Journal of Polymer*, February, 1956.

12. Kamerschen, David R., "An Estimation of the 'Welfare Losses' from the Monopoly in the American Economy," *Economic Journal*, , 4, Summer 1966.

SUMMARY

• Although economists have offered different views on the meaning of monopoly, for all practical purposes, monopoly refers to an absolute power of a firm to produce and sell a commodity which has no close substitute.

• The factors that lead to emergence of a monopoly include: (*i*) legal restriction on entry of new firms, (*ii*) ownership and control over essential raw material, (*iii*) large scale production and economies of scale, and (*iv*) managerial efficiency and product.

• In short run, cost and revenue curves of the monopoly firm are given. Given the cost and revenue condition, equilibrium output of a monopoly firm is determined at the profit maximizing, i.e., the level of output at which $MC = MR$. Once profit maximizing output is determined, price is automatically determined given the demand (*AR*) curve.

• In short run, a monopoly firm may not necessarily make abnormal profit or economic profit. It all depends on its cost and revenue curves. Give the cost and revenue curves, it may make losses too in short run. But it does make abnormal profit in the long-run.

• In long run, firm's scale of production increases. The monopoly firm gains from the economies of scale and its production cost decreases. Besides, demand for the monopoly product may also increase. So the firm has long-run cost and revenue curve. The rule of price and output determination in long run is the same as in short-run.

• Comparison of price and output under monopoly and perfect competition shows that output is large and price lower under perfect competition whereas it is other way under monopoly.

• It has been theoretically as well as emperically it has been proved that perfect competition contributes to social welfare whereas monopoly leads to deadweight loss to the society.

REVIEW QUESTIONS AND EXERCISES

1. What is monopoly? How does existence of a close substitute affect the monopoly power? What are the factors that create conditions for the emergence and survival of a monopoly?

2. Are the revenue and cost curves under monopoly different from those under perfect competition? Illustrate and explain the difference between *AR* and *MR* curves faced by monopoly and a competitive firm.

3. Write a note on the relationship between average revenue and marginal revenue under (*i*) perfect competition and (*ii*) monopoly.

4. Explain the equilibrium of a monopoly firm in the short run by using short run *AC*, *MC*, *AR* and *MR* curves. Why is monopoly price always higher than the competitive price?

5. A monopoly firm may earn normal or abnormal profits or may even incur losses in the short run, Do you agree with this statement? Give reasons for your answer.

6. Will a monopolist remain in business in the short run if it is just covering its average variable costs? Explain with the help of a diagram.

7. Suppose demand function for a monopoly product is given as $Q = 500 – 0.5 P$ and cost function is given as $C = 100 + 40 Q + Q 2$. Find the equilibrium level of price and output for the monopoly firm.

8. Compare monopoly and perfect competition with regard to the following: (*i*) price, (*ii*) output, (*iii*) welfare cost and (*iv*) relationship between *MC* and price.

9. Show the difference between the long-run equilibrium of a competitive firm and the long-run equilibrium of a monopoly firm with regard to the following: (*i*) price, (*ii*) profits and (*iii*) use of capacity.

10. What is price discrimination? Explain and distinguish between the first, second and the third degrees of price discrimination. Which one makes a general case for price discrimination and why?

11. What is a discriminating monopoly? What are the conditions which force the monopolist in practicing price discrimination? State the conditions under which price discrimination is possible and profitable as well.

12. Suppose a monopolist faces two markets, *A* and *B*, with their respective demand functions given as $QA = 500 – 0.5 P$ and $QB = 980 – 2 P$. Find the equilibrium price and output sales for each market.

13. What are the necessary conditions of price discrimination under monopoly? Show how a profit-maximizing discriminating monopolist allocates his output between two markets and charge different prices?

14. Show graphically the determination of profit maximizing equilibrium of discriminating monopolist. Is price discrimination socially desirable?

15. (*a*) Explain how profit maximization by a monopoly firm reduces public welfare vis-à-vis a competitive firm,

(*b*) Why is price discrimination under monopoly considered to be economically desirable?

16. Suppose a simple monopolist is in equilibrium. At the point of equilibrium, the coefficient of price elasticity is – 2 and marginal cost is `4.0. Calculate monopoly's equilibrium price. How will this price be affected by an increase in the fixed cost of the monopolist?

17. Calculate the ratio of prices charged by a discriminating monopolist in two markets, *A* and *B*, having price-elasticities of demand as – 0.5 and – 1.5, respectively.

18. What is meant by monopoly power? What are the methods of measuring the degree of monopoly power?

19. Write short notes on the following criteria of monopoly power.

(*a*) Concentration ratio,

(*b*) Excess profitability criterion, and

(*c*) Cross-elasticity criterion.

20. Which of the following statements are correct?

(*a*) A monopolist can charge any price to maximize profits.

(*b*) A monopoly firm can fix its price anywhere along the demand curve,

(*c*) If monopoly's $MC = 0$, it fixes its price where $e = 0$.

(*d*) The slope of monopoly's *MR* curve is twice that of *AR* curve.

(*e*) A monopolist is in equilibrium where $MC = MR$.

(*f*) A necessary condition for monopoly's long-run equilibrium is $AC = AR = MC = MR$.

(*g*) A monopolist produces always less than its optimum capacity.

(*h*) There is no unique relationship between price and supply under monopoly.

(*i*) Equilibrium price of a monopolist is always higher than that of a competitive firm.

(*j*) Price discrimination is possible only when demand curves are identical in two markets,

[**Ans.** (*d*) (*e*) (*h*) (*i*)]

21. Suppose market demand function for monopoly firm is given by = 100 – 5 *P*. Find the following.

(*a*) Price function or reverse demand function for the monopolist;

(*b*) *AR*-function of the monopolist;

(*c*) *MR*-function of the monopolist;

(*d*) *TR*- function of the monopolist.

22. Suppose a monopoly firm sells it product in two different markets with their respective demand functions given as follows.

$Q_1 = 500 – P_1$,

$Q_2 = 300 – P_2$

Firm's total cost function is given as $TC = 50000 + 100 Q$

Find:

(*a*) profit maximizing output,

(*b*) profit maximizing price,

(*c*) prices for the two markets, and

(*d*) total profit at equilibrium.

FURTHER READING

Bain, J.S. *Pricing Distribution and Employment* (New York: Holt Rhine, 1953), Chapters 5 and 9.

Basankso, D.A. and Braeutigam, R.R., *Microeconomies: An Integrated Approach* (New York: John Wiley & Sons, Inc., 2002), Chapters 11 and 12.

Blaug, M., *Economic Theory in Retrospect* (Homewood, IL; Richard D. Irwin, 1970).

Browning, K.E. and Browning, J.M., *Microeconomics: Theory and Applications* (New Delhi: Kalyani Publishers, 1986), 2nd Edn., Chapters 11 and 12.

HerBerger, A.C., 'Monopoly and Resource Allocation', *American Economics Review,* 1954 .

koutsoyiannis, A., *Modern Microeconomics* (London: Macmillan, 1954), Chapters 6 and 7.

lerner, A.P., 'The Concept of Monopoly and Measurement of Monopoly Power', *Review of Economic Studies*, 1953, 157–175.

MarsHall, A., *Principles of Economics* (London: Macmillan, 1920), 8th Edn., Book V, Chapter 14.

Perloff, J.M., *Microeconomics*, (New York, NY: Addison Wesley, 2001), 2nd Edn., Chapters 11 and 12.

Pigou, A.C., *The Economics of Welfare* (London: Macmillan, 1920).

Pindyck, R.S. and D.L. Rubinfeld, *Microeconomics* (London: Prentice Hall International, 2001), 5th Edn., Chapters 10 and 11.

roBinson, J., *The Economics of Imperfect Competition* (London: Macmillan, 1933), Chapters 2, 3, 15 and 16.

CHAPTER

16 Price and Output Determination under Monopolistic Competition

CHAPTER OBJECTIVES

The objective of this Chapter is to discuss how price and output are determined under monopolistic competition. The discussion of the subject covers the following aspects.

- Definition and features of monopolistic competition
- Measures of monopoly power of a firm in monopolistic market
- Comparison of monopolistic competition and perfect competition
- Price and output determination under monopolistic competition in short run
- Price and output determination under monopolistic competition in long run
- Non-price competition: competition by selling cost
- Critical appraisal of selling cost theory
- Summary

16.1 INTRODUCTION

In two preceding chapters, we have discussed the traditional theory of price and output determination under two kinds of the rarest of the rare markets, viz., perfect competition and monopoly. Until the 1920s, "the theories of perfect competition and monopoly constituted the 'classical' microeconomics from Marshall to Knight"[1] After WW I, however, there was rapid economic change in war-affected countries and also a drastic change in market structure. Because of change in market structure, the theories of price determination under perfect competition and monopoly lost their relevance. During the late 1920s and the early 1930s, economists expressed their dissatisfaction with perfect competition and pure monopoly models. In their opinion, these models were no more relevant to analyse the business behaviour of the real world. Piero Shraffa[2] was one of the first of those who pointed out the irrelevance and limitations of perfect competition and pure monopoly models to analyse the then existing market conditions. His view was later supported by

1. Ferguson, C.E., *Microeconomic Theory*, 2nd Edn (Homewood, IL, Richard D. Irwin), p. 317.
2. Piero Shraffa, " The Laws of Returns to Competitive under Competitive Conditions" *Economic Journal*, 1926, pp. 535–50.

Hotelling[3] and Zeuthen[4]. They claimed that neither perfect competition nor monopoly represented the real world market structure. In their opinion, the real business world conforms to a market structure which combines both perfect competition and monopoly, i.e., the real world market is represented by ***monopolistic competition***. Following this view, Joan Robinson[5] and Edward H. Chamberlin[6], formulated their own theories of *price determination under monopolistic competition* in 1933. However, Chamberlin's theory of price determination under monopolistic competition is considered to be more logical and theoretically sound. Although Chamberlin's theory has lost its relevance due to changing market conditions, it is respected for its theoretical soundness. In this Chapter, we discuss Chamberlin's theory of price and output determination under monopolistic competition.

16.2 MONOPOLISTIC COMPETITION: DEFINITION AND CHARACTERISTICS

16.2.1 Definition

By definition, monopolistic competition refers to a market structure in which a large number of sellers sell differentiated products, which are close substitutes for one another. Incidentally, a close substitute is one whose cross-elasticity is close to unity or greater. Monopolistic competition combines the basic elements of both perfect competition and monopoly.

The ***element of monopoly*** in monopolistic competition arises from the fact that each firm has an absolute right to produce and sell a ***branded*** or ***patented*** product. Other firms are prevented by laws from producing and selling a branded product of other firms, This gives a firm ***monopoly power*** over production, pricing and sale of its own-banded product. For example, consider toilet soap industry. There are a number of brand names available in the market, e.g., Lux, Liril, Palmolive, Fairglow, Pears, Fa, Rexona, Lifebuoy, Carmel, Godrej, Cinthol, Ponds, Dove, Dettol and so on. Each of these branded toilet soaps is produced and sold by a company having monopoly power over the product. Similarly, Maruti Udyog Limited has monopoly power for producing and selling cars under the brand name Maruti. No other car manufacturing company can produce and sell cars under this brand name, So is the case with all other car manufacturing companies.

The ***element of competition*** arises from the fact that each generic branded product has several close substitutes and firms selling branded products of the same generic category have to compete for the market share. Considering again our example of toilet soaps, all the companies producing and selling branded toilet soaps are in intensive competition for capturing the largest possible market share. One index of the competition between them is the amount that they spend advertising their product. These features of the toilet soap industry make it monopolistically competitive. Toothpaste industry with a number of branded product names (e.g. Binaca, Colgate, Close-up, Pepsodent, Forhans, Cibaca, Neem, Meswak, Signal, Promise, Prestige and so on) is another example of monopolistic competition. So is the case with major industrial products in India, e.g., electrical tubes and bulbs, TV sets, refrigerators, air conditioners, personal computers, textile goods, tea,

3. Harold Hotelling, "Stability and Competition", *Economic Journal*, 1929.

4. Zeuthen, A., *Problems of Monopoly and Economic Welfare* (London: Routledge, 1930).

5. Joan Robinson, *The Economics of Imperfect Competition* (London, Macmillan, 1933).

6. Edward E. Chamberlin, *Theory of Monopolistic Competition* (Cambridge, Mass, Harvard University, 1933).

coffee, cigarettes, soft drinks, cold creams, sham-poos, detergents, shaving blades, shaving cream, hair oils, hair dyes, shoes, wrist watches, steel, cement, mobile phones and so on. Some of the industries looking monopolistically competitive may be *oligopolistic* in which there are only a few sellers selling differentiated or homogenous products. The question as to what makes a market monopolistically competitive or oligopolistic will be

taken up in the next chapter which deals with *oligopoly markets*. Let us now look at the general characteristics of monopolistic competitions.

16.2.2 Characteristics of Monopolistic Competition

As mentioned earlier, monopolistic competition combines the elements of both perfect competition and monopoly power. Therefore, the main characteristics of monopolistic competition are the blend of perfect competition and monopoly. The main features of monopolistic competition *vis-à-vis* perfect competition and monopoly are described below.

1. ***Product Differentiation.*** Product differentiation is the basis of and the main distinctive characteristic of monopolistic competition that distinguishes it from monopoly and perfect competition. In case of monopoly, there is only one product and only one seller of the product, and under perfect competition, a large number of sellers sell *homogeneous product*. Under monopolistic competition, the firms differentiate their products from one another. Product are differentiated with respect to their shape, size, colour, design, minor qualitative differences, efficiency in use, some extra facility, packaging, after-sale-service, guarantee and warrantee and so on. Product differentiation may be real or fanciful and

spurious. The basic purpose of product differentiation is to make the consumers believe that a product is different from others and, thereby, to create brand loyalty of the consumers. Product differentiation affects firm's demand curve in a significant way.

2. ***Large Number of Sellers.*** Under monopolistic competition, the number of sellers is *large*. How large? It is difficult to specify number of firms: it may be 10, 20 or more depending on the size of the market. However, the question 'how large' can be answered in conceptual terms with reference to perfect competition. Under perfect competition, the number of sellers is so large that a firm becomes a *price taker*. In contrast, under monopolistic competition, the number of firms is only so large that a firm retains its power to be a *price maker*. The monopolistically competitive firms have the power to set the price of their product depending on the objective of the firm.

3. ***Free Entry and Free Exit.*** As in case of perfect competition, there is no barrier to the entry of new firms and exit of old ones from the industry. New firms are free to enter the monopolistically competitive industry and to quit at will. Entry of new firms reduces the market share of the existing ones and exit of firms does the opposite. These consequences of free entry and free exit lead to intensive competition among the firms for retaining as well as increasing their market share.

4. ***Selling Costs.*** Unlike firms under perfect competition and monopolies, under monopolistic competition firms make heavy expenditure on advertisement and other sales promotion measures for their product. This is an important feature that distinguishes monopolistic competition from perfect competition and monopoly. *Selling costs* include all the expenditure on advertisement, sales promotion schemes and salaries of sales personnel. Selling costs and their effect on firm's equilibrium will be discussed below in detail.

5. ***Downward Sloping Demand Curve.*** As in case of monopoly, a monopolistically competitive firm faces a downward sloping demand curve. The reason is that a monopolistically competitive firm can increase its price and still retain some buyers with brand loyalty and can increase the demand for its product by decreasing the price because of a relatively higher cross-elasticity of the competitive product.

16.3 PRICE AND OUTPUT DECISIONS IN THE SHORT-RUN

The price and output determination models of monopolistic competition was developed by Chamberlin as explained here.

Although monopolistic competition is characteristically close to perfect competition, pricing and output decisions under this kind of market are similar to those under monopoly. The reason is that a firm under monopolistic competition faces a downward sloping demand curve.7 This kind of demand curve is the result of (*i*) a strong preference of a section of consumers for the product and (*ii*) the quasi-monopoly of the seller over the supply. The strong preference or brand loyalty of the consumers gives the seller an opportunity to raise the price and

Fig. 16.1 *Price-Output Determination under*

yet retain some customers. Besides, since

Monopolistic Competition

each product is a substitute for the other, the firms can attract the consumers of other products by lowering their prices.

The short-term pricing and output determination under monopolistic competition is illustrated in Fig. 16.1. The *AR* and *MR* curves and *SMC* and *SAC* curves give short-run revenue and cost curves, respectively, faced by the monopolistic firm. As shown in the figure, firm's *MR* intersects its *MC* at point *N*. Point *N* determines the profit maximizing output at *OQ*. Given the demand curve, this output can be sold at price *PQ*. So the price is determined at *PQ*. At this output and price, the firm earns a maximum monopoly or economic profit equal to *PM* per unit of output and a total monopoly profit shown by the rectangle P_1PMP_2. The economic profit, *PM* (per unit) exists in the short-run because there is no or little possibility of new firms entering the industry. But 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 their products. Some firms may earn only a normal profit if their costs are higher than those of others. For the same reason, some firms may make even losses in the short-run.

16.4 PRICE AND OUTPUT DETERMINATION IN THE LONG-RUN

The mechanism of price and output determination in the long-run under monopolistic competition is illustrated graphically in Fig. 16.2. To begin the analysis, let us suppose that,

7. It has also been argued that the demand curve under monopolistic competition is indeterminate. However, for analytical convenience, it is assumed that the firms under monopolistic competition face an identical downward sloping demand curve.

at some point of time, firm's revenue curves are given as AR_1 and MR_1 and long-run cost curves as *LAC* and *LMC*. As the figure shows, MR_1 and *LMC* intersect at point *M* determining the equilibrium output at OQ_2 and price at P_2Q_2. At price P_2Q_2, the firms make a supernormal or economic profit of P_2T per unit of output. This situation is similar to short-run equilibrium.

Let us now see what happens in the long run. The supernormal profit brings about two important changes[8] in a monopolistically competitive market in the long run.

First, the supernormal profit attracts new firms to the industry. As a result, the

Fig. 16.2 *The Long-Run Price and Output*

existing firms lose a part of their market

Determination under Monopolistic Competition

share to new firms. Consequently, their demand curve shifts downward to the left. This kind of change in the demand curve is shown is Fig. 16.2 by the shift in *AR* curve from AR_1 to AR_2 and the *MR* curve from MR_1 to MR_2.

Second, the increasing number of firms intensifies the price competition between them. Price competition increases because losing firms try to regain or retain their market share by cutting down the price of their product. And, new firms in order to penetrate the market set comparatively low prices for their product. The price competition increases the slope of the firms' demand curve[9] or, in other words, it makes the demand curve more elastic. Note that AR_2 has a greater slope[10] than AR_1 and MR_2 has a greater slope than MR_1.

The ultimate picture of price and output determination under monopolistic competition is shown at point P_1 in Fig. 16.2. As the figure shows, *LMC* intersects MR_2 at point *N* where firm's long-run equilibrium output is determined at OQ_1 and price at P_1Q_1. Note that price at P_1Q_1 equals the *LAC* at the point of tangency. It means that under monopolistic competition, firms make only normal profit in the long-run. Once all the firms reach this stage, there is no attraction (i.e., super normal profit) for the new firms to enter the industry, nor is there any reason for the existing firms to quit the industry. This signifies the long-run equilibrium of the industry.

Mathematical Solution

To illustrate the price and output determination under monopolistic competition through a numerical example, let us suppose that the initial demand function for all the firms is given as

$$Q_1 = 100 - 0.5 P_1$$

8. The other important change is increase in advertisement cost.

9. For a detailed analysis of 'free entry' and 'price competition' and of their combined effects, see author's *Microeconomic Theory and Application* (Pearson Education, 2012), Ch. 18, *Appendix*.

10. Whether slope of the *AR* increases or not is a matter of empirical verification. However, the theory of pricing under monopolistic competition assumes that at least the firms believe that the demand curve for their product is more elastic than the market demand curve.

and price function as

$$P_1 = 200 - 2Q_1 \quad \ldots(16.1)$$

Given the price function (16.1), firms' TR_1 function can be worked out as

$$TR_1 = P_1 . Q_1 = (200 - 2Q_1)Q_1$$
$$= 200Q_1 - 2Q_1^2 \quad \ldots(16.2)$$

The marginal revenue function (MR_1) can be obtained by differentiating the TR_1 function (16.2). Thus,

$$MR_1 = 200 - 4Q_1 \quad \ldots(16.3)$$

Suppose also that firms face identical cost curves and that firms' long-run TC function is given as

$$TC = 1562.50 + 5Q - Q^2 + 0.05Q^3 \quad \ldots(16.4)$$

Given the firms' long-run TC function, LAC can be obtained as

$$LAC = \frac{TC}{Q} = \frac{1562.50 + 5Q - Q^2 + 0.05Q^3}{Q}$$
$$= \frac{1562.50}{Q} + 5 - Q + 0.05Q^2 \quad \ldots(16.5)$$

We get firms' LMC function by differentiating its TC function (16.4). Thus,

$$LMC = 5 - 2Q + 0.15Q^2 \quad \ldots(16.6)$$

Short-run "Status" Given the revenue and cost functions, let us now work out the *equilibrium* levels of output and price that maximize firms' profit assuming short-run conditions. The profit maximizing output can be obtained by equating MR_1 and LMC functions given in Eqs. (16.3) and (16.6), respectively, and solving for Q_1. That is,

$$MR_1 = LMC$$
$$200 - 4Q_1 = 5 - 2Q + 0.15Q^2 \quad \ldots(16.7)$$

For uniformity sake, let us replace Q in MC function as Q_1 and solve[11] Eq. (16.7) for Q_1.

$$200 - 4Q_1 = 5 - 2Q_1 + 0.15Q_1^2$$
$$195 = 2Q_1 + 0.15Q_1^2$$
$$Q_1 = 30$$

Thus, profit maximizing output in the short-run equals 30.

Let us now find firms' equilibrium price (P_1), LAC and supernormal profit. Price P_1 can be obtained by substituting 30 for Q_1 in the price function (16.1).

$$P_1 = 200 - 2Q_1$$
$$= 200 - 2(30) = 140$$

11. Alternatively, Eq. (16.7) can be solved for Q_1 by setting the equation equal to zero. Thus,

$$0.15Q_1^2 + 2Q_1 - 195 = 0$$

In order to eliminate fraction 0.15, let us multiply both sides by 20. We get an equation as

$$= 3Q_1^2 + 40Q_1 - 3900 = 0$$

By using quadratic formula, we get

− 40 ± 1600 + 12 × 3900 − 40 ± 1600 + 46800

Q =

− 40 + 220

=

=

6

6

6

Q = 30

Thus, firms' equilibrium price is determined at `140.

Firms' *LAC* can be obtained by substituting equilibrium output 30 for Q in function (16.5). Thus,

LAC = 1562.50 + 5 – 30 + 0.05 (30 × 30) = 72.08

30

Thus, *the short-run equilibrium condition* gives the following data.

Equilibrium output = 30

P 1 = 140

LAC = 72.08

Supernormal profit = *AR* 1 – *LAC* = 140 – 72.08 = 67.92 (per unit of output)

Long-run Status Let us now see what happens in the long-run. As already mentioned, the existence of supernormal profit attracts new firms to the industry in the long-run. Consequently, old firms lose a part of their market share to the new firms. This causes a leftward shift in their demand curve with increasing slope. Let us suppose that given the long-run *TC* function, firms' short-run demand and price functions given in Eq. (16.1) change in the long-run take the following form.

Q 2 = 98.75 – P 2

and

P 2 = 98.75 – Q 2

…(16.8)

To work out the ***long-run equilibrium***, we need to find the new *TR* function (*TR* 2) and the new *MR* function (*MR* 2) corresponding to the new price function (16.8). For this, we need to first work out the new *TR* function (*TR* 2).

TR 2 = P 2. Q 2 = (98.75 – Q 2) Q 2

= 98.75 Q

2

2 – Q 2

…(16.9)

We get *MR* 2 by *differentiating TR* function (16.9). Thus,

MR 2 = 98.75 – 2 Q 2

…(16.10)

The long-run equilibrium output can now be obtained by equating *MR* 2 with the *LMC* function (16.6). For the sake of uniformity, we designate Q in the *LMC* function as Q 2. The long-run equilibrium output is then determined where

MR 2 = *LMC*

or

98.75 – 2 Q

2

2 = 5 – 2 Q 2 + 0.15 Q 2

93.75 = 0.15 Q 2

2

625 = Q 2

2

Q 2 = 25

One of the conditions of the long-run equilibrium is that *AR* 2 or P 2 must be equal to *LAC*. Whether this condition holds can be checked as follows.

P 2 = *AR* 2 = *LAC*

Price functions (P 2) is given in Eq. (16.8) and *LAC* function in Eq. (16.5). By equating these functions, we get

98.75 – Q

1562.5

2 =

+ 5 – Q + 0.05 Q 2

Q 2

By substitution, we get

98.75 – 25 = 1562.5 + 5 – 25 + 0.05 (25)2

25

73.75 = 62.50 – 20 + 31.25 = 73.75

It is thus mathematically proved that in the long-run, firm's $P = AR = LAC$ and it earns only a normal profit.

16.5 NON-PRICE COMPETITION: SELLING COST AND GROUP EQUILIBRIUM

In the preceding section, we have presented Chamberlin's analysis of price competition and its effect on the firm's equilibrium output and profits under monopolistic competition. Chamberlin's analysis shows that price competition results in the loss of monopoly profits. All firms are losers: there are no gainers. Therefore, firms find other ways and means to ***non-price competition*** for enlarging their market share and profits. The two most common forms of non-price competition are ***product innovation*** and ***advertisement***. Product innovation and advertisement go on simultaneously. In fact, the successful introduction of a new product depends on its effective advertisement. Apart from advertisement expenses, under monopolistic competition firms incur other kinds of costs on competitive promotion of their sales, e.g., expenses on sales personnel, allowance to dealers, discounts to customers, expenses on displays, gifts and free samples to customers, additional costs on attractive packaging of goods, etc. All such expenses plus advertisement expenditure constitute firm's ***selling cost.***

Incurring selling cost increases sales, but at different rates, at different stages. Generally, sales increase initially at increasing rates, but eventually at decreasing rates. Consequently, the average cost of selling (*ASC*) initially decreases but ultimately it increases. The *ASC* curve is, therefore, *U*-shaped, similar to the conventional *AC* curve. This implies that total sales are subject to diminishing returns to increasing selling costs. Non-price competition through selling cost leads all the firms to an almost similar equilibrium. Chamberlin calls it "Group Equilibrium." We discuss here Chamberlin analysis of firm's group equilibrium.

Selling Cost and Group Equilibrium

To analyze group equilibrium of firms with selling costs, let us recall that the main objective of all firms is to maximize their total profit. When they incure selling costs, they do so with the same objective. Group equilibrium is analysed under the assumption cost and revenue curves remain the same. The analysis of group equilibrium is presented in Fig. 16.3. The curve *APC* represents firms' average

Fig. 16.3 *Selling Costs and Group Equilibrium*

production cost and competitive price is given at *OP* 3. None of the firms incurs any selling cost. And, as shown in Fig. 16.2, all the firms are 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 of the curve *APC* + *ASC* 1 and its total sale increases to *OQ* 4—the level of output at firm's *APC* is minimum. At output *OQ* 4, the firm makes supernormal profits of *P* 3 *PMP* 2. This abnormal profit is, however, earned only so long as other firms do not incur selling cost on their products. If other firms do advertise their products competitively and incure the same amount of selling cost, the initial advantage to the firm advertising first disappears and its output falls to OQ_2. In fact, all the firms reach equilibrium at point *A* and produce OQ_2 units. But their short-sightedness compels them to increase their selling cost further 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* rises to $APC + ASC_2$ which is tangent to the *AR* = *MR* line. The ultimate equilibrium point is shown by point *B*. Beyond point *B*, advertising is of no avail to any firm. Therefore, the equilibrium will be stable at point *B* where each firm produces OQ_3 and makes only normal profit.

16.6 CRITICAL APPRAISAL OF CHAMBERLIN'S THEORY

Chamberlin's theory of monopolistic competition propounded in the early 1930s is still regarded to be a major contribution to the theory of pricing. In fact, there is no better theoretical explanation of price determination under monopolistic competition. However, his theory has been criticized on both theoretical and empirical grounds. Let us now look into its theoretical weaknesses and empirical relevance.

First, Chamberlin assumes that monopolistic competitors act independently and their price maneouvering goes unnoticed by the rival firms. This assumption has been questioned on the ground that firms are bound to be affected by decisions of the rival firms since their products are close substitutes for one another and, therefore, they are bound to react.

Secondly, 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 decrease in their profits. Such an assumption can hardly be accepted.

Thirdly, Chamberlin's concept of industry as a 'product group' is ambiguous. It is also incompatible with product differentiation. In fact, each firm is an industry by virtue of its specialized and unique product.

Fourthly, his 'heroic assumptions' of identical cost and revenue curves are questionable. In the opinion of the critiques, each firm is an industry in itself, there is a greater possibility of variations in the costs and revenue conditions of the various firms.

Fifthly, 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.

Finally, so far as empirical validity of Chamberlin's concept of monopolistic competition is concerned, *it is difficult to find any example in the real world to which*

his model of monopolistic competition is relevant.[12] Most markets that exist in the real world may be classified under perfect or pure competition, oligopoly or monopoly.[13] It is, therefore, alleged that Chamberlin's model of monopolistic competition analyzes 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.[14]

Despite the above criticism, Chamberlin's contribution to the theory of price cannot be denied. Chamberlin was the first to introduce the concept of *differentiated product* and *selling costs* as a decision variable and to offer a systematic analysis 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 analyzing behaviour of firms, which later became the basis of the *kinked-demand* curve analysis.

SUMMARY

- Monopolistic competition refers to a market structure in which there are a large number of sellers of differentiated products which are close substitute for one another.
- The attempt to construct the theory of price determination under monopolistic competition was made in early 1930s. The reason was that price theories related to perfect competition and monopoly had become irrelevant because of drastic change in market conditions.
- The first attempt was made by Joan Robinson and by Chamberlin, both in 1933. Chamberlin's theory, though not relevant any more, is regarded theoretically sound and more realistic. Therefore, his theory constitutes a part of theory of price determination.

• Characteristics of monopolistic competition include (*i*) there is large number of sellers, (*ii*) product is differentiated, (*iii*) firms can enter and exit the industry, (*iv*) firms face downward sloping demand curve, and (*v*) firms incur selling cost.

• In short run, the system of price determination under monopolistic competition is exactly the same as under monopoly in short run. Given their objective of profit maximization, firms find their equilibrium at the output at which their *MR* = *MC*. Given their demand curve, price is determined accordingly and firms make abnormal profit.

• In the long run, the abnormal profit attracts new firms to the industry. With increase in the number firms, and share the market of existing firms. As a result, market becomes more competitive and firms make only normal profit.

• Chamberlin has pointed out another factor under monopolistic competition, i.e., firms go for incurring selling cost to retain and to promote their sales. According to Chamberlin, firms make ultimately only normal profit even with selling cost.

• Economists have pointed out several deficiencies in Chamberlin's theory of monopolistic competition. Nevertheless, Chamberlin's theory is regarded as an important contribution to pricing theory.

12. Cohen, K.J., and R.M. Cyert, *Theory of the Firm*, (New Delhi, Prentice-Hall of India, 1976), p. 230.

13. *Ibid*. , pp. 229–30.

14. *Ibid*.

REVIEW QUESTIONS AND EXERCISES

1. What is monopolistic competition? How does it differ from perfect competition and monopoly?
2. Monopolistic competition is the middle ground between perfect competition and monopoly; Explain the statement.
3. Theories of monopolistic competition and monopoly do not represent the real business world; whereas theory of monopolistic competition offers an explanation to price and output determination in a real business world. Explain and justify the statement.
4. What are the characteristics of monopolistic competition? Compare the characteristics of monopolistic competition with those of perfect competition?
5. There is an element of monopoly in monopolistic competition. What factors give monopoly; power to a firm in monopolistic competition?
6. What is meant by product differentiation? What is the purpose of product differentiation? How does it affect firm's demand curve?
7. In monopolistic competition, firms think that the demand curve for their individual product is different from that of the industry as whole. Illustrate graphically firm's perception of their individual demand curve.
8. What is meant by selling costs? How is selling cost different from advertisement cost? Why do firms in monopolistic competition incur selling costs?
9. According to Chamberlin, average selling cost (ASC) curve is U-shaped. What factors determine the shape of the ASC? Illustrate graphically how selling costs affect the overall cost structure of the firm.
10. The purpose of selling cost is to increase the demand for the product and to make demand curve more elastic. But there is limit to it. How can a firm find the optimum level of selling cost? Explain by using appropriate diagrams.
11. What is Chamberlin's concept of'product group'? How is the concept of 'product group' different from the traditional meaning of industry?
12. How does a monopolistically competitive firm adjust its price and output to arrive at its equilibrium? Explain and illustrate how a firm in monopolistic competition reaches its equilibrium in the short run? Does a firm in equilibrium in monopolistic competition always make a supernormal profit?
13. Suppose market demand curve is given Q = 500 – 0.5P. Work out the proportional demand curve for a firm assuming there are 50 firms in the industry. Is the elasticity of the proportional demand curve at a given price is the same as that of the market demand curve or different?
14. How are the long-run conditions different from short-run conditions for a firm in monopolistic competition? Illustrate and explain the long-run equilibrium of firm in monopolistic competition? How is firm's long-run equilibrium different from its short-run equilibrium?
15. What is meant by 'ideal output' and 'excess capacity'? Illustrate Chamberlin's measure of excess capacity. How does Harrod's measure of excess capacity differ from Chamberlin's measure?

16. What are the wastages of monopolistic competition? Is the 'traditional' concept of excess capacity an overstatement of one of the wastages of monopolistic competition? Give reasons in support of your answer.

17. What is the usual form of non-price competition in monopolistic competition? Explain and illustrate the firm's long-run equilibrium under non-price competition. Why do firms involved in non-price competition make only normal profit in the long run?

18. What does Chamberlin mean by 'group equilibrium'? How does selling cost affect the group equilibrium? Illustrate firm's equilibrium with selling costs.

19. Group equilibrium shows that firms incurring selling costs make only normal profit in the long run. It implies that selling cost is a waste. Do you agree with this statement? Give reason for your answer.

20. What are the drawbacks of Chamberlin's theory of monopolistic competition? What is the merit of his theory in spite of its drawbacks?

21. Write a note on the critical evaluation of Chamberlin's theory of monopolistic competition? Why is his theory not in use even though it is analytical sound?

FURTHER READING

Bain, J.S. (1971), 'Chamberlin's Impact on Microeconomic Theory', in R.E. Kuenne (ed), *Monopolistic Competition Theory: Studies in Impact—Essays in Honour of E.H. Chamberlin* (NY: Wiley Eastern, 1986), reprinted in H. Townsend (ed.), *Readings in Price Theory* (Middlesex, England: Penguin).

Browning, E.K. and Browning, J.M. (1994), *Microeconomic Theory and Applications* (New Delhi: Kalyani Publishers), 2nd Edn., Chapter 13.

ChamBerlin, E.H. (1956), *The Theory of Monopolistic Competition* (Cambridge, MA: Harvard University Press), 7th Edn.

Cohen, K.J. and Cyert, R.M. (1976), *Theory of the Firm* (New Delhi: Printice-Hall of India), Chapter 11.

Dempetz, H. (1959), 'The Nature of Equilibrium in Monopolistic Competition', *Jl. Pol. Eco.,* November.

goulD, J.P. and Lazear, E.P. (1993), *Microeconomic Theory* (Homewood, IL: Richard D. Irwin), 6th Edn., Chapter 12.

Koutsoyiannis, A. (1979), *Modem Microeconomics* (London: Macmillan), 2nd Edn., Chapter 8.

maDDala, G.S. and Miller, E. (1989), *Microeconomics: Theory and Applications* (New York, NY: McGraw-Hill Book Co.), Chapter 13.

roBinson, J. (1953), *The Economics of Imperfect Competition,* (Macmillan, 1953) and 'Imperfect Competition Revised,' *Eco. Jl,* (1953).

shrafta, P. (1926), 'The Laws of Returns under Competitive Conditions,' *Eco. J.* Reprinted in *Readings in Price Theory,* Am. Eco. Assn, 1944.

stigler, *G.J.* (1949), 'Monopolistic Competition in Retrospect', *Five Lectures on Economic Problems* (Longmans: Green).

triffin, R. (1939), *Monopolistic Competition and General Equilibrium* (Cambridge, MA: Harvard University Press).

CHAPTER

17 Price and Output Determination under Oligopoly

CHAPTER OBJECTIVES

The objective of this chapter is to discuss the various theories formulated to analyse the determination of price and output under oligopoly. The contents of this chapter included the following aspects.

- The meaning and sources of oligopoly
- Features of oligopolistic market
- An overview of theoretical models of oligopoly
- Important theoretical models of oligopoly, viz., Cournot model, kinked-demand curve model, and Chamberlin's model and
- Collusive models of duopoly – cartel and price leadership models

17.1 INTRODUCTION

Market conditions have been changing over time. Following the changes in market conditions, economists developed new theories of price determination. As noted in earlier chapters, theories of perfect competition and monopoly turned irrelevant due to change in market conditions – emergence of monopolistic competition. Market condition continued to change and monopolistic competition almost disappeared due to further change in market conditions and emergence of oligopolistic structure of market. Now economists have constructed new models and developed new theories of price and output determination under oligopoly. It may be added here that theories of oligopoly were developed by the classical economists, especially Augustin Cournot, F. Y. Edgeworth and J. Bertrand. Later on, economists developed many other theories of price determination under oligopoly. This Chapter presents a detailed discussion on the major theories of price and output determination under oligopoly.

17.2 OLIGOPOLY MARKET: DEFINITION AND SOURCES

17.2.1 Definition of Oligopoly[1]

Oligopoly is defined as a market structure in which there are *a few sellers* selling *homogeneous* or *differentiated* products. In case oligopoly firms sell a homogeneous product,

1. The term 'oligopoly' has been formed by two Greek words *Oligi* meaning 'a few' and *polan* meaning sellers'. Thus, *oligopoly* means a market of few sellers.

it is called *pure* or *homogeneous oligopoly*. For example, industries producing bread, cement, steel, petrol, cooking gas, chemicals, aluminium and sugar are industries characterized by *homogeneous* oligopoly. And, where firms of an oligopoly industry sell *differentiated products*, it is called *differentiated* or *heterogeneous oligopoly*. Automobiles, television sets, soaps and detergents, refrigerators, soft drinks, computers, cigarettes, etc. are some examples of industries characterized by *differentiated* or *heterogeneous oligopoly*.

Be it pure or differentiated, "Oligopoly is the most prevalent form of market organization in the manufacturing sector of the industrial nations..."[2]. In non-industrial nations like India also, a majority of big and small industries have acquired the features of oligopoly market. The market share of 4 to 10 firms in 84 big and small industries[3] of India is given in Table 17.1.

Table 17.1 *Distribution of Industries by Market Share*

Market share (%)	*No. of industries*
1 – 24.9	8
25 – 49.9	11
50 – 74.9	15
75 – 100	50
Total	84

As the data presented above shows, in India, in 50 out of 84 selected industries, i.e., in about 60 per cent industries, 4 to 10 firms have a 75 per cent or more market share which gives a *concentration ratio* [4] of 0.500 or above. All such industries can be classified under oligopoly.

17.2.2 Sources of Oligopoly

The factors that give rise to oligopoly are broadly the same as those for monopoly. The main sources of oligopoly are described here briefly.

1. *Huge capital investment.* Some industries are by nature capital-intensive, e.g., manufacturing automobiles, aircraft, ships, TV sets, computers, mobile phones, refrigerators, steel and aluminium goods, etc. Such industries require huge initial investment. Therefore, only those firms which can make huge investment can enter these kinds of industries and survive in the long run. In fact, a huge investment requirement works as a natural barrier to entry to the oligopolistic industries.

2. *Economies of scale.* By virtue of huge investment and large scale production, the large units enjoy *absolute cost advantage* due to economies of scale in production, purchase of industrial inputs, market financing, and sales organization. This gives the existing firms a comparative advantage over new firms in production and price competition. This also works as a deterrent for the entry of new firms.

3. *Patent rights to branded product.* In case of *differentiated oligopoly,* firms get their differentiated product patented which gives them an exclusive right to produce and market the patented commodity. This prevents other firms from producing the patented

2. Salvatore, D., *Managerial Economics*, (N.Y., McGraw-Hill, 1989), p. 475.

3. On the basis of data published by the CMIE in August 1999 issue of its *Industries and Market Share*.

4. The 'concentration ratio' is the measure of degree by which a small number of firms dominate the industry. It is the percentage share of dominant (4 to 12) firms in the total sales of the industry. The US Census of manufacturing uses 4, 8 and 12 firms for working out the concentration ratio.

commodity. Therefore, unless new firms have something new to offer and can match the existing products in respect of quality and cost, they cannot enter the industry. This keeps the number of firms limited.

4. *Control of some firms over certain raw materials.* Where a few firms acquire control over almost the entire supply of important inputs required to produce a certain commodity, new firms find it extremely difficult to enter the industry. For example, if a few firms acquire the right from the government to import certain raw materials, they have control over the entire input supply.

5. *Merger and acquisition.* Merger of rival firms or takeover of rival firms by the bigger ones with a view to protecting their jo int market share or to put an end to waste of competition has emerged, in modern times as an important factor that gives rise to oligopolies and strengthens the oligopolistic tendency in modern industries. Mergers and takeovers have been one of the main features of recent trend in Indian industries.

17.3 FEATURES OF OLIGOPOLY

Like other kinds of markets, oligopoly has its own characteristic features. Let us now look at the important characteristics of oligopolistic industries.

1. *Small number of sellers.* As already mentioned, there is a small number of *sellers* under oligopoly. How small is the number of sellers in oligopoly markets is difficult to specify precisely for it depends largely on the size of the market. Conceptually, however, the number of sellers is so small that the market share of each firm is large enough for a single firm to influence the market price and the business strategy of its rival firms. The number may vary from industry to industry. Some examples of oligopoly industries in India and market share of the *dominant* firms[5] in 1997–98 is given in Table 17.2.

Table 17.2 *Selected Industries and Their Market Share*

Industry	*No. of firms*	*Total market share* (%)
Ice-cream	4	100.00
Bread	2	100.00
Infant Milk food	6	99.95
Motorcycles	5	99.95
Passenger cars	5	94.34
Cigarettes	4	99.90
Fruit Juice, pulp & conc.	10	98.21
Fluorescent lamps	3	91.84
Automobile tyres		

8

91.37

Source: CMIE, *Industries and Market Share*, August 1999.

2. ***Interdependence of decision-making.*** The *most striking feature* of an oligopolistic market structure is the interdependence of oligopoly firms in their decision-making. The characteristic fewness of firms under oligopoly brings the firms in keen competition with each other. The competition between the firms takes the form of action, reaction and

5. Market share of individual firms vary to a great extent. For example, in 1997-98, Hindustan Lever had a share of 74% of the ice-cream market; Surya Roshni had 61% share in fluorescent lamp market; MUL had 76.1% market share in passenger cars; and *ITC* had 75.38% market share in cigarettes.

counter-action in the absence of collusion between the firms. For example, it has been generally observed that car companies change their prices following the change in price made by one of the companies. They have introduced new model in competition with one another. Since the number of firms in the industry is small, the business strategy of each firm in respect of pricing, advertising and product modification is closely watched by the rival firms and it evokes imitation and retaliation. What is equally important is that firms initiating a new business strategy anticipate and take into account the possible counter-action by the rival firms. This is called interdependence of oligopoly firms.

An illuminating example of strategic manoeuvering is cited by Robert A. Meyer.6

To quote the example, one of the US car manufacturing companies announced in one year in the month of September7 an *increase* of $180 in the price list of its car model. Following it, a few days later a second company announced an increase of $80 only and a third announced an increase of $91. The first company made a counter move and it announced a *reduction* in its enhanced price from $180 to $71. This is a pertinent example of interdependence of firms in business decisions under oligopolistic market structure. In India, when Maruti Udyog Limited (MUL), announced a price cut of `24,000 tò36,000 in early 2005 on its passenger cars, other companies followed suit. However, *price competition* is not the major form of competition among the oligopoly firms as price war destroys the profits. A more common form of competition is *non-price competition* on the basis of product differentiation, vigorous advertising and provision of survive.

3. ***Barriers to entry of new firms.*** Barriers to entry to an oligopolistic industry arise due to such market conditions as (*i*) huge investment requirement to match the production capacity of the existing ones, (*ii*) economies of scale and absolute cost advantage enjoyed by the existing firms, (*iii*) strong consumer loyalty to the products of the established firms based on their quality and service, and (*iv*) preventing entry of new firms by the established firms through price cutting. However, the new entrants that can cross these barriers can and do enter the industry, though only a few, that too mostly the branches of *MNCs*.

4. ***Indeterminate price and output.*** Another important feature, though a controversial one, of the oligopolistic market structure is the indeterminateness of price and output. The characteristic fewness and interdependence of oligopoly firms makes derivation of their demand curve a difficult proposition. Therefore, price and output are said to be indeterminate. However, price and output are said to be determinate under collusive oligopoly. But, there too, collusion may last or it may break down. *An opposite view is that price under oligopoly is sticky*, i.e., if price is once determined, it tends to stabilize.

17.4 OLIGOPOLY MODELS: AN OVERVIEW

As already mentioned, under oligopolistic conditions, rival firms indulge in an intricate pattern of actions, reactions and counter-actions showing a variety of behavioural patterns. As Baumol puts it, "Under [these] circumstances, a very wide variety of behaviour pattern becomes possible. Rivals may decide to get together and cooperate in the pursuit of their objectives,... or, at the other extreme, may try to fight each other to the death. Even if they enter an agreement, it may last or it may break down."8 The economists have, therefore,

6. *Microeconomic Decisions*, (Houghton Miflin Company, Boston, 1976), p. 249.

7. The month in which automobile manufacturers introduce new models.

8. Baumol, W.J., *Economic Theory and Operations Analysis*, (New Delhi, Prentice Hall of India, 4th edn., 1985), p. 410.

found it extremely difficult to make a systematic analysis of price and output determination under oligopoly.9 This has, however, not deterred the economists from their efforts to build theoretical model to analyse the behaviour of oligopoly firms.

In accordance with the wide variety of behavioural patterns, the economists have developed a variety of analytical models based on different behavioural assumptions. The widely known oligopoly models include Cournot's duoply model (1838), Bertrand's leadership model (1880), Edgeworth's duoply model (1897), Stackelberg's model (1933), Sweezy's kinked demand curve model (1939), Neumann and Margenstern Game Theory model (1944) and Baumol's sales maximization model (1959). None of these models, however, provides a universally acceptable analysis of oligopoly, though these models do provide an insight into oligopolistic behaviour.

In this chapter, we discuss some selected oligopoly models with the purpose of showing the behaviour of oligopoly firms and working of the oligopolistic markets. The analytical models discussed here are selected on the basis of how price and output are determined under price competition among oligopoly firms and cartel system. Specifically, we discuss here the following oligopoly models.

(*i*) Cournot's duopoly model,

(*ii*) Sweezy's kinked demand curve model,

(*iii*) Price leadership models:

(*a*) Price leadership by low-cost firm,

(*b*) Price leadership by dominant firm and

(*c*) Price leadership by barometric firm,

(*iv*) Collusive model: The Cartel Arrangement.

The Game Theory model of oligopoly, Prisoner's Dilemma and Baumol's sales revenue maximization model, which merits a detailed discussion, will be discussed in the next chapter.

17.5 COURNOT'S MODEL OF OLIGOPOLY

Augustin Cournot,10 a French economist, was the first to develop a formal oligopoly model in 1838 in the form of a duopoly model. Cournot developed his model with the example of two firms, each owning a well of mineral water and water being produced at zero cost. To illustrate his model, Cournot made the following assumptions.

(*a*) There are two firms, each owning an artesian mineral water well;

(*b*) Both the firms operate their wells at zero cost;11

(*c*) Both of them face a demand curve with constant negative slope;

(*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.

9. Baumol, W.J., *op. cit.* , p. 410.

10. Cournot, Augustin, *Research into the Mathematical Principles of the Theory of Wealth*, Translation by Nathaniel T. Bacon, (New York, Macmillan, 1897).

11. Under zero cost condition, the total revenue is the same as the total profit.

On the basis of this model, Cournot has concluded that each seller ultimately supplies one-third of the market and both the firms charge the same price. And, one-third of the market remains unsupplied.

Cournot's duopoly model is illustrated in Fig. 17.1. The demand curve for mineral water is given by the *AR* curve and their *MR* by the *MR* curve. To begin with, let us suppose that firm *A* is the only seller of mineral water in the market. By assumption, its *MC* = 0. Following the profit maximizing rule, it sells quantity *OQ* where its *MC* = 0 = *MR*, at price *OP* 2. Its total profit is *OP*

Fig. 17.1 *Price and Output Determination*

2 *PQ*. This is the maximum profit seller *A* can make given the demand curve.

under Duopoly: Cournot's Model

Now let another firm *B* enter the market. The market open to *B* is the market unsupplied by A. This market equals QM which is *half* of the total market.12 That is, B can sell its product in the remaining half of the market. B assumes that A will not change its price and output because A is making maximum profit. In other words, B assumes that A will continue to sell OQ at prices OP 2. Thus, the market available to firm B is QM and the relevant part of the demand curve is PM. The MR curve corresponding to B's demand curve PM is given by PN. B's MR curve, PN, it bisects QM at point N where $QN = NM$. In order to maximize its revenue, B sells QN at price OP 1 = $P'N$. Its total revenue is maximum at $QRP'N$ which equals its total profit. Note that B supplies only $QN = (1/2)/2 = 1/4$ of the market.

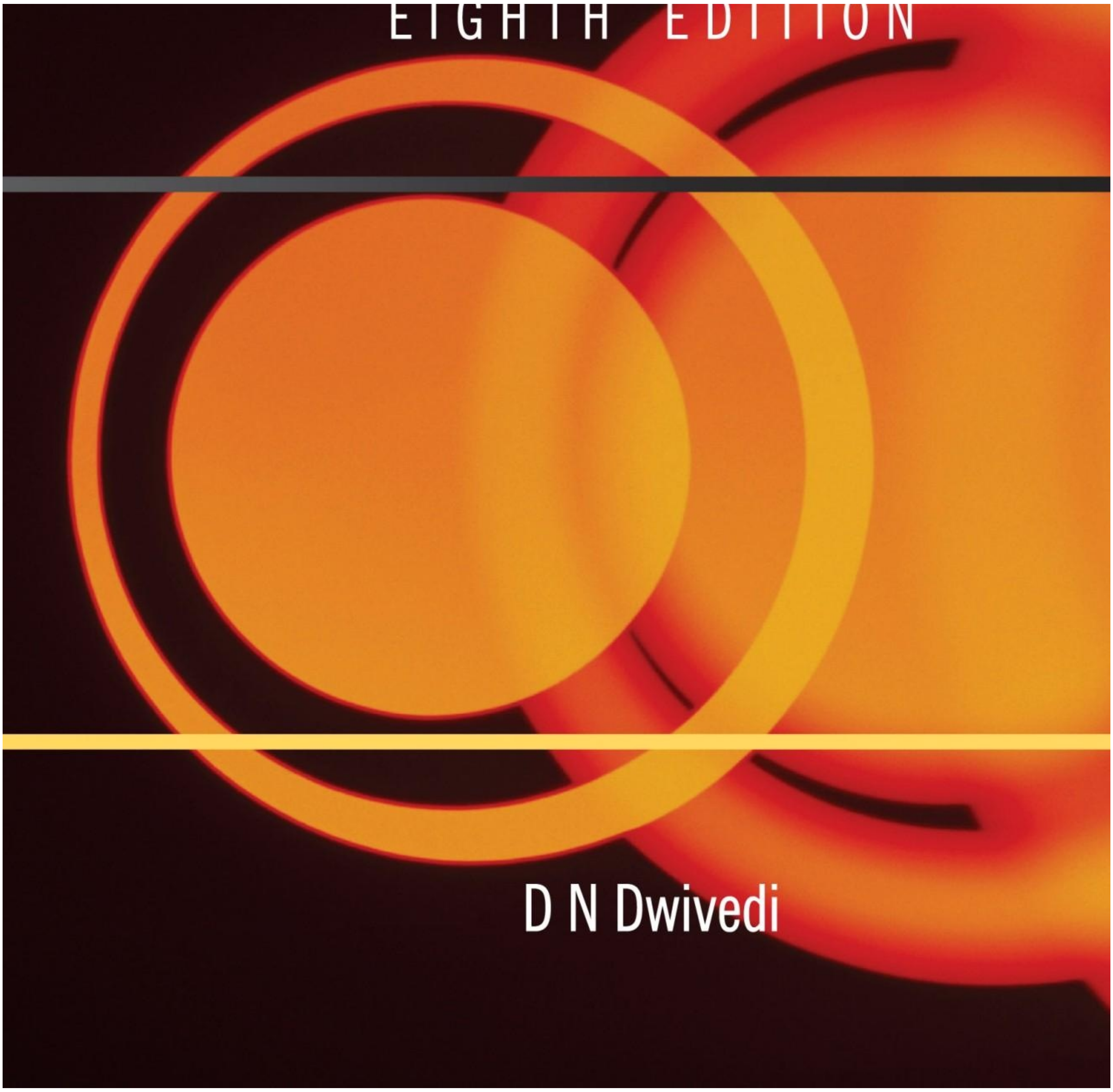
With the entry of firm B, price falls to OP 1. Price falls becuase A's customers will switch over to firm B. Therefore, firm A is forced to reduce its price. Faced with this situation, firm A adjusts its price and output to the changed conditions. Firm A assumes that firm B will not change its output QN and price OP 1 as it is making maximum profit. Accordingly, firm A assumes that firm B will continue to supply $QN = 1/4$ of the market. Thus, firm A assumes that it has 3/4 (= 1 – 1/4) of the market available to it. To maximize its profit, firm A supplies 1/2 of the remaining 3/4 of the market, i.e., $1/2 \times 3/4 = 3/8$ of the market. It is noteworthy that A's market share has fallen from 1/2 to 3/8.

Now it is firm B's turn to react. Following Cournot's assumption, firm B assumes that firm A will continue to supply only 3/8 of the market and the rest of the market is open to him, which equals $1 – 3/8 = 5/8$. To maximize his profit under the new conditions, firm B supplies half of the remaining market, i.e., firm firm B supplies $1/2 \times 5/8 = 5/16$ of the market.

It is now for firm 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, firm A continues to lose his market share and firm B continues to gain. Eventually, a situation is reached when their market shares equal 1/3 each. Any further attempt to adjust output produces the same result. The firms, therefore, reach their equilibrium position where each one supplies one-third of the market and both charge the same price and one-third of the market remains unsupplied.

12. Note that where $MR = 0$, price elasticity, $e = 1$, i.e., $PM/PD = 1 = QM/OQ$. This means, $PM = PD$ and $QM = OQ$.



VIKAS®

The actions and reactions and equilibrium of the firms *A* and *B*, according to Cournot's model, are presented in Table 17.3.

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 beyond one-third of the market as shown in the last row of the table.

Table 17.3 *Determination of Market Share*

Period

Frim A's Market share

Firm B's Market share

I

1

1

(1) =

1 ⎛ 1 ⎞ 1

| |=
2
2
2 ⎝ 2 ⎠ 4
1 ⎛
1 ⎞ 3
1 ⎛
3 ⎞
5
II
| 1− |=
| 1− |=
2 ⎝
4 ⎠
8
2 ⎝
8 ⎠ 16
III
1 ⎛
5 ⎞ 11
| 1− |=
1 ⎛
11 ⎞ 21
2 ⎝
16 ⎠ 32
| 1−
|=
2 ⎝
32 ⎠ 64
1 ⎛
43 ⎞
85
IV
1 ⎛
21 ⎞
43
| 1−
|=
| 1−
|=
2 ⎝
64 ⎠ 128
2 ⎝
128 ⎠ 256
1 ⎛
1 ⎞ 1
1 ⎛
1 ⎞ 1
N
| 1− |=
| 1− |=
2 ⎝
3 ⎠ 3
2 ⎝
3 ⎠ 3

Note: In the calculation of market share, number 1 represents the total market.

Duopoly Model Extended to Oligopoly

Cournot's model of duopoly can be extended to a general oligopoly model. For example, if there are three sellers in the industry, each one of them will be in equilibrium when each firm supplies 1/4 of the market. The three sellers together supply 3/4 of the total market, 1/4 of the market remaining unsupplied. Similarly, when there are four firms each one of them supply 1/5th of the market and 1/5th of the market remains 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.

Mathematical Treatment of Cournot's Model

Cournot's model may also be presented mathematically. Let us suppose that the market demand function for mineral water is given by a linear function as

$Q = 90 - P$

…(17.1)

As noted above, under zero cost condition, profit is maximum where $MC = MR = 0$

and when $MR = 0$, the profit maximizing output is 1/2 (Q).

Thus, when seller A is a monopolist in the market, his profit-maximizing output (QA), according to the profit maximizing rule under zero cost condition, is determined at half of the total market. That is, A's initial market share

$QA = 1/2\ (90 - P)$

…(17.2)

When another seller, B, enters the market, his profit maximizing output

$QB = 1/2\ [(1/2(90 - P)]$

…(17.3)

Thus, the respective shares of sellers A and B are fixed at QA and QB.

The division of market output may be expressed as

$Q = QA + QB = 90 - P$

…(17.4)

The demand function for A may now be expressed as

$QA = (90 - QB) - P$

…(17.5)

and for B as

$QB = (90 - QA) - P$

…(17.6)

Given the demand function (17.5), the market open to A (at $P = 0$) is $90 - QB$. The profit maximizing output for A will be half of the market size, i.e.,

Q

$- QB$

$A = 90$

…(17.7)

2

and similarly for B, it will be

$90 - Q$

Q

A

$B =$

…(17.8)

2

The Eqs. (17.7) and (17.8) represent the reaction functions of sellers A and B, respectively. For example, consider Eq. (17.7). The profit maximizing output of A depends on the value of QB, i.e., the output which B is assumed to produce. If B chooses to produce 30 units, (i.e., $QB = 30$), then A's output = [(90 – 30)/2] = 30. If B chooses to produce 60 units, A's output = (90 – 60)/2 = 15. Thus, Eq. (17.7) gives A's reaction function. Similarly Eq. (17.9) gives B's reaction function.

The reaction functions of A and B are graphed in Fig. 17.2. The reaction function shown by line AM shows how A will react on the assumption that B will not react to changes in his output once *B*'s output

is fixed. The reaction function *BD* shows

a similar reaction of *B*. The two reaction

functions intersect at point *E*. It means

that the assumptions of *A* and *B* coincide

at point *E* and here ends their action and

Fig. 17.2

reaction. Point *E* is, therefore, the point of

Reaction Function and

Equilibrium: Cournot Model

stable equilibrium. At this point, each seller

sells only 30 units. The same result can be

obtained by equating the two reaction Eqs. (17.7) and (17.8).

The market equilibrium takes place where

$= 90 - Q$

$- Q$

B

90

A

=

2

2

Since, QB = (90 – QA)/2, by substitution, we get first term as

$90 - (90 - QA)/2$

$QA =$

2

$2\ QA = 90 - (90 - QA)/2$

$2\ QA = 180 - 90 - QA$

$QA = 30$

Similarly, it can be shown that QB = 30. Thus, both the sellers are in equilibrium. At equilibrium, both the sellers produce and sell 30 units each. The market output will be 60 units. Given the market demand curve, market price will be $P = 90 - Q = 90 - 60 =$ `30.

Criticism Although Cournot's model yields a stable equilibrium, it has been criticized on the following grounds.

First, Cournot's behavioural assumption [assumption (*d*) above] is naïve as it implies that firms continue to make wrong assumption about the competitor's behaviour. In other words, Cournot's assumption that each seller continues to assume that his rival will not change his output even though he observes time and again that his rival firm does change its output, is not practically current.

Second, his assumption of zero cost of production is unrealistic though dropping this assumption does not alter his position.

17.6 SWEEZY'S MODEL OF OLIGOPOLY: KINKED-DEMAND CURVE MODEL

The concept and form of the kinked-demand curve was first developed and used by Chamberlin in his theory of monopolistic competition.[13] Later, Hall and Hitch[14] used kinked-demand curve to explain rigidity of prices in oligopolistic market. But, neither Chamberlin nor Hall and Hitch used kinked demand curve as a tool of analysis in their respective theories. Paul M. Sweezy[15] used the kinked-demand curve to establish that price remains stable in oligopolistic market. Sweezy's model is described below.

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 the behaviour of the oligopolistic firms. It must, however, be noted at the outset that *kinked-demand curve analysis does not deal with price and output determination*. It only seeks to establish that once a price-quantity combination is determined, an oligopoly firm does not find it profitable to change its price even when there is a considerable change in the cost of production and change in demand for the product.

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 go for price cutting, so that the price-raising firm loses a part of its market to the rival firms. This behaviour is true of all the firms. The oligopoly firms would, therefore, find it more desirable to maintain the prevailing price and output. Let us now see how Sweezy has proved this point of view by using kinked demand curve technique.

13. Chamberlin, Edward Hasting, *Theory of Monopolistic Competition,* (Cambridge, MA., Harvard University Press, 1933).

14. Hall, R.L. and Hitch, C.I., "Price Theory and Business Behaviour," *Oxford University Papers*, 1939, pp 12–45.

15. Sweezy P.M., "Demand under Conditions of Oligopoly," 1939, pp. 568–573.

Sweezy's Kinked-Demand Curve Model In order to analyse the effects of possible reactions of the rival firms on the demand for the product of the firm initiating the change in price, let us make the following assumptions.

1. There are four oligopoly firms—A, *B, C* and *D;*
2. Market demand is divided between the firms and all the firms face a uniform demand curve;
3. All the firms are in equilibrium at a point of time, all maximizing their profit.

Sweezy's kinked-demand curve model is presented in Fig. 17.3 on the basis of these assumption. Let us suppose that the individual demand curve is given by the curve *MN* and the price is initially fixed at *PQ*. Given the price at *PQ*, let firm *A* take the lead in changing its price. Let us now examine the effect of various kinds of possible reactions of the rival firms on demand for A's product.

Fig. 17.3 *Kinked-Demand curve analysis*

• **Reaction (** ***i*****).** When firm *A* increases or decreases its price, the rival firms follow the suit. Then, firm *A* finds itself moving along the demand curve *MN*. It does not gain or lose.

• **Reaction (** ***ii*****)**. If rival firms do not react to price changes made by the firm *A*, its demand curve will be *DD′*. To explain it further, when firm A raises its price and rival firms do not follow, firm A loses part of its market to the rival firms and moves along the *PD* part of the demand curve. And, when firm *A* cuts its price down and rival firms do not follow, then it captures a part of the rival's market share and finds itself moving along the *PD′* part of the demand curve. This is what firm *A* would like to achieve.

• **Reaction (** ***iii*****).** When firm *A* raises its price and rival firms do not follow, then firm *A* loses a part of its market share to the rival linns. Then the relevant demand curve for firm *A* is *DP.* But, when firm *A* decreases its price, rival firms react by cutting down their own prices by an equal amount of even more. This counter move by the rival firms prevents firm *A* from taking any advantage of price cut.

Therefore, firm *A* is forced to move along the *PN* part of the demand curve.

Thus, the two relevant segments of demand curve for firm *A* (below point *P)* is *PN* and *PD* beyond point *P*. When the two relevant segments of the two demand curves *(MN* and *DD′)* are put together, the demand curve for *A*′ s product takes the form of the curve *DPN.* The demand curve *DPN* has a *kink* at point *P*. It is, therefore, called *a kinked-demand curve*.

Consider now the relationship between *AR* and *MR,* We know that slope of the *AR* curve is twice the slope by the *MR* curve. The *MR* curve drawn on the basis of this relationship, will take a shape as shown of *DJKL* in Fig. 17.3. It is discontinuous between points *J* and *K,* at output *OQ*. The segment *DJ* of the *MR* curve corresponds to *DP* segment of the demand curve and *KL* segment of *MR* curve corresponds to *PN* segment of the demand curve.

Now let us see why price gets stabilized at *PQ* even when there is change in cost curves. Suppose *MC* curve of the firms is given as *MC* 1 which intersects *MR* at point *K*, Point *K* satisfies the necessary condition for profit maximization (*MR* = *MC*). Therefore, oligopoly firms are in equilibrium at output *OQ* and they are making maximum profit at price *PQ.* Now, if *MC* curve shifts upwards to *MCn* or to any level between points *J* and *K,* their profit would not be affected because profit maximization condition remains undistributed. Therefore, they have no motivation for increasing or decreasing their price.

It is always beneficial for them to stick to the price *PQ* and output *OQ*. Thus, both price and output are *stable*. This is how price gets stabilized under oligopoly. The oligopoly firms would go for changing their price and output only if *MC* rises beyond point *J* or falls below point *K* (in Fig. 17.3)

To conclude, the basic point that Sweezy's model establishes is the rigidity of price in oligopolistic market. As shown in Fig. 17.3, once price is determined in oligopoly, it remains rigidly fixed because profit maximizing firms' have no incentive to change price even if *MC* curve shifts upwards or downwards within the range between points *J* and *K*. Therefore, price remains constant over a period of time.

Algebraic Solution to Sweezy's Model

Let us suppose that usual individual demand function[16] (D_1) corresponding to demand curve *MN* and firm's own stipulated demand function (D_2) with price manipulation corresponding to *DD′* in Fig. 17.3 and individual total cost (*TC*) function are given as follows.

(*i*) $D_1 : Q_1 = 100 - 0.5P_1$ ...(17.10)

(*ii*) $D_2 : Q_2 = 160 - P_2$ ...(17.11)

(*iii*) $TC = 300 + 10Q + 0.5Q^2$ ...(17.12)

The demand functions (17.10) and (17.11) are shown by D_1 and D_2 curves in Fig. 17.4. Demand curves D_1 and D_2 intersect at point *K*. The kinked demand curve is drawn and marked by BKD_1. What we need now is to work out MR_1 and MR_2 corresponding to the two demand functions, D_1 and D_2 and *MC* from the cost function.

To work out MR_1 and MR_2, we need to find TR_1 and TR_2. Given the demand functions (17.10) and (17.11), P_1 and P_2 can be obtained as

$P_1 = 200 - 2Q_1$ ...(17.13)

16. All the firms in oligopoly market are supposed to face identical demand function.

and

$P_2 = 160 - Q_2$ ...(17.14)

The TR_1 and TR_2 functions can be worked out by using price functions (17.13) and (17.14) as follows.

$TR_1 = P_1 . Q_1$

$= (200 - 2Q_1)Q_1$

$= 200Q_1 - 2Q_1^2$ ...(17.15)

and

$TR_2 = P_2 . Q_2 = (160 - Q_2)Q_2$

$= 160Q_2 - Q_2^2$ ...(17.16)

By differentiating TR_1 and TR_2 functions (17.15) and (17.16), we can derive the MR_1 and MR_2 functions, respectively, as given below.

$MR_1 = 200 - 4Q_1$ ...(17.17)

and

$MR_2 = 160 - 2Q_2$ ...(17.18)

The MR_1 and MR_2 functions are shown by truncated lines MR_1 and MR_2 in Fig. 17.4. The *MR* curve corresponding to the kinked demand curve is drawn through points *BLM* and along the line MMR_1.

Fig. 17.4 *Sweezy's Kinked Demand Model*

As regards *MC* curve, it can be obtained by differentiating the *TC* function (17.12).

Thus,

2

∂TC $\partial(300 + 20Q + 0.5Q)$

$MC =$

=

∂Q

∂Q

$= 20 + Q$

…(17.19)

Having derived the MR_1, MR_2 and MC functions, we can now illustrate the conclusions of the kinked demand curve analysis. Let us first find price (P) and quantity demanded (Q) at kink point K. At the kink point (K), price P is determined by market conditions. And, at the point of intersection (K), demand indicated by D_1 and D_2 curves are equal at the given price. Let us assume that, given the price, $Q_1 = Q_2 = Q$. Now Q and P at the kink point K can be known as follows. Since at the point of intersection of D_1 and D_2 curves, $Q_1 = Q_2 = Q$, by substituting Q for Q_1 and Q_2 in price functions (17.13) and (17.14), we can work out quantity demanded at the point of kink by equating the price functions as follows.

$P_1 = P_2$

$200 – 2Q = 160 – Q$

$Q = 40$

By substituting 40 for Q in any of the price functions, given in Eq. (17.13) or Eq. (17.14), we can get price (P) at the point of intersection (P) of demand curves MN and DD' in Fig. 17.3. We know that at the point of intersection, $P_1 = P_2$. So when we get P_1 or P_2, we get P. Thus, by using price function (P_1), we get,

$P = 200 – 2Q_1$

$= 200 – 4(40)$

= `120

This price can be verified from Fig. 17.4.

Having worked out P and Q, let us now verify the main thesis of Sweezy's model that the variation in MC within a range will not affect the price. The lower limit of MC variation is given by point M at the MR_1 at price P and the upper limit by point L at MR_2 at the same price. Thus,

$MR_1 = 200 – 4Q_1 = 200 – 4(40)$ = `40

and

$MR_2 = 160 – 2Q_2 = 160 – 2(40)$ = `80

Thus, the lower and upper limits of MC variation that will not affect the price at $Q = 40$ lie between `40 and `80. At $Q = 40$, $MC = 10 + 40$ = `50. Now let the cost of production increase and cost function change

from

$TC = 300 + 10Q + 0.5Q^2$

to

$TC = 400 + 30Q + 0.5Q^2$

Then

$MC = 30 + Q$

…(17.20)

Given the MC function in Eq. (17.20), at $Q = 40$, $MC = 30 + 40$ = `70. Since $MC = 70$ is within the lower and upper range of variation in MC, price will not change. This proves Sweezy hypothesis that prices once determined, tend to be stable in the oligopoly market.

Criticism of Sweezy's Model

As mentioned earlier, Sweezy's model is considered to be the best known model that explains relatively more satisfactorily the behaviour of the firms in oligopoly. On the face it, it appears to be logically sound and realistic. However, economists have criticized Sweezy's model on both theoretical and empirical grounds, which are summarized below.

1. ***Sweezy's Model does not Explain Price Determination.*** The basic function of price theory is to explain price and output determination in a particular kind of market. Sweezys model, however, does not explain price and output determination.

His model only assumes the price to be given at a point to time. It explains only why price once determined tends to be sticky even if there are changes in cost conditions to a certain extent. Sweezy's model is, therefore, regarded as an *ex-post rationalization* rather than *ex-ante* explanation of market equilibrium.

2. ***This Model does not Determine the Point of Kink.*** This is a criticism related to non-determination of price. The kinked-demand curve analysis explains why 'kink' appears on the demand curve. "It does not explain how and at what level of price and output, the point of kink is determined"17. George Stigler doubts even the existence of the kinked demand curve. He had investigated seven oligopolistic industries in the US and found 'little evidence' of reluctance to price hike made by the firms. Stigler's view is supported by Julian Simon.18 The lack of empirical evidence makes Sweezy's model a purely hypothetical one, not as realistic as it appears on the face of it.

3. ***Price Rigidity is not Supported by Empirical Facts.*** Sweezy's claim of price rigidity in oligopoly does not stand the test of empirical verification. Empirical facts reveal a surprising lack of price stability in oligopoly markets. Empirically, monopoly prices have been found to be more stable than oligopoly prices. Economists' opinion is, however, divided on the issue of price rigidity in oligopoly. While Stigler Liebhafsky has questioned price rigidity in oligopoly market, finds considerable evidence of price rigidity in oligopolistic industries of the US. Cohen and Cyert argue that kink in the demand curve and price rigidity may exist for a short period, for lack of inter-firm information, especially when new and unknown rivals enter the market. They are of the opinion that kink is dearly not a stable long-run equilibrium.

4. ***Sweezy's Conclusion Conflicts with Marginal Productivity Theory.*** In Sweezy's model *MC* curve can shift up and down (say, between finite points *J* and *K* in Fig. 17.3), while *MR* remains the same. This argument is in conflict with *marginal productivity theory* of factor pricing as this means that factor prices do not necessarily equal the marginal revenue (*MR*) productivity.

17.7 COLLUSIVE MODELS OF OLIGOPOLY

The oligopoly models discussed in the previous sections are based on the *assumption* that the oligopoly firms act *independently;* they are in competition with one another; and there is *no collusion* or any kind of agreement between the firms. The oligopoly models of this category are called ***non-collusive models***. In reality, however, oligopoly firms are found to be in some kind of collusion or agreement—open or secret, explicit or tacit, written or unwritten, and legal or illegal—with one another for at least *three major reasons*.

First, collusion eliminates or reduces the degree of competition between the firms and gives them some monopolistic powers in their price and output decisions.

17. Stigler, George, "The Kinky Oligopoly Demand Curve and Rigid Prices," *Jl. of Pol. Eco*., October 1947. pp. 432–46.

18. Simon, Julian "A Further Test of the Kinky Oligopoly Demand Curve", *Am. Eco. Rev*., December 1969, pp. 971–75.

Secondly, collusion reduces the degree of uncertainty surrounding the oligopoly firms and ensure profit maximization.

Thirdly, collusion creates some kind of barriers to the entry of new firms.

The models that deal with the collusive oligopolies are called ***collusive oligopoly models***. Collusion between firms may take many forms depending on their relative strength and objective of collusion, and on whether collusion is legal or illegal. There are, however, two major forms of collusion between the oligopoly firms: (*i*) cartel, i.e., firms' association and (*ii*) price leadership agreements.

On the basis of these forms of collusion between oligopoly firms, collusive oligopoly models are classified as:

1. Cartel models, and
2. Price leadership models.

In this section, we will discuss these two types of oligopoly collusive models.

17.7.1 Price and Output Determination under Cartel Models

Cartel Models of Collusion. A *cartel* is a formal collusion of the oligopoly firms in an industry with a purpose. A general purpose of cartels is to centralize certain areas of managerial decisions and functions of individual firms of the industry, with a view to promoting common benefits. Under cartel system, "the firms jointly establish a cartel organization to make price and output decisions, to establish production quotas for each firm and to supervise market activation of the firms in the industry". Cartels are formed with the purpose of (*i*) eliminating uncertainty in business, and (*ii*) restraining harmful interfirm competition.

Cartels may be in the form of *open* or *secret collusion.* Whether open or secret, cartel agreements are *explicit* and formal, in the sense, that agreements are enforceable on member firms not observing the Cartel Rules or dishonouring the. agreements. 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.

Although a cartel performs a variety of functions for its members, the two important functions of cartels are following.

1. Fixing price for joint profit maximization, and
2. Market sharing between its members.

Accordingly, the economists have developed two cartel models: (*i*) joint profit maximization model, and (*ii*) market sharing model. These cartel models are discussed here.

1. Joint Profit Maximization by the Cartel

Let us suppose that a group of firms producing a homogeneous commodity forms a cartel aiming at joint profit maximization. The firms constitute a cartel management board with powers to take decisions on the following aspects.

1. The total quantity to be produced;
2. The price at which cartel output has to be sold; and
3. The share of each firm in the total output.

The cartel board is provided with cost data of individual firms. Besides, it is supposed to obtain the **necessary** data required to estimate the market demand function and to derive the *AR* and *MR* curves. The cartel board calculates the *MC* and *MR* for the industry. In a sense, the cartel board holds the position of a multi-plant monopoly. It determines the price and output for each firm in the manner a multi-plant monopoly determines the price and output for each of its plants.

The model of price and output determination19 for each firm is presented in Fig. 17.5.

For the sake of analytical convenience, let us suppose that there are only two firms, A and *B,* in the cartel. Their **respective** cost curves are given in the first two panels of Fig. 17.5. In the third panel, *AR* and *MR* **curves** represent the revenue conditions of the *industry.* The *MC* curve in the industry panel is the summation of *MC* curves of the individual firms, i.e., *MC = MCA + MCB*.

Given these cost and revenue conditions, the determination of cartel output is shown in the third panel of Fig. 17.5. As the panel shows, the *MC* and *MR* curves intersect at point *C* determining the industry output at *OQ*. The market price is determined at *PQ = OPC.* The industry output *OQ* is so allocated between firms *A* and *B* that their individual *MC = MR.* The share of each firm in the industry output, *OQ*, can be obtained by drawing a line from point *C* and parallel to X-axis through *MCB* and *MCA.* The points of intersection *CA* and *CB* determine the equilibrium level of output for frims *A* and *B*, respectively. Thus, the share of firms, *A* and *B*, is determined at *OQA* and *OQB*, respectively, where *OQA + OQB = OQ.* At these levels of outputs, they maximize their respective profits.

Fig. 17.5 *Price and Output Determination Under Cartel*

Problems in Joint Profit Maximization

Although the solution to joint profit maximization by cartel appears to be theoretically sound, William Fellner20 gives the following reasons why profits may not be maximized jointly.

First, it is difficult to estimate market demand curve accurately because each firm thinks that the demand for its own product is more elastic than the market demand curve as its product is a perfect substitute for the product of other firms.

Secondly, 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.

19. A mathematical model of joint profit maximization in given in Appendix to the chapter.

20. William Fellner, "Competition among the few: Oligopoly and similar market structure" *EJ.*, 1949.

Thirdly, cartel negotiations take a long time. During the period of negotiations, 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 increases beyond 20 or so, cartel formation becomes difficult, or even if it is formed, it tends to break down soon or it becomes ineffective.

Fourthly, there are 'chiselers' who have a strong temptation to give hidden or undisclosed concessions to their customers. This tendency in the cartel members reduces the prospect of joint profit maximization.

Fifthly, 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 follow the cartel price agreement.

Sixthly, another reason for not charging the cartel price is the fear of entry of new firms. A 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, yet 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.

2. Market Sharing by Cartel

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 agreement, and (*ii*) quota system.

1. **Non-price Competition Agreement.** The non-price competition agreements are usually associated with loose cartels. Under the non-price competition agreement 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 obligatory condition 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. However, 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 without reducing their price below the level of cartel price.

Whether this kind of agreement works or breaks down depend 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 a lower price. This may lead to a price war and the cartel may break down.

2. **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 by which quota is fixed. In practice, however, quota for each firm is determined on the basis of the following factors.

(*a*) Bargaining ability of a firm and its relative importance in the industry,

(*b*) The relative sales or market share of the firms in pre-cartel period, and

(*c*) Production capacity of the firms on the base period. The choice of the base period depends on the bargaining ability of the firm.

Methods of Quota Fixation

Fixation of Quota is a difficult problem. Nevertheless, some theoretical guidelines for market sharing are suggested by the economists. A reasonable criterion for ideal market sharing can be to share the total market between the cartel members in such proportions that the industry's *MC* equals the *MC* of individual firms. This criterion is illustrated in Fig. 17.6 assuming an oligopoly industry consisting of only two firms, *A* and *B*. The *AR, MR* and *MC* curves of the two firms are given in panels (*a*) and (*b*), respectively. The summation of their *AR, MR* and *MC* curves are given in panel (*c*). These curves make the industry *AR, MR* and *M* C curves as shown by the *ARM, MRM* and *MC* curves in panel (*c*). The *MRM* and *MC* curves intersect at point *C* determining the profit maximizing output of the industry at *OQ*. The quota for each firm is determined by finding the output of each firm at which their individual *MR* is equal to industry *MC*. A horizontal line drawn form point *C* through *MRB* and *MRA* determines their quota. As panel (*a*) of Fig. 17.6 shows market share for firm *A* is determined at *OQA* and panel (*b*) shows market share of firm *B* at *OQB*. Thus, the industry output *OQ* is shared between the two firms *A* and *B*, as *OQA* and *OQB*, respectively. Note that $OQ = OQA + OQB$. At output *OCA*, *MC* of firm *A* equals industry's marginal cost, *MC*, and at output *OQB*, *MC* of firm *B* equals industry's *MC*. Thus, under quota system, the quota for firms *A* and *B* would be fixed as *OQA* and *OQB,* respectively.

As Fig. 17.6 shows both the firms fix the same price, *OPM*. However, 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.

Figure 17.6 *Quota Allocation Under Cartel Agreements*

Another reasonable criterion for market sharing under quota system is *equal market share for equal firms.* This criterion is applicable where all firms have *identical* cost and revenue curves. This criterion also leads to a monopoly solution. It resembles Chamberlin's duopoly model.

To illustrate equal market sharing through quota allocation, let us assume that there are only two firms, *A* and *B.* Suppose their *AR*, *MR* and *MC* curves are given as in Fig. 17.6 (a) and (b), respectively, the market revenue and cost curves, which are obtained by

summing the individual revenue and cost curves, are presented in part (c) of the figure.

As shown in panel (c) of Fig. 17.6, *MRM* and *MC* Intersect at point *C*. Thus, the industry equilibrium output is determined at *OQ*. The share of each firm, which maximizes their profits, is so determined that $OQ = OQ$ A + *OQB*. Given the identical cost and revenue conditions, $OQA = OQB$. That is, market is divided equally between firms *A* and *B*. This result can be obtained also by drawing an ordinate from the point where price line (*PM*) intersects the *MRM*, i.e., from point *R*. The market output *OQ* is divided equally between firms *A* and *B*.

It may be noted 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.

17.7.2 Price Leadership Models

Price leadership is an informal position of a firm in most oligopolistic industries. Price leadership may emerge spontaneously due to technical reasons or out of a tacit or explicit agreement between the firms to assign a 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*, i.e., the largest firm in the industry. The dominant firm takes lead in determining the price and in making price changes and the smaller ones follow.

Sometimes, price leadership is **barometric** by a firm whose decisions and actions serve as a barometer for the decisions and actions taken by other firms. In the *barometric price leadership* one of the firms, not necessarily the dominant one, takes the lead generally in announcing a change in price, particularly when such a change is due but is not effected due to uncertainty in the market.

The price leadership is found under both product homogeneity and product differentiation. There may be, however, price differences commensurate 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 stabilization. Achievement of this objective establishes an 'effective price leadership'. Such a leadership can, however, exist and work effectively only under the following conditions.

(*i*) There is a small number of firms in oligopoly;

(*ii*) Entry to the industry is restricted;

(*iii*) Products are, by and large, homogeneous;

(*iv*) Demand for industry is inelastic or has very low elasticity, and

(*v*) Firms have almost similar cost curves.

Price Leadership Models: Given the conditions for effective working of price leadership, the economists have envisaged the following kinds of price leadership models.

1. Price leadership by low cost firm,

2. Price leadership by dominant firm, and
3. Price leadership by barometeric firm.

Let us now discuss price determination under different kinds of price leadership models.

1. Price Leadership by Low-Cost Firm

The price and output decisions under the leadership of a low-cost firm is illustrated in Fig. 17.7. Suppose all the oligopoly firms face identical revenue curves as shown by *AR* and *MR* curves, but they have different cost curves. The largest firm is the low-cost firm and has its cost curves as shown by AC_1 and MC_1. All the rival firms, smaller in size, have higher cost and their identical cost curves are as shown by AC_2 and MC_2. The largest firm has the lower costs because 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 go by the profit maximization rule and fix its

Fig. 17.7 *Price Leadership by a Low-Cost Firm*

equilibrium output at OQ_2 and fix its price at OP_2 (= LQ_2). At this level of output its *MC* = *MR* and hence its profit is 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. But, if they charge a higher price, OP_3, they would lose their customers to the low-cost firm. 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 the price to OP_1 (= JQ_2). The low cost firm can sell its entire output OQ_2 at OP_1 although, at price OP_1, it will make only normal profit. The low cost firm would, however, not do so as it may invite anti-monopoly laws.

Mathematical Solution. Suppose there are two oligopoly firms—Firm 1 and Firm 2—selling homogeneous products and, therefore, they face the same demand curve, but it is expressed differently for the sake of computational convenience. Their demand curves are given as follows.

Firm 1:

$Q_1 = 50 - 0.5P_1$ and $P_1 = 100 - 2Q_1$ …(17.21)

Firm 2:

$Q_2 = 50 - 0.5P_2$ and $P_2 = 100 - 2Q_2$ …(17.22)

Suppose also that Firm 1 is a low-cost firm and Firm 2 is a high-cost firm. Their respective cost functions are given as follows.

(*i*) Firm 1:

$TC_1 = 100 + 20Q_1 + 2Q_1^2$ …(17.23)

and

AC

2

1 = (100 + 20 Q 1 + 2 Q 1)/ Q 1

…(17.24)

(*ii*) Firm 2:

TC

2

2 = 48 + 36 Q 2 + 2 Q 2

…(17.25)

and

AC

2

2 = (48 + 36 Q 2 + 2 Q 2)/ Q 2

…(17.26)

Let us now see how oligopoly firms will set their price and output independently

acting as monopolies.

Like all other firms, **Firm 1** will determine its output at the level that maximizes its

total profit, P1.We know that total profit is maximum where

P1 = TR 1 – TC 1 is maximium.

For the Firm 1, TC 1 is given in Eq. (17.23). What we need to find is firm's TR 1.

TR 1 = P 1 × Q 1 = (100 – 2 Q 1) Q 1

= 100 Q

2

1 – 2 Q 1

…(17.27)

By substituting TR 1 and TC 1 into the profit equation, we get the profit function for

Firm 1 as

P

2

2

1 = 100 Q 1 – 2 Q 1 – (100 + 20 Q 1 + 2 Q 1)

= 80 Q

2

1 – 4 Q 1 – 100

…(17.28)

The profit maximizing output can be obtained by taking the first derivative of the

profit function (17.28) and setting it equal to zero. Thus,

δΠ1

δ Q = 80 – 8 Q 1 = 0

…(17.29)

1

By solving Eq. (17.29), we get Q 1 = 10.

Alternatively, profit maximizing Q 1 can be obtained by finding and equating MC 1

and MR 1. Firm's MC 1 can be obtained by differentiating TC 1 function (17.23) and MR 1

by differentiating its TR 1 function (17.23), as given below.

Given the TC function (17.23), and TR function (17.27),

MC 1 = 20 + 4 Q 1

and

MR 1 = 100 – 4 Q 1

Given the MC 1 and MR 1 functions, Q 1 can be obtained as follows.

MC 1 = MR 1

20 + 4 Q 1 = 100 – 4 Q 1

8 Q 1 = 80

Q 1 = 10

Profit maximizing price (P 1) and average cost (AC 1) of Firm 1 can now be obtained

by substituting 10 for Q 1 in price functions (17.21) and (17.24), respectively. Thus,

P 1 = 100 – 2 Q 1 = 100 – 2(10) = `80

and

AC

2

1 = (100 + 20 Q_1 + 2 Q_1)/ Q_1

= [100 + 20(10) + 2(10)2]/10 = `50

Let us now see how Firm 2 determines its profit maximizing price. **Firm 2** will also set its output at the level that maximizes its total profit, P2.

P2 = TR_2 – TC_2

Firm's TC_2 is given in Eq. (17.25). Its *TR*, i.e., TR_2 can be obtained as follows.

$TR_2 = P_2 \times Q_2 = (100 – 2 Q_2) Q_2 = 100 Q_2 – 2 Q_2$2

By substitution, we get profit function as

P

2

2

2 = 100 Q_2 – 2 Q_2 – (48 + 36 Q_2 + 2 Q_2)

= 64 *Q*

2

2 – 4 Q_2 – 48

…(17.30)

The profit maximizing output can be obtained by taking the first derivative of the profit function (17.30) and setting it equal to zero. Thus,

δΠ2

δ

= 64 – 8 *Q*

Q

2 = 0

…(17.31)

2

By solving Eq. (17.31), we get Q_2 = 8.

Profit maximizing price (P_2) and average cost (AC_2) of Firm 2 can now be obtained by substituting 8 for Q_2 in price functions (17.22) and (17.26), respectively. Thus,

P_2 = 100 – 2 Q_2 = 100 – 2(8) = `84

and

AC

2

2 = (64 + 64 Q_2 + 2 Q_2)/ Q_2

= [64 + 36(8) + 2(82)]/8 = `58

To summarize, given the market demand function and individual cost functions, the two firms will set their price and output in the absense of collusion between them as follows.

Firm 1: Q_1 = 10 and P_1 = 80

and Firm 2: Q_2 = 8 and P_2 = 84

If Firm 2 sets the price of its product at `84, it will lose its market to Firm 1. This kind of market condition may create condition for price leadership by Firm 1. Under the *price leadership model*, therefore, Firm 1, a low-cost firm, acting as price leader will set the price of its product at `80 and Firm 2 will also set the price of its product at `80, the price set by Firm 1. Note that the per unit profit of Firm 2 is reduced from `26 tò22. But this happens when a high-cost firm has to accept the price-leadership of the low-cost firm.

2. Price Leadership by a Dominant Firm

In case of price leadership by a dominant firm, it is assumed that there exists a large-size 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. In that case, the large firm gains the status of a monopoly which may invite legal problems. The dominant firm, therefore, compromises with the existence of the rival firms in the market. It uses its dominance to set its price so as to maximize its profit. The smaller firms recognize their weak position and accept the price set by the dominant firm.

Price leadership and market sharing between the dominant firm and the rival firms as a group is illustrated in Fig. 17.8. Suppose market demand curve is given by *DDM* and total supply by the small firms by the curve *SS* in panel (*a*) of the figure. Given the market demand and supply by the small firms, the problem the dominant firm is how to determine its price and output that will maximize its profit. To solve this problem, the dominant firm works out its demand curve by deducting the quantity supplied jointly by the small firms at different prices from the corresponding market demand. The dominant firm considers the residual of the market as the demand for its own product. Thus, at a given price, the market share of the dominant firm equals the market demand *less* the share of small firms.

In the absense of the dominant firm, market 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 to the dominant firm is zero. When price falls to OP_2, total demand increases to P_2F. Of this demand, small firms supply only P_2C. The market left for the dominant firm equals $P_2F - P_2C = CF$. Following this process, the market-share of the dominant firm at other prices can be easily obtained. Note that the gap between demand curve *DDM* and supply curve P_1Ss below point *E* in Fig. 17.8(*a*) gives the measures the demand for the dominant firm.

Fig. 17.8 *Price Leadership by a Dominant Firm*

The information so derived and plotted graphically gives P_3DD as the demand curve for the dominant firm in Fig. 17.8 (*b*). Since the relation between *AR* and *MR* is known, the *MR* curve for the dominant firm can be derived as *MRD* [Fig. 17.8(*b*)]. If the *MC* curve of the dominant firm is assumed to be given as *MCD*, its profit maximizing output will be *OQD* and price $PQD = OP'$.

Once the dominant firm sets its price at OP', the small firms have to accept this price. As shown in panel (*b*), at price OP', the market demand is equal to $P'B$, of the market demand $P'B$, the market share of the dominant firm is *AB* and joint market share of small firm is equal to $P'B - AB = P'A$. For small firms, therefore, profit maximizing joint output is $P'A$.

Mathematical Treatment of Dominant Firm Model

If market demand function and joint supply function of small firms are known, then the profit maximising market share of the dominant firm and of the small firms can be easily worked out. Suppose there are six firms—one of them being dominant—in an industry supplying a nearly homogeneous product and market demand function for the product is given as

$$QM = 100 - 2P \quad \text{...(17.32)}$$

and the joint supply function of *five small firms* is given as

$$QS = 10 + P \quad \text{...(17.33)}$$

Given the demand and supply functions (17.32) and (17.33), respectively, the market equilibrium without the dominant firm can be obtained by equating the demand and supply functions. Thus, the market is in equilibrium where

$$QS = QM$$

$$10 + P = 100 - 2P$$

$$P = 30$$

The market supplied by five small firms can be obtained by substituting 30 for *P* in either the demand or supply function. By using demand function, we get

$$QS = 100 - 2P = 100 - 2(30) = 40$$

This means that five small firms jointly supply 40 units at $P = 30$. Thus, the market equilibrium without the dominant firm is established at $QM = QS = 40$ at $P = 30$.

Let us now see how the dominant firm works out the demand function for its product and sets its price. The demand function for the dominant firm can be obtained by deducting the equilibrium quantity ($QS = 40$) supplied by the small firms from the market demand function (17.32). Thus,

$$QD = QM - QS = (100 - 2P) - 40$$

$$= 60 - 2P \quad \text{...(17.34)}$$

The dominant firm's profit maximizing output (*QD*) and price (*PD*) can be obtained by finding its *MCD* and *MRD* and equating them. Let us now find *MC* and *MR* of the dominant firm.

Suppose total cost function (*TCD*) of the dominant firm is given as

$$TC_D = 50 + 6\,Q_D + 0.25\,Q_D^2 \qquad \ldots(17.35)$$

Its marginal cost function (*MCD*) can be obtained by differentiating the *TCD* function (17.35).

$$MC_D = \frac{\partial TC_D}{\partial Q_D} = 6 + 0.5\,Q_D \qquad \ldots(17.36)$$

The *TR* function (*TRD*) of the dominant firm can be obtained as follows. Given the *QD* function (17.34), the price function for the dominant firm (*PD*) can be obtained as

$$P_D = 30 - 0.5\,Q_D \qquad \ldots(17.37)$$

Given the price function (17.37), *TRD* (= *PD* × *QD*) can be obtained as

$$TR_D = (30 - 0.5\,Q_D)\,Q_D$$

$$= 30\,Q_D - 0.5\,Q_D^2 \qquad \ldots(17.38)$$

Dominant firm's *MR* function can be obtained by differentiating the *TR* function (17.38) as

$$MR_D = \frac{\partial TR_D}{\partial Q_D} = 30 - Q_D$$

Now that the dominant firm's *MCD* and *MRD* functions have been obtained, we can work out its profit maximizing *QD* and *PD* as follows. At equilibrium,

$$MC_D = MR_D$$

$$6 + 0.5\,Q_D = 30 - Q_D$$

$$1.5\,Q = 24$$

$$Q = 16$$

Given the dominant firm's equilibrium output (*QD* = 16), its equilibrium price (*PD*) can be obtained by substituting 16 for *QD* in its price function (17.37). That is,

$$P_D = 30 - 0.5\,(16)$$

$$P_D = 22 \;(`)$$

To conclude, the dominant firm fixes its output at 16 and price at `22. This price has to be accepted by the small firms. Thus, *PD* = `22 becomes the market price.

The final market share of the dominant and five small firms can be worked out as follows. The total demand at price *P* = 22 can be obtained by substituting 22 for *P* in the market demand function (17.32).

Total Demand = 100 – 2 (22) = 56

Of the total demand of 56 units at pricè22, only 16 units will be supplied by the dominant firm and the remaining part of the market demand, i.e., 56 – 16 = 40, will be shared by the five small firms.

Critical Appraisal of Dominant Firm Model

The dominant-firm 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 leadership of the dominant firm. That is, small firms produce and supply the quantity and charge the price set by the dominant firm. Besides, the model requires that the dominant firm should be both large and a low-cost firm. For, if a firm does not enjoy the economies of large scale and, consequent upon it, the advantage of 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 large nor is a low-cost firm. But such cases are found mostly under recessionary conditions when a relatively smaller firms reduce their price to survive in the market.

Furthermore, if a leading firm loses its cost advantages, it also loses its leadership.

Such cases are frequent in the real business world. Leadership also changes following the innovation of products and techniques of production by the relatively small firms. Lastly, it is assumed that the entry of new firms is prevented either by low-cost of the existing firms or by initial high cost of new firms. In practice, however, many firms having the capacity to diversify their products enter the industry with relatively low initial cost. For these reasons, dominant firm leadership model is not considered to be a realistic one as it is based on unrealistic assumptions. For the same reasons, the solution given by this leadership model may not be stable.

3. The Barometric Price Leadership

Another form of price leadership is *barometric price leadership*. In this form, a firm initiates well publicized changes in price and price changes are generally followed by the rival firms. This kind of price leadership may not necessarily come from the largest firm of the industry. A barometric firm is supposed to have a better knowledge of the prevailing market conditions and has an ability to predict them more precisely than any of its competitors. These qualities of the barometric firm should have been established and recognized over time by the rival firms. The 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 serve 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 large 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 the 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 that has a proven ability to make reasonably good forecasts.

Third, Kaplan21 *et. al.* observe that barometric price leadership often develops as a reaction to a long economic warfare in which all the firms are losers.

17.7.3 Nature of Non-Price Competition in Oligopoly Markets

It is obvious from the above discussion that oligopolists may be reluctant to wage price wars and encroach upon each other's market share. That is, 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 non-price competition are of diverse form. Yet, there are two important techniques of non-price competition.

First, non-price competition involves product differentiation that 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 market of other sellers by advertising their products. Advertising is also necessary to retain market-share in the face of tough competition between the firms.

SUMMARY

- Oligopoly is a market structure with a *few sellers* selling homogeneous or differentiated products.
- Oligopoly market arises because of (*i*) huge capital investment, (*ii*) economies of scale to large scale firms, (*iii*) patent rights to the branded product, (*iv*) control of some firms over certain raw materials, (*v*) merger and acquisition, etc.
- Features of oligopoly include (*i*) small number of firms, (*ii*) inter-dependence of decision making, (*iii*) barriers to entry of new firms, and (*iv*) indeterminate price and output.

21. Kaplan, A.D.H., Joel B. Dirlam, and Robet F. Lanzillotti, *Pricing in Big Business,* (Washington, D.C. the Brooking Institution, 1958), p. 206. Quoted in Cohen and Cyret, *op. cit*.

- Economists of different generations have constructed different kinds of theories to explain the determination of price and output under oligopoly, including those developed by Counrnot, Bertrand, Edgeworth, Stackelberg, and Paul M. Sweezy and later collusive models of oligopoly, including market sharing models and price leadership models.

- According to Cournot's duopoly model, in case of duopoly, market get divided equally in three parts. Each firm supplies one-third of the market at a given price and one third of the market remains unsupplied. His duopoly model can be extended to oligopoly of larger number of firms. For example, if there are three firms, the market gets divided in four parts. Each firm supplies one-quarter of the market and one-quarter of the market remains unsupplied.
- Sweezy's kinked demand curve model does not explain the determination of the price. It establishes that once market price is determined, it remains rigid over time. Price does not change even if there is significant change in cost of production.
- Collusive models include (*i*) cartel market models, and (*ii*) price leadership models. Under cartel models, firms are supposed to form a cartel of legal form. Under cartel system, market gets divided between the firms in such a way that all firms maximize their profit given the demand curve for their product.
- Under price leadership models, price is determined either by a low-cost firm or by a dominant firm. Under price leadership of low cost firm, the profit maximizing price of the product is determined by the low-cost firm and that price is acceptable to all other firms. Under the leadership of the dominant firm, the dominant firm finds its market share as total market less the share of small firms. The dominant firm determines its own profit maximizing price and that price is acceptable to all small firms.
- The theories of price determination under oligopoly are, however, found to be inapplicable to real life market conditions.

REVIEW QUESTIONS AND EXERCISES

1. What are the characteristics of oligopoly? How is an oligopoly market different from monopolistically competitive market in respect of price and output determination? Is price determinate in an oligopoly market?
2. What are the assumptions of theory of oligopoly? How are these assumptions different from those of theory of monopolistic competition?
3. Explain Cournot's model of duopoly, Illustrate graphically whether price is determinate and stable in Cournot's model of duopoly Show that his model of duopoly can be applied to oligopoly?
4. Suppose a demand function is given as $Q = 100 – 2P$. Using Cournot's model, find the equilibrium output and price. Is equilibrium of the firm stable in Cournot's model?
5. Explain and illustrate Chamberlin's model of oligopoly (small group). How is Chamberlin's model different from Cournot's model of oligopoly? Which of the two models offers a more realistic explanation of oligopoly markets?
6. What are the assumptions behind the existence of kinked demand curve? How does it reflect behaviour of the oligopoly firms? Why do rival firms react to a price cut but not to price rise in oligopoly markets?
7. Kinked demand curve model establishes that price once determined in oligopoly does not change even if there is change in cost of production. Using kinked demand curve model show that change in cost of production does not lead to change in price in oligopoly.

8. Explain and illustrate graphically Sweezy's kinked demand curve theory of rigidity in oligopoly, Are prices in oligopoly really sticky? What are the weaknesses of this model?
9. Using kinked demand curve model, show that price remains rigid at its current level in an oligopoly market even if there is an upward shift in the market demand curve. What happens to the price in a buoyant oligopoly market?
10. What is meant by cartel? Why do firms in an oligopoly market form cartel? Illustrate and explain joint profit maximization by the cartel. Does cartel system work efficiently?
11. Suppose there are only two firms in an oligopoly industry, F_1 and F_2, facing the demand and cost functions given as follows.

Demand function: $Q = 25 – 0.5P$

Cost function F_1: $C_1 = 50 + 10Q_1 + Q_1^2$

Cost function F_2: $C_2 = 24 + 10Q_2 + Q_2^2$

Find the profit maximization output for firms F_1 and F_2 and maximum joint profit.

12. Suppose an industry is characterized by oligopoly and is dominated by a large firm. The supply curve for the small firms in an oligopoly market is perfectly elastic. Does the dominant firm have any role in this kind of market? Does it still have a scope to determine its price like a monopolist?

13. Suppose there are only two firms in an oligopolistic industry, F_1 and F_2. While F_1 is a low cost and F_2 is a high cost firm. The firm F_1 acts as a price leader. Using an appropriate price-leadership model, explain and illustrate graphically price determination by the low-cost firm. Should F_2 accept the price leadership of F_1? Give reasons for your answer.

14. How does a firm in oligopoly market gain the status of a dominant firm? Suppose there are four firms, *A*, *B, C* and *D*, in an oligopoly market, firm *D* being the dominant one. Explain and illustrate price determination by the dominant.

15. Suppose in an oligopoly market, the joint demand curve for small firms A, *B* and C is given as $Q = 100 - 2P$ and their joint supply curve is given as $Q = 5 + 2P$. Derive the demand curve for the dominant firm, *D*. Find the price determined by the dominant firm assuming its *MC* function given as $MC = 5 + 0.5Q$.

16. The general demand curve for oligopoly firms is given by the demand function

$D_1 = 50 - 0.5P_1$

The firms however believe that their individual demand function is

$D_2 = 80 - P_2$

Their identical cost function is given as

$TC = 150 + 10Q + 0.05Q^2$

(i) Find the initial level of price and output and

(ii) What is the range of variation in *MC* which will not affect the price and output?

17. Suppose there are two oligopoly firms—Firm 1 and Firm 2. Firm 1 is a low-cost firm whereas Firm 2 is a high-cost firm. Both the firms face an identical demand curve given by the demand function as

$Q = 50 - 0.5P$

The cost functions of the two firms are given, respectively, as

$TC_1 = 100 + 20Q_1 + 2Q_1^2$

and

$TC_2 = 48 + 36Q_2 + 2Q_2^2$

Find the following:

(*a*) Price and output of Firm 2 prior to Firm 1 working as the price leader, and

(*b*) Price and output of Firm 2 after it accepts the price leadership of Firm 1.

18. Suppose there are five firms in an oligopoly market, one of them being a dominant one. The market demand function is given as

$Q_M = 200 - 2P$

and the supply function of 4 small firms together is given as

$Q_s = 20 + P$

The cost function of the dominant firm reads as follows.

$TC_D = 100 + 12Q + 0.25Q^2$

Find the following:

(a) Total supply by 4 small firms and their supply price before the dominant firm fixed its own price.

(b) Price and output of the dominant firm, and

(c) Total market share of four small firms after dominant firm fixes its price.

FURTHER READING

Bain, J.S. (1960), "Price Leaders, Barometers and Kinks', *Journal of Business*.

Bishop, R.L. (1960), 'Duopoly; Collusion or Warfare?' *American Economic Review.*

Browning, E.K. and Browning, J.M. (1986), *Microeconomic Theory and Applications* (New Delhi:

Kalyani Publishers), 2nd Edn., Chapter 13.

ChamBerlin, E.H. (1933), *The Theory of Monopolistic Competition* (Cambridge, MA; Harvard University Press), 8th Edn.

Cohen, KJ. and Cyert, R.M. (1976), *Theory of The Firm* (New Delhi: Prentice-Hall of India), Chapter 12.

Fellner, W.J. (1949), *Competition Among the Few: Oligopoly and Similar Market Structure* (New York: Alfred A. Knopf).

gould, J.P. and E.P. Lazeur (1993), *Microeconomics Theory* (Homeswood, IL: Richard D. Irwin), 6th Edn., Chapter 13.

hall, R.L. and Hitch, C.J. (1939), 'Price Theory and Business Behaviour', *Oxford Economic Papers*.

henderson, J.M. and Quant, R.E. (1974), *Microeconomic Theory: A Mathematical Approach* (New York: McGraw-Hill).

Kaplan, A.D.H., Dirlam, J.B. and Lanzillotti, R.F. (1958), *Pricing in Big Business* (Washington, DC: The Brookings Institutions).

Koutsoyiannis, A. (1979), *Modern Microeconomics* (London: Macmillan), 2nd Edn., Chapters 9 and 10.

lanzillotti, R.F. (1957), 'Competitive Price Leadership: A Critique of Price Leadership Model', *Rev. Eco. and Stat.*

maChlup, E (1952), *The Economics of Sellers's Competition* (Baltimore, MD: John Hopkins).

rothsChil, K.W. (1949), 'Price Theory and Oligopoly', *Economic journal*, LVII.

stigler, G.J. (1947), 'The Kinked Oligopoly Demand Curve and Rigid Prices,' *journal of Political Economy,* IV.

stigler, G.J. (1964), 'A Theory of Oligopoly,' *journal of Political Economy.*

Sweezy P.M. (1939), 'Demand Under Conditions of Oligopoly,' *journal of Political Economy,* 568–573.

APPENDIX TO CHAPTER 17

Mathematical Model of Joint Profit Maximization by Cartel under Oligopoly

The cartel model of joint profit maximization has already been explained and illustrated graphically in section 17.7 of the Chapter. Here, the cartel model presented mathematically.

Let us assume that there are only two firms, F_1 and F_2, forming a cartel for joint profit maximization under the following conditions.

(*i*) The market demand function is given as

$Q = 50 - 0.5P$, and

price function as

$P = 100 - 2P$.

(*ii*) Cost function of F_1 is given as

$C_1 = 100 + 20Q_1 + Q_1^2$

Cost function of F_2 is given as

$C_2 = 48 + 36Q_2 + 2Q_2^2$

(*iii*) Total profit of F_1 is given as Π_1 and that of F_2 as Π_2. Thus,

Joint Total Profit = $\Pi_1 + \Pi_2 = \Pi$.

The total profit of firms F_1 and F2 can be measured as

$\Pi_1 = R_1 - C_1$

and

$\Pi_2 = R_2 - C_2$

where R_1 = total revenue of F_1 and R_2 = total revenue of firm F_2.

Since $R_1 = P \cdot Q$ and $R2 = P \cdot Q2$, by substitution, total joint profit can be redefined as

$\Pi = P \cdot Q_1 + P \cdot Q2 - (CI + C2)$

$= P(Q_1 + Q_2) - (C_1 + C2)$

...(A17.1)

By substituting price and cost functions into the profit equation (A17.1), we get

$\Pi = [100 - 2(Q$

2

2

1 + Q 2)] (Q 1 + Q 2) – (l00 + 20 Q 1 + 2 Q 1 + 48 + 36 Q 2 + 2 Q 2)

= 100 Q

2

2

2

1 – 2 Q 1 Q 2 – 2 Q 1 + 100 Q 2 – 2 Q 2 – 2 Q 1 Q 2 – 148 – 20 Q 1 – 36 Q 2 – 2 Q 2

= 80 Q

2

2

1 + 64 Q 2 – 4 Q 1 – 4 Q 2 – 4 Q 1 Q 2 – 148

…(A17.2)

Since Π = Π + Π2, profit maximization requires that Π1 and Π2 are maximized. The first-order condition of profit maximization requires that partial derivative of the individual profit functions, i.e., ∂Π/∂ Q, is equal to zero. Although, individual profit functions are not known, the partial derivative of the individual firm's profit function can be obtained by differentiating profit function (A17.3) with respect to Q 1 and Q 2 and setting them equal to zero as shown below.

∂Π = 80 – 8 Q

∂

1 – 4 Q 2 = 0

(A17.3)

1

Q

and

∂Π = 64 – 8 Q

∂

2 – 4 Q 1 = 0

(A17.4)

2

Q

Equations (A17.3) and (A17.4) are two simultaneous equations with two unknowns (Q 1 and Q 2) which can be solved for Q 1 and Q 2 as follows. Equations (A17.3) and (A17.4) can be written, respectively, as follows.

8 Q 1 + 4 Q 2 = 80

(A17.5)

4 Q 1 + 8 Q 2 = 64

(A17.6)

By solving these equations for Q 1 and Q 1, we get Q 1 = 8 and Q 2 = 4.

This means that cartel will maximize profit by dividing the market between firms F 1 and F 2 in proportion of 8 and 4. This will maximize their joint profit. Maximum profit can be obtained by substituting 8 and 4 for Q 1 and Q 2, respectively, in profit function (A17.2), reproduced below.

Π = 80(8) + 64(4) – 4(8)2 – 4(4)2 – 4(8)(4) – 148

= 640 + 256 – 256 – 64 – 128 – 148

= 300

Thus, joint profit is maximized at `300.

CHAPTER

18 Game Theory and Strategic Behaviour of Oligopoly Firms

CHAPTER OBJECTIVES

The objective of this Chapter is to discuss an advanced theory of firms' behaviour in oligopoly, known as game theory. The content of this Chapter includes the following aspects of game theory and firms' strategic behaviour.

- Basics of game theory and its application to oligopoly
- Prisoners' dilemma – the nature of problems faced by oligopoly firms
- Application of game theory to oligopolistic strategy
- Meaning and application of Nash Equilibrium

• Strategic behaviour of oligopoly firms

18.1 INTRODUCTION

In Chapter 17, we have discussed the theories of price and output determination under various models offered by the economists under different conditions of oligopoly market structure, including cartel models of joint profit maximization and collusive models of price leadership. The traditional theories of oligopoly, as discussed in the preceding chapter, have not been found theoretically strong enough to explain realistically the interdependence of oligopolistic markets and strategic actions, reactions and counteractions of the oligopoly firms. While traditional theories were evaluated and their deficiencies were pointed out, other academicians – including mathematicians and economists – were making efforts to find reasonable explanation to strategic behaviour of oligopoly firms. They have made significant contributions to explain the strategic behaviour of the oligopoly firms.

The first most important contribution to this field was made by a mathematician John von Neumann and an economist Oskar Morgenstern[1] in 1944. Their contribution was in the form of ***game theory.*** The game theory brings out the strategy used by the oligopoly firms to determine the best possible action to maximize their predetermined objective. Although many other economists have contributed to game theory, Martin Shubik[2] is regarded as

1. J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behaviour* (Princeton, N.J., Princeton University Press, 1944).

2. Martin Shukik, *Strategy and Market Structure* (John Wiley, 1959) and his *Game Theory in Social Sciences* (Cambridge, Mass., MIT Press, 1982).

the 'most prominent proponent of the game theory approach'. A more recent and in-depth work on the game theory and its application of economics and management problems can be found in the work of Prajit K. Dutta[3].

In this chapter, we discuss the ***game theory approach*** to explain the strategic actions and reactions of oligopoly firms.

18.2 BASICS OF GAME THEORY

Before we proceed to discuss the game theory, it is helpful to understand the meaning and purpose of the game theory and some basic terms and tools used in the analysis and application of game theory.

1. *The Game Theory*. In all kinds of games there are two teams. In all games, the objective of the players of each team is to win the game. To win the game, players make their play-strategy and take action in anticipation of possible reactions of the opposite team and plan their own counter action. This concept of strategic play has been applied by von Neumann and Morgenstern to strategic play of oligopoly firms. As a player, each firm formulates its strategic play and estimates its effects on its objective, called ***pay-off.*** The pay-off may be *positive, negative* or *zero-sum* for a firm taking strategic action. Accordingly, if a strategic action taken by a firm may yield some gains to the firm and counteraction by the rival firm neutralises the gain, it is a ***zero-sum game.*** If both the firms—action-taking firm and rival firms—gain from the strategic action taken by a firm, it is a ***positive-sum game.*** And, if both the firms, action-taking firm and rival firms, make losses from the strategic action taken by a firm, it is a ***negative-sum game.***

2. *Interdependence.* The game theory has been formulated on the basis of a realistic assumption of *interdependence* of oligopoly firms. It implies that decision-making of the firms under oligopoly is *interdependent*. That is, while taking a business decision – be it price determination, advertising, introduction of a new product or brand, setting-up a new production unit, or any other issue – oligopoly firms take into account the possible action and reaction of the rival firms. This kind of behaviour of the oligopoly firms shows their *interdependence.*

3. *Strategy.* The term 'strategy' means the course of action to be taken by the oligopoly firms with the purpose of gaining most from an action under the condition of unknown reaction of the rival firms. For example, suppose there are two firms – *A* and *B*. Firm *A* plans to cut down the price of its product. But Firm *A* is not sure of the possible reaction of the rival Firm *B*. There are two possible reactions of the rival Firm *B*: (*i*) it may cut down its own price, and (*ii*) it may not cut down the price. The price cutting firm *A* will assess its gain and losses under these conditions and chose the best option. This is the strategy of the price cutting firm.

4. ***Pay-off Matrix.*** The 'pay-off matrix' is tabular recording of gains and losses of a firm taking an action under different kinds of anticipated reactions of the rival firms. Recall the above example of two firms, *A* and *B*. Firm A estimates its gains and losses in terms of increase in the sales of its product under the following conditions:

(*i*) Estimated increase in its sales when Firm *A* does not reduce its price;

(*ii*) Estimated increase in its sales when Firm *A* reduces its price and Firm *B* does not react;

3. Prajit K. Dutta, *Strategies and Games* (Cambridge, Mass., MIT Press, 1980).

(*iii*) Estimated increase in its sales when Firm *A* reduces its price and Firms *B* also cuts down the price of its product; and

(*iv*) Estimated change in its sales when Firm *A* does not cut down its price but Firm *B* does cut down its price.

When all these estimates are recorded in a cross-sectional tabular form, it produces a *pay-off matrix.*

5. *Dominant Strategy.* As noted above, *strategy* means the course of action planned by an oligopoly firm with the purpose of gaining most from its action. The pay-off of the strategy may be high or it may be low depending on the counteraction taken by the rival firm. A strategic action that yields the best outcome whatever the reaction of the rival firms is called ***dominant strategy.*** In the context of game theory, *dominant strategy* can be defined as *the strategy that gives the best payoff no matter what counteraction is taken by the rival firm.*

Having looked at the 'basics' of the game theory, we proceed now to discuss the game theory and its application to business decision-making. In game theory, the decision-making problem of the oligopoly firms is best exemplified by, what game theorists call, the *prisoners' dilemma.* We begin our discussion with the concept of *prisoners' dilemma* – an example of dilemma faced by the oligopoly firms in decision-making.

18.3 PRISONERS' DILEMMA: THE PROBLEM OF OLIGOPOLY FIRMS

The nature of the decision-making problems faced by the oligopoly firms is exemplified in game theory by *prisoner's dilemma*. To illustrate *prisoners' dilemma,* let us suppose that two persons, *A* and *B,* are partners in illegal activities. They are arrested under the suspicion of being involved in cricket match-fixing. They are lodged in separate jails with no possibility of communication between themselves. They are interrogated by CBI officials under the following conditions disclosed to each of them in isolation.

1. If you confess your involvement in match fixing, 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 you confess and your partner does not confess and you turn approver, then you get 2-year imprisonment and the other person will get 10-year imprisonment.

Given these conditions, each suspect has two options open to him: (*i*) to confess or (*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 match-fixing. But, there is no certainty that if one denies his involvement, the other will also deny—the other one 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 and turns approver (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 Table 18.1 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 18.1 *Prisoners' Dilemma: The Pay-off Matrix*

B's Options

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 *B* will do, nor *B* knows what *A* will do. When they both confess, each one 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 is supposed to reason out the case in this way: It 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 then I will get 10 years’ imprisonment. That will be the worst scenario. It is quite likely that suspect *B* also reasons out their case in the same manner, even if he too is innocent. If they both confess, they would get jail-term for 5 years and would avoid 10 years’ 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 firms are confronted with in the formulation of their business strategy with respect to such problems as strategic advertising, price cutting or cheating the cartel if there is one. 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 following questions arise. Will the rival firms increase ad-expenditure or will they not? 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 planning to increase ad-spending will have to find the answer to these queries under the conditions of uncertainty. To find a reasonable answer, the firm will have to anticipate actions, reactions and counter-actions by the rival firms and chalk out its own strategy. It is in case of such problems that the case of prisoners’ dilemma becomes an illustrative example.

18.4 APPLICATION OF GAME THEORY TO OLIGOPOLISTIC STRATEGY

Let us now apply the game theory to our example of ‘whether or not to increase ad-expenditure’, assuming that there are only two firms, *A* and *B*, i.e., the case of a duopoly. We know that in all games, the players have to anticipate the moves of the opposite player(s) and formulate their own strategy to counter them. 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 two kinds of reactions of the rival firm and their pay-offs.

(*i*) The counter moves by the rival firm in response to increase in ad-expenditure by this firm, and

(*ii*) The *pay-offs* of this strategy under two conditions:

(*a*) when the rival firm does not react, and

(*b*) the rival firm does make a counter move by increasing its ad-expenditure.

In order to find solution to its problem, the firm anticipates the possible reactions of the rival firms and estimates their posible outcomes. The firm will then take decision 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 that the possible outcomes of the ad-game under the alternative moves are given in the pay-off matrix presented in Table 18.2.

Table 18.2 *Pay-off Matrix of the Ad-Game*

(Increase in sales in million `)

		B's Options			
		Increase Ad		*Dont't increase*	
		A	*B*	*A*	*B*
A's Strategy	Increase Ad	20	10	30	0
	Don't increase	10	15	15	5

As the matrix shows, if Firm *A* decides to increase its ad-expenditure, and Firm *B* counters *A*'s move by increasing its own ad-expenditure, *A*'s sales go up by `20 million and those of Firm *B* by `10 million. And, if Firm *A* increases its advertisement and *B* does not, then *A*'s sales increase by `30 million and there are no sales gain for Firm *B*. One can similarly find the pay-offs of the stategy 'Don't increase' in case of both firms. As shown in Table 18.2, if Firm *A* does not increase its Ad-spending and Firm *B* does increase its Ad-spending, then *A*'s sales increases by `10 million and of *B* by `15 million.

Given the pay-off matrix, the question arises: What strategy should Firm *A* choose to optimize its gain from extra ad-expenditure, irrespective of counter-action 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, sales of Firm *A* could increase by `30 million and sales of Firm *B* do not increase. But there is a greater possibility that Firm *B* will go for counter-advertising in anticiption 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 deal of uncertainty. Under these conditions, the first option gives the dominant strategy for Firm *A*.

18.5 NASH EQUILIBRIUM

In the preceding section, we have used a very simple example to illustrate the application of game theory to an oligopolistic market setting, with the following simplifying assumptions.

(*i*) The strategy formulation is a one-time affair,

(*ii*) Only one firm initiates the competitive warfare and other firms only react to action taken by one firm, and

(*iii*) The 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 counter-actions are regular phenomena. Under these conditions, a *dominant strategy* is often non-existent. To analyze this kind of situation, John Nash,[4] an American mathematician, developed a technique, which is known by his name 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 18.2. 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 is given in Table 18.3. 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 18.3 *Nash Equilibrium: Pay-off Matrix of the Ad-Game*

(Increase in sales in million `)

	B's Options			
	Increase AD		Dont't increase	
A's Strategy	*A*	*B*	*A*	*B*
Increase Ad	20	10	30	0
Don't increase	10	15	25	5

It can be seen from the pay-off matrix (Table 18.3) that Firm *A* no longer has 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 expenditure, Firm *B* will have to follow 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 its ad-outlay, Firm *A* gains more in terms of increase in its sales

4. The technique of finding equilibrium where there is no 'dominant strategy', called 'Nash equilibrium' was developed by John Nash, an American mathematician, in 1951.

(`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 do likewise. 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.

CONCLUDING REMARKS

What we have presented here is an elementary introduction to the game theory. It can be used to find equilibrium solution to the problems of oligopolistic market setting under different assumptions regarding the behaviour of the oligopoly firms and market conditions.5 However, despite its merit of revealing the nature and pattern of oligopolistic warfare, game theory often fails to provide a determinate solution.6

SUMMARY

• The *game theory* formulated by Neumann and Morgenstern, is the theory that seeks to explain the rational behaviour oligopoly firms under the condition of actions taken by a firm and reactions of the rival firms.

• If a strategic action taken by a firm yields a gain to the firm and an equal loss to the rival firm, it is known as a ***zero-sum game.*** If all the firms gain from the strategic action taken by a firm, it is a ***positive-sum game.*** And, if both action-taking and rival firms make losses from the strategic action taken by a firm, it is a ***negative-sum game.***

• When an action-taking firm works out its own gains and gains and losses of the rival firm under different kinds of reactions of the rival firm and presents in a cross-sectional table, it is called ***payoff matrix.***

• Prisoners' dilemma in taking a decision under different conflicting conditions is used to illustrate the decision making problem of the oligopoly firms.

• The decision-making process of prisoners is applied to explain the decision-making process of oligopoly firms. A payoff matrix is prepared to find the best option open to oligopoly firms. The best possible option available to a firm is called *dominant strategy*.

• According to John Nash, a dominant strategy is generally non-existent because of continuous change in economic conditions caused by action, reaction and counteractions by the firms. To resolve the problem, Nash formulated a technique to find the equilibrium position – a position that the ultimate result of all actions, reactions and counter actions. It is called *Nash equilibrium.*

• *Nash equilibrium* is a technique to establish the best achievement of all the firms and no firm can do better. Thus, Nash equilibrium refers to situation in which all firms gain the best and no improvement is possible.

5. The interested reader may refer to the advanced references (see Further Reading of this Chapter).

6. Browning, E.K. and J.K., Browning, *Microeconomic Theory and Application* (London, Scott, Foresman & Co., 1989), 3rd Edn., p. 413.

REVIEW QUESTIONS

1. What is the reason for the formulation of the game theory? What is the objective of the game theory?
2. In what way does the game theory help in resolving the decision-making problem of the oligopoly firms? Does the game theory offer a realistic solution to the problem of oligopoly firms?
3. What is meant by (*i*) strategy, (*ii*) payoff, (*iii*) payoff matrix, and (*iv*) dominant strategy? Explain these concepts with some example.
4. What is meant by prisoners' dilemma? How is prisoners' dilemma used to explain the problem of oligopoly firms? How does it help in resolving the problem?
5. What is meant by payoff matrix? Illustrate the formation of the payoff matrix by using an example of business problem of oligopoly firm.
6. What is meant by dominant strategy? How is dominant strategy worked out?
7. What is meant Nash Equilibrium? How was the Nash equilibrium formulated? What purpose does serve?

FURTHER READING

Camerer, C., "Progress in Behavioural Game Theory", *Journal of Economic Perspective,* Summer 1997.

Dixit, A. and S. Skeath, *Games of Strategy* (New York: Norton, 2004).

Dutta, Prajit K., *Strategy and Games* (Cambridge, Mass.: MIT Press, 1999).

FrieDman, James W., *Games Theory with Application to Economics* (New York: Oxford University Press, 1990).

HirsCHey, Mark, *Economics for Managers* (Delhi: Thomson-South Western, 2006), 8th Edn, Ch 14.

tHomas, Christopher R. and Maurice S. Charles, *Managerial Economics: Concepts and Application* (New Delhi: Tata McGraw-Hill Publishing Company, 2005), 8th Edn, Ch. 13.

CHAPTER

19 Alternative Theories

of the Firm

CHAPTER OBJECTIVES

After having discussed the traditional theories of firm in the preceding chapters, we move in this chapter to discuss the 'alternative theories of firm' formulated by the economists of the next generation. The objective of this chapter is to discuss the following important alternative theories of firm.

• Baumol's theory of sales maximization

• Marris's theory of growth maximization

• Williamson's theory of maximization of managerial utility function and

• Cyert-March's Behavioural theory of firm

19.1 INTRODUCTION

Although the conventional theories of firm still hold their ground firmly, several *alternative theories of firm* were proposed during the early 1960s by the economists of the next generation, notably by Simon, Baumol, Marris, Williamson, Berle and Means, Galbraith, and Cyert and March. These economists have questioned the validity of the *profit maximization hypothesis* and the relevance of the conventional theory to modern business, mainly on empirical grounds. The arguments against the profit maximization hypothesis have already been discussed in Chapter 2.

Another major drawback of the conventional theories is that these theories do not recognize the dichotomy between the ownership and management and its role in setting the goal for the firm. Berle and Means[1] were first to point out in 1932, the fact that, in modern business management, there is a separation between management and ownership and its impact on setting the objective of the firm. The proponents of the recent theories of firm argue that the dichotomy between the ownership and management and the shift in decision-making powers from the owners (of the firm) to its managers give managers an opportunity to exercise their discretion in setting the goals for the firm, especially in case of large business corporations. The managers of large business corporations set the goals for the firm which in their judgment are feasible and desirable for the firm's survival and growth. In view of these facts, some economists formulated their own hypotheses

1. Berle, A.A. and G.C., Means, *The Modern Corporation and Private Property* (Commerce Clearing House, N.Y., 1932).

and studied extensively the objectives, motivations and behaviour of firms afresh and developed their own theory of firm. As a result, there are now a number of ***alternative theories of firm*** postulating different objectives of business firms. The alternative theories of the business firms are sometimes classified under the following categories.

(*i*) Managerial theories of firm,

(*ii*) Growth maximization theories of firm,

(*iii*) Maximization of managerial utility theories and

(*iv*) Behavioural theories of firm.

We will however discuss here only those alternative theories of firm which have gained considerable ground in economic literature and have a greater relevance to business decision making on empirical grounds. The theories of this category include:

(*i*) Baumol's theory of sales revenue maximization,

(*ii*) Marris's theory of maximization of firm's growth rate,

(*iii*) Williamson's theory of maximization of managerial utility function,

(*iv*) Cyert-March theory of satisficing behaviour.

We will discuss here only the basic elements of these alternative theories of firm.[2] Our objective here is to make the readers aware of the recent developments in the theory of firm rather than dealing with the alternative theories at length.

19.2 BAUMOL'S THEORY OF SALES REVENUE MAXIMIZATION

Baumols's theory of sales maximization is one of the most important alternative theories of firm's behaviour. The basic premise of Baumol's theory is that *sales maximization,* rather than *profit maximization,* is the plausible goal of the business firms. He argues that there is no reason to believe that all firms seek to maximize their profits. Business firms, as noted above, pursue a number of incompatible objectives and it is not easy to single out one as the most common objective pursued by the firms. However, research conducted at by Baumol revealed that most managers seek to maximize sales revenue rather than profits. He argues that, in modern business, management is separated from ownership, and managers enjoy the discretion to pursue goals other than profit maximization. Their discretion eventually falls in favour of sales maximization. According to research findings of Baumol, business managers pursue the goal of sales maximization for the following reasons.

First, financial institutions consider sales as an index of performance of the firm and tend to provide finance to the firm with growing sales.

Secondly, while profit figures are available only annually at the end of the final accounting year, sales figures can be obtained easily and more frequently to assess the performance of the management. Maximization of sales is more satisfying for the managers than the maximization of profits that go into the pockets of the shareholders.

Thirdly, salaries and slack earnings of the top managers are linked more closely to sales than to profit. Therefore, managers aim at maximizing sales revenue.

Fourthly, the routine personnel problems are more easily handled with growing sales. Higher payments may be offered to employees if sales figures indicate better performance.

2. For a more detailed discussion on these theories, see A. Koutsoyiannis, *Modern Microeconomics* (Macmillan, 2nd Edn.), Chs. 15–18.

Profits are generally known after a year. To rely on profit figures means, therefore, a longer waiting period for both the employees and the management for resolving labour problems.

Fifthly, where profit maximization is the goal and it rises in one period to an unusually high level, this becomes the standard profit target for the shareholders that managers find very difficult to maintain in the long run. Therefore, managers tend to aim at sales maximization rather than profit maximization.

Finally, sales growing at a rate higher than the rate of market expansion indicate growing market share, a greater competitive strength and better bargaining power of a firm in a collusive oligopoly. In a competitive market, therefore, sales maximization is found to be a more reasonable target.

To formulate his theory of sales maximization, Baumol3 has developed two basic models: (*i*) Static Model and (*ii*) Dynamic Model–each with and without advertising. His static models with and without advertising are discussed below4.

19.2.1 Baumol's Model Without Advertising

Baumol assumes cost and revenue curves to be given as in conventional theory of pricing. Suppose that the total cost (*TC*) and the total revenue (*TR*) curves are given as in Fig. 19.1. The total profit curve, *TP*, is obtained by plotting the difference between the *TR* and *TC* curves. Profits are zero where *TR* = *TC*. This point is indicated by the points of intersection of *TR* and *TC* curves.

Fig. 19.1 *Sales Revenue Maximization*

Given the *TR* and *TC* curves, there is a unique level of o utput at which total sales revenue is maximum. The total sales revenue is maximum at the highest point of the *TR* curve. At this point, slope of the *TR* curve (i.e., $MR = \partial TR / \partial Q$) is equal to zero. The highest point on the *TR* curve can be obtained easily by drawing a line parallel to the

3. Baumol, W.J. *Business Behaviour, Value and Growth* (New York, Macmillan, 1959, Revised edition, Harcourt Brace and World Inc., 1967).

4. As summarised in his *Economic Theory and Operations Analysis*, 4th Edn.

horizontal axis and tangent to the *TR* curve. The point *H* on the *TR* curve in Fig. 19.1 represents the total maximum sales revenue. A line drawn from point *H* to output axis shows that sales revenue is maximized at output *OQ* 3. It implies that a sales revenue maximizing firm will produce output *OQ* 3 and its price equals *HQ* 3/ *OQ* 3.

Profit Constraint and Revenue Maximization

At output *OQ* 3, the firm maximizes its total revenue. At this output, the firm makes a total profit equal to *HM* = Total Revenue *HQ* 3 less Total cost *MQ* 3. Since total *TP* curve gives the measure of total profit at different levels of output, profit *HM* = *TQ* 3. If this profit is enough or more than enough to satisfy the stockholders, the firm will produce output *OQ* 3 and charge a price = *HQ*/ *OQ* 3. But, if profit at output *OQ* 3 is not enough to satisfy the stockholders, then the firm's output must be changed to a level at which it makes a satisfactory profit, say OQ_2, which yields a profit $LQ_2 > TQ_3$.

Thus, there are two types of probable equilibrium: ***one*** in which the profit constraint does not provide an effective barrier to sales maximization,5 and ***second*** in which profit constraint does provide an effective barrier to sales maximization. In the second type of equilibrium, the firm will produce an output that yields a satisfactory or target profit. It may be any output between OQ_1 and OQ_2. For example, if minimum required profit is OP_1, then the firm will stick to its sales maximization goal and produce output OQ_3 which yields a profit much greater than the required minimum. Since actual profit (TQ_3) is much greater than the minimum required, the minimum profit constraint is not operative. However, if required minimum profit level is OP_2, output OQ_3 will not yield sufficient profit to meet the target profit. The firm will, therefore, produce an output which yields the required minimum level of profit OP_2 (= LQ_2). Given the profit target OP_2, the firm will produce OQ_2 where its profit is just sufficient to meet requirement of minimum profit. As can be seen in Fig. 19.1, output (OQ_2) is less than the sales maximization output OQ_3. Evidently, the profit maximization output, OQ_1 is less than the sales maximization output OQ_2 (with profit constraint).

19.2.2 Baumol's Model with Advertising

We have shown above how price and output are determined in a static model without advertising. In an oligopolistic market structure, however, determination of price and output is subject to non-price competition. Baumol has reconstructed his model with *advertising* as the typical form of *non-price competition* and suggests that the various forms of non-price competition may be analyzed on similar lines.

In his analysis of advertising, Baumol makes the following assumptions.

(*a*) Firm's objective is to maximize sales, subject to a minimum profit constraint;

(*b*) Advertising causes a shift in the demand curve and hence the total sales revenue (*TR*) rises with an increase in advertisement expenditure (*A*) i.e., $\partial TR/\partial A > 0$;

(*c*) Price remains constant—a simplifying assumption, and

(*d*) Production costs are independent of advertising. This is rather an unrealistic assumption since increase in sales may put output at a different cost structure.

5. The equilibrium is associated with output OQ_3 and profit TQ_3.

Baumol's model with advertising is presented in Fig. 19.2. The *TR* and *TC* are measured on the *Y*-axis and total advertisement outlay on the *X*-axis. The *TR* curve is drawn on the assumption that advertising increases total sales in the same manner as price reduction.

Fig. 19.2 *Sales Revenue Maximisation*

The *TC* curve includes both production and advertisement costs. The total profit curve is drawn by subtracting *TC* from *TR*. The profit so estimated is shown by the curve *PT*. As shown in Fig. 19.2 profit maximizing advertisement expenditure is *OAp* which maximizes profit at *MAp*. Note that *MAp* = *RC*. Assuming that minimum profit required is *OB*, the sales maximizing advertisement outlay would be *OAc*. This implies that a firm increases its advertisement outlay until it reaches the target profit level which is lower than the maximum profit. This also means that sales maximizers advertise more than the profit maximizers to capture a large market share.

Criticism of Baumol's Model

Although Baumol's sales maximization model is found to be theoretically sound and empirically practicable, economists have pointed out the following shortcomings in his model.

First, it has been argued that in the long-run, Baumol's sales maximization hypothesis and the conventional hypothesis would yield identical results, because the *minimum required* level of profits would coincide with the *normal* level of profits.

Second, Baumol's theory does not distinguish between firm's equilibrium and industry equilibrium. Nor does it establish industry's equilibrium when all the firms are sales maximizers.

Third, it does not clearly bring out the implications of interdependence of the firm's price and output decisions. Thus, Baumol's theory ignores not only actual competition between the firms but also the threat of potential competition in an oligopolistic market.

Fourth, Baumol's claim that his solution is preferable to the solutions offered by the conventional theory, from a social welfare point of view, is not necessarily valid.

19.3 MARRIS'S THEORY OF GROWTH RATE MAXIMIZATION

According to Marris6, the objective that managers of a corporate firm set for themselves is to maximize the firm's *balanced growth rate* subject to managerial and financial constraints. To prove his point of view, he developed a model of firm's growth rate maximization. Marris defines firm's growth rate (*Gr*) as

$$Gr = GD = Gc \quad ...(19.1)$$

where *GD* = growth rate of demand for firm's product, and

Gc = growth rate of capital supply to the firm.

Equation (19.1) implies that a firm achieves a *balanced growth rate* when the *growth rate* of demand for its product equals the growth rate of capital supply to the firm. In maximizing firm's growth rate, managers are faced with two constraints: (*i*) managerial constraints, and (*ii*) financial constraints. These constraints arise because managers and owers have different utility functions.

Managerial con straints arise due to (*a*) limits to managers' ability to manage and to achieve optimum efficiency, and (*b*) managers' own job security. ***Financial constraints*** arise due to conflict between managers' own utility function which they attempt to maximize and owners' utility function. As already mentioned in Chapter 2, Marris defines managerial utility (*Um*) and owners' utility (*Uo*) functions as follows.

Manager utility function: Um = f (salary, power, status, job security)

Owner utility function: Uo = f (profit, capital, output, market share, public reputation)

Apparently, there is a divergence and, to some extent, a conflict between the manager's and owner's utility functions. However, Marris argues that the divergence between *Uo* and *Um* is not so wide as it is made out in managerial theories of firm. He claims that the two utility functions con verge into one variable, i.e., *a steady growth in the size of the firm*, however defined. Marris defines steady growth rate of the firm for managers and owners in terms of two different variables—for managers in terms of *Gd*, i.e., growth in demand for firm's product, and for owners in terms of *Gc*, i.e., the growth of firm's capital (*Gc*). Thus, he redefines manager's and owner's utility functions as follows.

$$Um = f(Gd) \quad ...(19.2)$$

and

$$Uo = f(Gc) \quad ...(19.3)$$

According to Marris, managers try to maximize utility functions (19.2) and (19.3) in such a way that *Gd* = *Gc*. This is what Marris calls the '*balanced growth rate*'. The firm reaches its equilibri um when ' *balanced growth rate*' is achieved. This is what Eq. (19.1) implies. The manager's objective is to maximize *balanced growth rate* (*Gr*) such that *Gd* = *Gc*. Thus, the firm is in equilib rium where

$$Gr(\max) = Gd = Gc \quad ...(19.4)$$

Marris redefines *Gd* and *Gc* in Eq. (19.4) in operational terms as given below:

$$Gd = f(d, k) \quad ...(19.5)$$

6. Marris, Robin, "A Model of the Managerial Enterprise",. *Qly. Jl. of Eco*., 1963, reprinted in R. Marris and A. Wood (eds.), *The Corporate Economy* (Macmillan, 1967), and in his *Theory of Manageri al Capitalism* (Macmillan, 1964).

where *d* = diversification7 of product, and *k* = success rate of new products,

and

$$Gc = r(P) \quad ...(19.6)$$

where *r* = financial security ratio assumed to be a constant proportion of profit (Π).

In Marris's mode, *r* is assumed to be determined subjectively by the managers. To elaborate on his theory, Marris has developed an elaborate model8, which is not within the purview of this book. We now turn to another aspect of Marris' theory of firm, i.e., the manager's financial policy.

Financial Policy for Balanced Growth

In their effort to strike a balance between their own and the owner's utility functions, managers adopt a *prudent financial policy*. In formulat ing a prudent financial policy, managers use the following three criti cal ratios.

(*i*) Debt ratio or Leverage (*r* 1) = Value of debts

Total assets

(*ii*) Liquidity ratio (*r* 2) = Liquid assets

Total assets

Retained profits

(*iii*) Profit retention ratio (*r* 3) = Total profit

Managers keep ***debt ratio*** (*r* 1) within a manageable limit by avoid ing high debt liabilities including interest and debt repayment. According to Marris, the reason for this strategy is that a high debt ratio might lead to bankruptcy or insolvency and a low debt ratio means relying heavily on own resources which imposes a limit on capital growth.

Likewise, high and low ***liquid ity ratios*** (*r* 2) are avoided. The reason is a *high liquidity ratio* invites the risk of takeover of liquidity by the dominant group of owners who could use the liquidity for their other ventures. *Low liquidly ratio* is avoided because it implies low financial leverage and low ability to meet payment obligations which often leads to loss of prestige and sometimes even to insolvency.

The ***retention ratio*** is maintained at a medium level which prevents the change of top management (i.e. , job security aspect) and keeps share prices reasonably high. *Low retention ratio* is avoided because it means high distribution of profits which may attract takeover by raiders. *High retention ratio* is avoided because it involves the risk of replacement of the top management.

In brief, a prudent financial policy is devised by constructing 'a financial security ratio' *r*, which is a weighted average of the three *financial ratios*.

Shortcomings of Marris's Theory

Marris's theory is regarded as an important contribution to the theory of firm in so far as it introduces financial ratios as decision variables in determining the firm's goal. Besides, his theory provides a reconciliation between the conflicting utility functions of the managers and owners. However, Marris's theory has been criticized for its following short comings.

7. Here 'diversification' means, generally, the product di fferentiation.

8. For a comprehensive summary, see Koutsoyiannis, *op. cit.* , Ch. 16.

One, Marris assumes cost structure and price to be given. There fore, he assumes implicitly that profit is given too. This assumption is not realistic. If fact, price determination has been the major point of conten tion in the theory of firms whereas Marris ignores this aspect completely. This is one of the serious drawbacks of his theory.

Two, most industries are oligopolistic and hence firms' business decisions are interdependent. Marris's theory does not account for this interdependence in firms' decisions. This implies that product differentiation by rival firms goes unnoticed or is ignored in the firm's decision-making. His theory has, therefore, a limited appli cability.

Three, in an oligopolistic industry, if all the firms seek simul taneously to maximize their growth rate, it imposes a serious limi tation on the growth in demand for firms' product and the supply of capital. Marris's theory does not account for this factor.

19.4 WILLIAMSON'S MODEL OF MAXIMIZATION OF MANAGERIAL UTILITY FUNCTION

Berle-Means Model: A Prelude to Williamson's Model

Williamson's model of maximization of managerial utility function is a culmination of the managerial utility model of Berle and Means. As already mentioned (Chapter 2), A.A. Berle and G.C. Means9 were the first business economists to point out, in 1932, that management is separated from ownership in the large multi-product business corporations and this influences the role of business managers in setting the goals of the large corpora- tions. They argued that owners (the shareholders) look for high dividends and, therefore, they might be interested in profit maxi mization. But, for lack of corporate democracy, the owners have little or no role to play in policy decisions.

On the other hand, managers have different motives, desires and aspirations which they seek to maximize rather than maximizing profit. Besides, since corporate managers can generate the necessary capital internally by means of retained earnings and they do not need to venture into the capital market for debt capital, their decisions and actions are not subject to scrutiny. The managers, therefore, feel free to pursue their own interest in the corporate firms.

J.K. Galbraith[10] developed Berle-Means hypothesis further and exam ined the issue extensively which is known as the Berle-Means-Gaibriath hypothesis. It claims (*i*) that manager-controlled firms have lower profits than owner-controlled firms, and (*ii*) that professional managers have no interest in maximizing profits. While some empirical studies by some other economists[11] support these claims, some others[12] do not. The issue remains controversial.

9. Berle, A.A. and G.C., Means, *The Modern Corporation and Private Property* (Commerce Clearing House, New York, 1932).

10. Galbraith, J.K., *American Capitalism: The Concept of Countervailing Power* (Houghton Miffun, Boston, 1952); *The Affluent Society* (Houghton Mifflin, Boston, 1958); and *The New Industrial Estate* (Houghton Mifflin, Boston, 1967).

11. McGurie L.W., J.S. Chiu, A.O. Elbing, "Executive Incomes, Sales and Profits". *Am. Eco. Rev.* , September 1962; P. Shelton, John, "Allocative Efficiency Vs. 'X-Efficiency: Comments, *Am. Eco. Rev*., December 1967; R.J. Mansen, Chiu, and D.E. Colley, "The Effect of Separation of Ownership and Control on the Performance of the Large Firm", *Qly, Jl. of Eco.* , August 1968.

12. Kamerschen, David R., "The Influence of Ownership and Control on Profit Rates", *Am. Eco. Rev*., June 1968; P., Palmer, John, "The Separation of Ownership from Control in Large U.S. Corporations", *Qly. Rev. of Eco and Bus.* , Autumn 1972; and H., Demsetz, "Where Is the New Industrial Estate?", *Economic Inquiry*, March 1974.

Williamson's Model

Williamson[13] made further improvements in the Berle-Means hypothesis. We discuss Williamson's hypothesis here in some detail.

Williamson's model of maximization of managerial utility function is regarded as another important contribution to the managerial theory of firms' behaviour. To build his model Williamson postulates as follows.

(*i*) Management is di vorced from ownership;

(*ii*) Managers enjoy discretionary powers to set the goals of the firm they manage; and

(*iii*) Managers maximize their own utility function rather than maximizing profi t.

Williamson's *managerial utility function* includes both quantifiable and unquantifiable variables. *Quantifiable variables* are also called *pecuniary variables* which include managers' salary, slack earnings and perks, etc., and *unquantifiable variables* include power, prestige, job security, status, professional excellence and discretionary powers to spend money.

Williamson's model of managerial utility function (Um) can be ex pressed as follows.

Maximize

$Um = f(S, M, ID)$

...(19.7)

subject to a minimum profit

where S = staff salary (management and administration), M = managerial monetary emoluments (including perks, etc.), and ID = discretionary investment.

In Eq. (19.7), S, M and ID are important decision variables in the managerial utility function and, therefore, need some elabora tion. The variable S includes all payments to managerial and administrative staff on account of salary. It increases with expansion and promotion of the supporting staff for the top managers. It reflects the power, prestige, status and professional success of the management. Also, it enhances the market value of the managers. Variable M includes managers' gross emoluments which comprises salary and slack earnings in the form of luxurious residence, office, car, travel grants and entertainment. Variable ID refers to the in vestment that managers make on their own discretion in addition to routine investment meant for the operation of the business to make a certain minimum profit. ID reflects manager's powers, a sense of fulfillment and satisfaction.

Assumptions Williamson makes the following assumptions in his model of managerial utility maximization.

(*i*) Demand function: $Q = f(P, S, e)$

where Q = output, P = price, S = staff expenses, and e = environ mental factors causing an upward shift in the demand curve;

(*ii*) Cost function: $C = f(Q)$ where $\partial C/\partial Q > 0$;

(*iii*) Profit measures:

(*a*) Actual profit = $\Pi = R - C - S$

where R = revenue, C = cost of production, and S = staff salary,

(*b*) Reported profit = $\Pi_R = \Pi - M$

13. Williamson, O., "Managerial Discretion and Business Behaviour", *Am. Eco. Rev.* , 1963, reprinted in M. Gilbert (ed.), *The Modern Business Enterprise* (Penguin, 1973).

where M = managerial emoluments,

(*c*) Minimum profit = $\Pi_0 = \Pi_R - T$

where T = tax and $(\Pi_0 + T) \leq \Pi_R$, and

(*d*) Discretionary profit = $\Pi_D = \Pi = \Pi_0 - T$

A Simple Version of Williamson's Model

Given the assumptions and the parameters, we present here only the simple version of Williamson's model.14 The simple version of the model assumes that 'managerial emoluments' are equal to zero, i.e., $M = 0$. With this assumption, the managerial utility function (19.7) can be written as

Maximize

$$U_m = f(S, I_D) \qquad ...(19.8)$$

Subject to

$$\Pi > \Pi_0 + T$$

The term I_D in Eq. (19.8) is defined as $\Pi - (\Pi_0 + T)$. That is,

$$I_D = \Pi - \Pi_0 - T \qquad ...(19.9)$$

Equation (19.9) implies that managers set aside a part of *actual profit* (Π) as owners' 'minimum profit' (Π_0) and a part for tax payment (T). The balance of the actual profit is available to the managers for the purpose of 'discretionary investment' (I_D).

Note that I_D in Eq. (19.9) is the same as *discretionary profit* (Π_D) given in assumption (*iii* d) above. It means that

$$I_D = \Pi_D$$

By substitution, the managerial utility function (19.8) can be rewritten as

Maximize

$$U_m = f(S, \Pi_D) \qquad ...(19.10)$$

where

$$P_D = \Pi - \Pi_0 - T$$

Equation (19.10) gives the final form of the managerial utility function in the simple version of the model. It must, however, be noted here that there is substitutability between S and Π_D. It means that, given the actual profit (Π), S can be increased only by reducing Π_D, and *vice versa*. Therefore, in their attempt to maximize their utility function (19.10), managers find an *optimum* combination of S and Π_D. This is the point of firm's equilibrium. The firm's point of equilibrium is shown below graphically.

Firm's Equilibrium: Graphical Presentation

Williamson's simple model of firm's equilibrium is presented graphically in Fig. 19.3. To begin with, let us recall that there is *substi tutability* between S and Π_D. This implies that managers can attain a certain level of utility (U) from the various combinations of S and Π_D. This possibility can be shown by an indifference curve as depicted by U_1 in Fig. 19.3(a). The indifference curve U_1 presents the various combina tions of S and Π_D that yield the same level of managerial satisfac tion. By the same logic, an indifference map can be constructed assuming different levels of actual profits (Π) and the associated level of managerial utility, as shown by the indifference curves U_2, U_3 and U_4 in Fig. 19.3(a). The higher the indifference curve, the higher the level of mana gerial satisfaction at different levels of actual profit.

14. Williamson's model is elaborated in two stages—in a simple version and in its general form. We describe here William son's model in its simple version.

Fig. 19.3 (a) *Profit-Salarly Indifference Curve*

Fig. 19.3 (b) *Actual Profit Curve*

The problem now is how to find the optimum point on the indifference map. This task is accomplished by finding the relationship between *S* and Π *D* and the total actual profit (Π). We know that Π = *TR* – *TC* and *TR* = *P* × *Q*. Therefore, by assuming usual demand and cost functions, we can imagine that P increases over some level of output and then it begins to decline. This behaviour of actual profit (Π) is shown by the curve marked Π in Fig. 19.3(b). By combining manager's indifference map and the profit function, one can obtain the optimum combination of *S* and Π *D*, i.e., the point of firm's equilibrium. The combination of indifference map and profit function is shown in Fig. 19.4. The equilibrium of the firm lies at the point at which the high est indifference curve is tangent to the Π-curve. As shown in Fig. 19.4, point *E* is the point of firm's equilibrium. Point *E* denotes a situation in which managerial utility function (*Um*) is maximized subject to a minimum profit of *EM*.

Criticism. Williamson's model, like other models of this category, suffers from certain weaknesses of its own. This model does

Fig. 19.4 *Equilibrium of the Firm:*

not deal satisfactorily with the problem of

Williamson's Model

interdependence of firms under oligopolistic competition. Williamson's model is said to hold only where rivalry is not strong. In the case of strong rivalry, profit maximization hypothesis has been found to be more appro priate.

19.5 BEHAVIOURAL MODEL OF CYERT AND MARCH

As mentioned in Chapter 2, the behavioural model of Cyert and March15 is an extension and modified version of Simon's 'satisficing behaviour' model of corporate firms16. The Cyert-March model can be appreciated better in contrast to other alternative theories of firm. Traditional theory of firm assumes 'profit maximization' as the sole goal of business firms. Managerial utility models emphasize the role of the dichotomy between the ownership and

15. Cyert, R.M. and J.G., March, *A Behavioural Theory of the Firm* (Prentice-Hall, 1963).

16. Not discussed here.

the management in setting business goals and claim that managers maximize their utility function. They argue that managers use their discretion to set goals for themselves differ ent from profit maximization. They set such goals for themselves as maximization of sales revenue, maximization of firm's growth rate, maximization of manager's own utility function, and so on. In contrast to managerial utility models, Cyert and March look at large multiproduct corporations not as an ordinary firm, but as a *coalition of different but related interest groups including owners, managers, workers, input suppliers, customers, bankers, and tax authorities.* All these groups have their own inter est in the corporations and their interests are often in conflict with one another.

- Owners (the stockholders) are interested in maximum profit possible;
- Managers aim at high salary, power and perks;
- Workers are interested in high pay pack ets, bonus, safe working conditions, insurance and other facili ties;
- Customers are interested in high quality goods and lower prices;
- Input suppliers are interested in continuity and growth in demand for their supplies at higher prices;
- Bankers expect and want their loans and advances to be secure and repaid on time; and
- Tax authorities expect honest and regular tax payments.

Obviously there is a conflict—more or less—between the interests of the different interest groups. One of the important managerial tasks is the goal formation for the firm reconciling these conflicting interests. Let us now look at the aspiration levels of different interest groups and the process of goal formation.

Aspiration Levels and Process of Goal Formulation

Goal formulation by reconciling conflicting interests is a complicated task. Cyert and March argue that managers have a crucial task in formulating a goal for the firm that reconciles the conflicting and competing interests of the different inter est groups so as to ensure a smooth functioning of the corpora tion. In reconciling conflicting and competing interests, managers look at the factors that determine the demands of the various interest groups from the corporation. The demands of the various interest groups are determined largely by their 'aspiration levels', past performance of the firm, and information available to the interest groups. For example, managers' demand for a higher salary depends on the level of their aspirations, and their aspirations depend on their experience about the achievements of their aspirations. In a dynamic society, business environment and conditions continue to change. Environmental changes alter the achievements and, therefore, the level of their aspirations and their demands. That is, in a dynamic society—aspirations, achievements and goals of the corporations keep changing continu ously.

Setting the Goals: The Satisficing Behaviour

Now the question arises: How are the goals set? The goals of large multiproduct corporations are set by the top manage ment. Since interest groups are many and their aspirations and expec tations are many and competing, a single goal cannot be set as it will not satisfy all concerned. Therefore, the top management sets a set of diversified

goals. As mentioned already, according to Cyert and March, the top management sets the following five main goals:

(*i*) Production goal,

(*ii*) Inventory goal,

(*iii*) Sales goal,

(*iv*) Market share, and

(*v*) Profit goal.

These goals are determined through a process of continuous bargaining between the coalition groups. The top management attempts in the process of bargaining to bring about a reconci liation between the conflicting goals. However, so long as the firm is able to achieve the above goals, top management finds it helpful in reconciling the 'aspirations' of the interest groups. How the achievement of these goals satisfies the different coalition groups is described here briefly.

Production goal aims at continuity in production irrespective of any seasonal variability of demand. This goal is achieved by prevent ing (*a*) underutilization of capacity in one period and its overu tilization in another period, and (*b*) lay-off of labour in one period and 'rush recruitment' in another. This helps in prevent ing undue variation in the cost of production and the problem of labour unrest and dissatisfaction. As a result, owners, managers and workers are satisfied.

Inventory goal aims at maintaining a balanced inventory of both raw materials and finished goods. A balanced inventory of inputs and raw materials ensures continuity of production and supply of goods to the customers and also keeps the suppliers of inputs satisfied.

Sales and market share goals aim at promotion and enhancing the market share of the firm. Sales are promoted through competi tive advertising and a pricing strategy. Sales promotion and increase in market shares keep top management and owners satis fied.

Profit goal is so determined that it satisfies the owners (the shareholders), the bankers and other financiers of the firm. Besides, the profit goal aims at making adequate financial provision for future projects.

However, setting the goals is an extremely complicated and difficult task. What the top management aims at, in practice, is to achieve an overall satisfactory performance. This they call the firms' 'satisficing behaviour'. This is, according to Simon–Cyert-March, *a bounded rational behaviour*. The practical methods of the 'satis ficing behaviour' are to bring a reconciliation between the con flicting and competing aspirations. The methods that are generally used to find a 'satisficing behaviour' model are the following.

(*i*) Budget allocation and delegation of authority,

(*ii*) Regular payment of dues to related interest groups,

(*iii*) Allocation of funds for R&D as 'side payment',

(*iv*) 'Slack payments' to deserving groups,

(*v*) Allocation of priorities to demand from different groups and meeting them in the same sequence, and

(*vi*) Decentralization of decision-making powers at different levels of managerial functions.

Shortcomings of the Cyert-March Model

The behavioural model of Cyert and March has been criticized on the following grounds.

(*i*) It provides only a simulation of managerial technique rather than providing a behavioural model.

(*ii*) It does not analyze and reveal how a firm reaches its equilibrium level in its 'satisficing behaviour'.

(*iii*) More importantly, it does not deal with the interde pendence in the case of oligopolist firms.

(*iv*) This model has no predictive power whatsoever.

(*v*) At its best, it presents managerial behaviour rather than economic behaviour of the firms.

19.6 CONVENTIONAL *VS*. ALTERNATIVE THEORIES OF THE FIRM

We have described in this chapter some important alternative theories of the firm – alternative to conventional theories. A question that may be asked now is: Do the alternative theories replace the conventional theory of firm? Or to what extent do the alternative theories really offer an alterna tive and more appropriate explanation to firms' behaviour? There are no simple answers to these questions. One thing is clear that the conventional theory of firm based on profit maximization hypothe sis is not the only theory applicable to a multitude of firms—large and small, owner-managed and manager-managed, single-product and multi-product, local and multinational, private and public undertakings, and alternative theories do provide alternative explanations to the firm's behaviour.

As regards the validity and plausibility of the alternative theories, this issue can be examined on both theoretical and empirical grounds. The theoretical plausibility of a theory depends on its power to predict. There is a general consensus that the conven tional theory has greater explanatory and predictive power than the alternative theories of firm. As regards the empirical valid ity, the empirical evidence in support of the alternative theo ries is not unambiguous. In fact, the multitude of alternative theories is in itself an evidence against them. On the contrary, the empirical evidence against the conventional theory is not clear and strong. In conclusion, it can be said that the alternative theories of firm are still in a state of testable hypotheses and they do not offer a replacement to the conventional theory of firm.

SUMMARY

- Alternative theories were formulated to explain the behaviour of the firms due to (*i*) changing market conditions, especially the growth of large scale corporations, (*ii*) separation of ownership and management, (*iii*) difference in the objectives of owners, the shareholders, and managers, and (*iv*) inapplicably of traditional theories of firm.
- The alternative theories based on different business objectives postulated by the economists are classified as (*a*) managerial theories, (*b*) growth maximization theories, (*c*) theories of maximization of managerial utility function, and (*d*) behavioural theories of firm.
- The important theories of various categories discussed in this chapter include (*i*) Baumol's theory of sales maximization, (*ii*) Marris's theory of growth-rate maximization, (*iii*) Williamson's theory of maximization of managerial utility function, and (*iv*) Cyert-March theory of satisficing behaviour.

- According to Baumol's theory business managers seek to maximize sales because (*i*) sales are taken as index of firm's performance by the financial institutions, (*ii*) sales figures are available frequently while profits are available annually, (*iii*) sales give index of performance of managers for their benefits, (*iv*) maintaining high profit level is a difficult proposition, and (*v*) growing sales are a good indicator of firm's performance. When owners put a target profit, managers adjust sales maximization accordingly.
- According to Marris's theory, managers seek to maximize growth rate of the firm. Growth of the firm is measured on the basis of (*i*) growth of sales, and (*ii*) growth capital supply to the firm. There may be constraints in adjusting these two growth rates, but managers find a compromise growth rate.

• According to Williamson, the objective of managers of big corporations is to maximize 'managerial utility function'. Managerial utility function includes (*a*) *pecuniary variables* such as managers' salary, slack earnings and perks, etc., (*b*) *status related variables* such as power to spend money, prestige, job security, status in the organisation, professional excellence, etc. All these variables have to be associated with cost and profit constraints. Williamson has developed a model to find the equilibrium between the managerial utility and cost-profit functions.

• Cyert and March have tried to combine the interest of all associated with large multiproduct corporations, including owners, managers, workers, customers, banks and tax authorities. These categories of people have different objectives and expectations, often conflicting with one another. According to Cyert and March, managers set under this condition certain specific goals in regard to (*i*) production, (*ii*) inventory, (*iii*) sales, (*iv*) market share, and (*v*) profit. Although finding a compromise solution is an extremely complex task, according to Cyert and March, managers do find a reconciling solution in the form of a *bounded rational behaviour.*

REVIEW QUESTIONS

1. What lies at the foundation of the alternative theories of business firms? Do the alternative theories really offer an alternative explanation to firms' behaviour?
2. Explain Baumol's theory of sales revenue maximization? In what way is this theory superior to the conventional theory based on profit maximization hypothesis?
3. Explains Baumol's model of price and output determina tion with and without advertisement. Does this model offer a more appropriate explanation to price and output determination than the conventional theory?
4. Explain Williamson's model of managerial utility maxi mization. How does this model explain the equilibrium of the firm?
5. How does Marris define the balanced growth of the firm? How do managers arrive at the balanced growth? What kind of financial policy do the managers adopt to secure their stake in the firm?
6. What is the basic postulate of the behavioural model of Cyert and March? How does the top management determine the aspirational goals of the firm? Is this model a genuine theory of business firms?
7. How will you judge whether alternative theories really offer a more plausible alternative to the conventional theory of the firm?

FURTHER READING

Baumol, William J, *Business Behaviour, Value and Growth* (Harcourt Brace Javanovich, New York, 1967).

Berle, A.A. and G.C., means, *The Modern Corporation and Private Property* (Commerce Clearing House, N.Y., 1932).

Cyert, R.M. and J.G., marCh, *A Behavioural Theory of the Firm* (Prentice-Hall, N.Y., 1963).

GalBraith, J.K., *American Capitalism*: *The Concept of Countervailing Power* (Houghton Mifflin, Boston, 1952).

GalBraith, J.K., *The Affluent Society* (Houghton Miffln, Boston, 1958).

Koutsoyiannis, A., *Modern Microeconomics* (Macmillan, 1979), Chs. 15–18.

marris roBin, "A Model of the Managerial Enterprise", *Qly. Jl. of Eco*. May 1963.

Williamson O.E., "Managerial Discretion and Business Behaviour", *Am. Eco. Rev.*, 1963, reprinted in M. Gilbert (ed.), *The Modern Business Enterprise* (Penguin, 1973).

Williamson, O.E., *The Economics of Discretionary Behaviour: Managerial Objectives in a Theory of Firms* (Markham, Chicago, 1967).

CHAPTER 20 Pricing Strategies and Practices

CHAPTER OBJECTIVES

The objective of this chapter is to present a brief discussion on the practical methods adopted by firms to determine the price of their product under different conditions faced by them. The strategic methods and pricing practices that are discussed in this chapter are following.

• Cost-Plus pricing method
• Multiple product pricing
• Pricing in life cycle of a product

- Pricing against established products
- Transfer pricing
- Competitive price bidding
- Peak load pricing
- Administered pricing and
- Export pricing and dumping

20.1 INTRODUCTION

From 14 to 17 chapters, 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. We begin our discussion by explaining the *cost-plus pricing method* of product pricing.

20.2 COST-PLUS PRICING

Cost-plus pricing is also known as **'mark-up pricing', 'average cost pricing'** and **'full cost pricing'**. The cost-plus pricing is the most common method of pricing used by the

manufacturing firms. The general practice under this method is to add a 'fair' percentage of profit[1] margin to the average variable cost (*AVC*). The formula for setting the price is given as

$$P = AVC + AVC(m) \qquad \ldots(20.1)$$

where *AVC* = average variable cost, and *m* = mark-up percentage, and *AVC* (*m*) = gross profit margin (*GPM*).

The mark-up percentage (*m*) is fixed so as to cover average fixed cost (*AFC*) and a net profit margin (*NPM*). Thus,

$$AVC(m) = AFC + NPM \qquad \ldots(20.2)$$

The general procedure followed by the firms for arriving at *AVC* and price fixation may be summarized as follows.

The **first step** in price fixation is to estimate the average variable cost. For this, the firm has to ascertain the volume of its output for a given period of time, usually one accounting or fiscal year. To ascertain the output, the firm uses figures of its 'planned' or 'budgeted' output or takes into account its normal level of production. If the firm is in a position to compute its optimum level of output or the capacity output, the same is used as *standard output* in computing the average cost.

The **next step** is to compute the total variable cost (*TVC*) of the 'standard output.' The *TVC* includes direct cost, i.e., the cost of labour and raw material, and other variable costs e.g., electricity and transportation cost, etc. These costs added together give the *total variable cost*. The 'Average Variable Cost' (*AVC*) is then obtained by dividing the total variable cost (*TVC*) by the 'standard output' (*Q*), i.e.,

$$AVC = \frac{TVC}{Q}$$

After *AVC* is obtained, a 'mark-up' in the form of some percentage of *AVC* is added to it as profit margin and the price is fixed. While determining the mark-up, firms always take into account 'what the market will bear' and the degree of competition in the market.[2]

20.2.1 Mark-up Pricing and Marginalist Rule Compared

The mark-up or average-cost pricing method appears to be a 'rule of thumb' totally different from the marginalist rule of pricing. Fritz Machlup has, however, shown that mark-up pricing is not incompatible with the marginalist rule of pricing. Rather, it is very much compatible with marginalist rule of pricing. According to Machlup, when we look into the logic of mark-up pricing, it appears quite similar to the marginalist rule of pricing. Let us now compare the two rules of pricing.

Recall that, according to the marginalist rule, profit is maximum at the level of output where $MC = MR$ and that the mark-up pricing method is given by

$$P = AVC + AVC\,(m)$$

or

$$P = AVC\,(1 + m) \qquad \ldots(20.3)$$

Let us now see how Machlup has proved that the mark-up pricing ultimately converges with the marginalist rule of pricing, at least under constant cost conditions.

1. The 'fair' percentage of profit margin is usually determined on the basis of the firm's past experience and the practice of the rival firms or on the basis of current rate of returns in the industry.

2. Silbertson, A., "Price Behaviour of Firms", *Economic Journal*, September 1970.

Recall that profit is maximum at the level of output at which

$$MC = MR \qquad \ldots(20.4)$$

and, as shown earlier, $MR = P\left(1 - \dfrac{1}{e}\right)$

or

$$MR = P\left(\frac{e-1}{e}\right) \qquad \ldots(20.5)$$

where e is price elasticity of demand.

By substituting Eq. (20.5) in Eq. (20.4), we may restate the necessary condition of profit maximization as

$$MC = P\left(\frac{e-1}{e}\right) \qquad \ldots(20.6)$$

If MC is constant, $MC = AVC$. By substituting AVC for MC, Eq. (20.6) may be rewritten as,

$$AVC = P\left(\frac{e-1}{e}\right) \qquad \ldots(20.7)$$

By rearranging the terms in Eq. (20.7), we get

$$P = AVC \div \left(\frac{e-1}{e}\right)$$

return

o

$$P = AVC\left(\frac{e}{e-1}\right)$$

...(20.8)

Now, consider Eq. (20.6). If $MC > 0$, then $P\left(\frac{e-1}{e}\right)$ must be greater than 0 For $P\left(\frac{e-1}{e}\right)$ to be greater than 0, e must be greater than 1. This implies that profit can be maximized only when $e > 1$. The logic of this conclusion can be provided as follows. Given the Eq. (20.5) and Eq. (20.6), if $e = 1$, $MR = 0$, and if $e < 1$, $MR < 0$, it means that if $MR < 0$ and $MC > 0$, or in other words, when MR π MC, then the rule of profit maximization breaks down. Thus, profit can be maximized only if $e > 1$, and $MC > 0$.

Now if $e > 1$, then the term $e/(e - 1)$ will always be greater than 1. Let e be greater than 1 by an amount, say m. Then

$$\frac{e}{e-1} = (1 + m) \qquad ...(20.9)$$

By substituting term $(1 + m)$ from Eq. (20.9) for $e/(e - 1)$ in Eq. (20.8), we get

$$P = AVC\,(1 + m) \qquad ...(20.10)$$

where m denotes the mark-up rate.

Note that Eq. (20.10) is exactly the same as Eq. (20.3). This means that the mark-up rule of pricing converges with the marginalist rule of pricing. In other words, it is proved that the mark-up pricing method leads to the marginalist rule of pricing. However, m in Eqs. (20.3) and in (20.9) need not be the same.

20.2.2 Limitations of Mark-up Pricing Rule

The cost-plus pricing has certain *limitations*, which should be borne in mind while using this method for price determination.

First, cost-plus pricing assumes that a firm's resources are optimally allocated and the standard cost of production is comparable with the average cost of the industry. In reality, however, resources may not be optimally allocated. The cost estimates based on these assumptions may be an overestimate or an underestimate. Under these conditions, pricing may not be commensurate with the objective of the firm.

Secondly, in cost-plus pricing, generally, historical cost rather than current cost data are used. This may lead to under-pricing under increasing cost conditions and to over-pricing under decreasing cost conditions, which may go against the firm's objective.

Thirdly, if variable cost fluctuates frequently and significantly, cost-plus pricing may not be an appropriate method of pricing on regular basis, as it may necessitate frequent change in price.

Finally, it is also alleged that cost-plus pricing ignores the demand side of the market and is solely based on supply conditions. This is, however, not true, because the firm determines the mark-up on the basis of 'what the market can bear' and it does take into account the elasticity aspect of the demand for the product, as shown above.

20.3 MULTIPLE PRODUCT PRICING

Most microeconomic models of price determination are based on the assumption that a firm produces a single, homogeneous product. In actual practice, however, production of a single homogeneous product by a firm is an exception rather than a rule. Almost all firms have more than one product in their line of production. Even the most specialized firms produce a commodity in multiple models, styles and sizes, each so much differentiated from the other that every model or size of the product may be considered a different product. For example, the various models of refrigerators, TV sets, cell phones, computers and car models etc. produced by the same company may be treated as different products for at least pricing purpose. The various models are so differentiated that consumers view them as different products[3] and, in some cases, as close substitutes for each other. It is for this reason that each model or product has different *AR* and *MR* curves and that one product of the firm competes against the other product. The pricing under these conditions is known *as multi-product pricing* or *product-line pricing*.

The major problem in pricing multiple products is that each product has a separate demand curve. But, since all the products are produced under one establishment by interchangeable production facilities, they have only one joint and one inseparable marginal cost curve (*MC*). That is, while revenue curves, *AR* and *MR*, are separate for each product, cost curves, *AC* and *MC*, are inseparable. Therefore, the marginal rule of pricing cannot be applied straightaway to fix the price of each product separately. The problem, however, has been provided with a solution by E.W. Clemens.[4] The solution is similar to the technique employed to illustrate third degree price discrimination under profit maximization assumption. As a discriminating monopoly tries to maximize its revenue in all its markets, so does a multi-product firm in respect of each of its products.

3. For Example, the different models of Maruti cars, viz., Maruti 800, Zen, Maruti 1000, Esteem, Maruti Van and Wagon-R, etc. are treated as a different product by the consumers.

4. Clemens, E.W., 'Price Discrimination and Multiple Product Firm', *Review of Economic Studies* (1950–51), reprinted in *Industrial Organisation and Public policy*, Am. Eco. Assn. (Richard D. Irwin, Illinois, 1959).

To illustrate the multiple product pricing, let us suppose that a firm has four different products— *A*, *B*, *C* and *D* in its line of production. The *AR* and *MR* curves for the four branded products are shown in four segments of Fig. 20.1. The marginal cost for all the products taken together is shown by the curve *MC*, *which is the factory marginal cost curve*. Let us suppose that when the *MRs* for the

Fig. 20.1 *Multi-Product Pricing*

individual products are horizontally summed up, the aggregate *MR* (not given in the figure) passes through point *E* on the *MC* curve. If a line parallel to the *X*-axis, is drawn from point *E* to the *Y*- axis through the *MRs*, the intersecting points will show the points where *MC* and *MRs* are equal for each product, as shown by the line *EMR*, the *Equal Marginal Revenue* line. The points of intersection between *EMR* and *MRs* determine the output level and price for each product. The output of the four products are given as *OQa* of product *A*; *QaQb* of *B*; *QbQc* of *C*; and *QcQd* of *D*. The respective prices for the four products are: *PaQa* for product *A*; *PbQb* for *B*; *PcQc* for *C*, and *PdQd* for *D*. These price and output combinations maximize the profit from each product and hence the overall profit of the firm.

20.4 PRICING IN THE 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. Fig. 20.2 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 no considerable
increase or decrease in the sales.
After the saturation stage, comes the
stage of ***decline*** in which total sales
begin to decline for such reasons
as (*i*) increase in the availability
of substitutes, and (*ii*) the loss of

Fig. 20.2 *Life-Cycle of a Product*

distinctiveness of the 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. We have first discussed the pricing of a product in its initial stage as pricing of a new product and then the pricing method in the 'maturity' and 'decline' stage.

20.4.1 Pricing a New Product

A new product may be either a new brand name added to the existing ones or an altogether new product. Pricing a new brand for which there are many substitutes available 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 *new brand*, market provides adequate information regarding cost, demand, and availability of market, etc. Pricing in this case depends on the nature of the market. However, problems arise in pricing a *new product* without close substitutes 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, generally two kinds of pricing strategies are suggested in pricing a new product, viz., (*i*) skimming price policy, and (*ii*) penetration price policy.

(i) Skimming price policy. The *skimming price policy* is adopted where close substitutes of a new product are not available. This pricing strategy is intended to skim the cream off 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 reduction, especially in case of consumer durables. The initial high price would generally be accompanied by heavy sales promoting expenditure. This policy succeeds for the following reasons.

First, in the initial stage of the introduction of the product, demand is relatively inelastic because of consumers' desire for distinctiveness by the consumption of a new product.

Second, cross-elasticity is usually very low for lack of a close substitute.

Third, step-by-step price-cuts help *skimming consumers'* surplus available at the lower segments of demand curve.

Fourth, high initial prices are helpful in recovering the development costs.

The *post-skimming strategy* includes the decisions regarding the time and size of price reduction. The appropriate occasion for price reduction is the time of *saturation* of the total sales or when strong competition is apprehended. As regards the rate of price reduction, when the product is on its way to losing its distinctiveness, the price-cut has to be appropriately larger. But, if the product has retained its exclusiveness, a series of small and gradual price reductions would be more appropriate.

(ii) Penetration price policy. In contrast to skimming price policy, the penetration price policy involves a reverse strategy. This pricing policy is adopted generally 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 is intended to maximize the profits in the long-run. Therefore, the firms pursuing the penetration price policy set a low price of the product in the initial stage. As the product catches the market, price is gradually raised up. The success of penetration price policy requires the existence of the following conditions.

First, the short-run demand for the product should have an *elasticity greater than unity*. It helps in capturing the market at lower prices.

Secondly, *economies of large-scale production* should be available to the firm with the increase in sales. Otherwise, increase in production would result in increase in costs which might reduce the competitiveness of the price.

Thirdly, the *potential market* for the product ought to be fairly large and have a good deal of future prospects.

Fourthly, the product should have *a high cross-elasticity* in relation to rival products for the initial lower price to be effective.

Finally, the product, by nature should be such that it can be easily accepted and adopted by the consumers.

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, slow growth of purchasing power, consumers' hesitation, etc., penetration price policy would be unsuitable. The reason is 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. Penetration pricing policy has to be followed when lead time, i.e. *,* the period of distinctiveness, is fairly long. If cost-structure shows a decreasing trend over time, 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.

20.4.2 Pricing in Maturity Period

Maturity period is the second stage in the life-cycle of a product. It is a stage between the growth period and 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). It can be defined for all practical purposes as 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 Dean5 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". He should rather move in the direction of "product improvement and market segmentation".

20.4.3 Pricing a Product in Decline

The product in decline is one that enters the post-maturity stage. During this stage, the total sale of the product starts declining. The first step in pricing strategy at this stage is obviously to reduce the price with the objective of retaining sales at some minimum level. The product should be reformulated and remodelled to suit the consumers' preferences. It is a common practice in the book trade. When the sale of a hard-bound edition reaches saturation, paper-back edition is brought into the market. This facility is, however, limited

5. *Managerial Economics., op. cit.,* p. 425.

to only a few commodities. As a final step in the strategy, the advertisement expenditure may be reduced drastically or withdrawn completely, and the residual market may be relied on. This, however, requires a strong will of the producer to retian the low market share.

20.5 PRICING IN RELATION TO ESTABLISHED PRODUCTS

Many producers enter the market often with a new brand of a commodity for which several substitutes are available. For example, cold drinks like Coke and Spot were quite popular in the market during 1980s when new brands like Limca, Thums Up, Double Seven, Mirinda, Pepsi, Teem, Campa, etc., were introduced in the market over time. Many other models of motor cars appeared in the market despite the popularity of Maruti cars. So has been the case with many consumer goods. Besides, a new entrant to the market faces the problem of pricing its 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*.

Generally, the following three types of pricing strategies are adopted in pricing a product in relation to its well established substitutes.

(*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.

20.5.1 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 by adding a new product to its line of production with the objective of utilizing its unused capacity. Also, when the firm expects to face 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 the success of this pricing method, 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 *penetrating pricing*. ***Second,*** this technique of pricing 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.

20.5.2 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. This 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.

20.5.3 Pricing Above the Prevailing Market Price

Sometimes some firms price their product above the on-going or prevailing market price of the competitive products. 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 a prestigious good or deemed to be a product of much superior quality. Consumers of such goods prefer shopping in shopping malls 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 sell even the ordinary goods at a price much higher than the market price. This practice is common among the sellers of readymade garments. Besides, a firm may set a high price for its product if it pursues the 'skimming price strategy'. This pricing strategy is more suitable for innovative products especially when the firm is sure of the distinctiveness of its product. The demand for the commodity must have a low cross-elasticity in respect of competing goods.

20.6 TRANSFER PRICING

Large size firms often divide their production process into different product divisions or their subsidiaries. Also, growing firms add new divisions or departments to the existing ones. The firms then transfer some of their production activities to other divisions. The goods and services produced by the new divisions 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 parent body. 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.[6] We will discuss here briefly his technique of transfer pricing.

To begin with, 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. Now the problem for the company is how to price the product of its subsidiary under the following conditions.

(*i*) Both parent and subsidiary companies have their own profit functions to maximize, and

(*ii*) The refrigerator company has the option of using all the compressors produced by its subsidiary and/or to sell the compressors in a competitive market and its demand is given by a straight horizontal line.

Given these condition of the model, transfer pricing is discussed under two conditions.

(*i*) The parent company uses the entire output of its subsidiary and there is no external market for the compressors, and

(*ii*) There does exist a competitive market for the compressor and refrigerator company sells also in the open market.

6. Hirshleifer, Jack, in his "On the Economics of Transfer Pricing", *Journal of Business*, July 1956.

Let us begin our analysis of transfer pricing model by assuming that there is no external market for the compressors. We will later drop this assumption and assume that there is an external market for the compressors and discuss the technique of transfer pricing under both the alternative conditions.

20.6.1 Transfer Pricing without External Market

When refrigerator company uses its entire compressor ouput, it 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 refrigerator company sells its refrigerators in a competitive market at a constant price, the demand for its product is given by a straight horizontal line as shown by the line *ARr* = *MRr* in Fig. 20.3.

The marginal cost of intermediate good, i.e., compressor, is shown by *MCc* curve and that of the refrigerator body by *MCb*. The *MCc* and *MCb* added vertically give the combined marginal cost curve, the *MCt*. At output *OQ*, for example, *TQ* + *MQ* = *PQ*.

The *MCt* intersects line *ARr* = *MRr* at point *P*. An ordinate drawn from point *P* down to the horizontal axis determines the most profitable outputs of refrigerator bodies and compressors, each at *OQ*. Thus, the output of both—refrigerator bodies and compressors is simultaneously determined. Since at *OQ* level of output, the firm's *MCt* = *MRr*, the refrigerator company maximizes its profits from the final product, the refrigerators.

Now, let us find the price of

Fig. 20.3 *Price Determination of the Final Product*

the compressors. The question that

(Refrigerators)

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 profit maximization rule: profit is maximum where *MC* = *MR*. This rule requires equalizing *MC* and *MR* in respect of compressors. The marginal cost curve for the compressors is given by *MCc* in Fig. 20.3. The firm therefore has to obtain the marginal revenue for its compressors. The marginal revenue of the compressors (*MRc*) can be obtained by subtracting marginal cost of the final good (refrigerator) from the *MRr*.7 Thus,

MRc = *MRr* – (*MCt* – *MCc*)

…(20.11)

For example, in Fig. 20.3, at output *OQ*, *MRr* = *PQj MCt* = *PQj* and *MCc* = *MQ*. By substituting these values in Eq. (20.11), we get

7. The *MRc* can also be obtained by asking the compressor division how much it will supply at different transfer prices. *MRr* – *Tp* gives *MRc*, also known as "marginal net revenue".

MRc = *PQ* – (*PQ* – *MQ*) = *MQ*

And, as shown in Fig. 20.3,

MRc = *PQ* – *PM* = *MQ*

Alternatilely, since in Fig. 20.3,

PQ – *MQ* = *PM*, and *PM* = *TQ*,

therefore,

MRc = *PQ* – *TQ* = *PT* and *PT* = *MQ*

We may recall here that, when price is constant, *ARr* = *MRr,* and *MRr* is constant

Fig. 20.4 *Determination of Transfer Price* as shown in Fig. 20.3. Where *MRr* is constant, *MCt* is a rising function. Under this condition, *MRr* – *MCc* will be a decreasing function. Note that the vertical distance between *ARr* = *MRr* line and *MCc* curve goes on decreasing with increase in production of the goods as shown in Fig. 20.3. When *MRc* (which equals *MRr* – *MCc*) is obtained for different levels of output and graphed, it yields a curve like *MRc* curve shown in Fig. 20.4. The *MCc* curve (which is the same as *MCc* curve in Fig 20.3) intersects the *MRc* at point *P*. At point *P*, *MRc* = *MCc* and output is *OQ*. Thus, the price of compressors is determined at *PQ* in Fig. 20.4. This price enables the compressor division to maximize its own profit.

20.6.2 Transfer Pricing with External Perfectly Competitive Market

We have discussed above 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 refrigerator company can buy compressors from other sellers if the compressor division fails to meet its total demand. The external market may be perfectly or inperfectly competitive. In this section, we discuss transfer pricing under the assumption that the external market is perfectly competitive.

The method of transfer pricing with external market is illustrated in Fig. 20.5. Since the compressor market is perfectly competitive, the demand for compressors is given by a straight horizontal line as shown by the line *P* 2 *D*. In that case *AR* = *MR*. The marginal cost curve of the compressors is shown by *MCc*. The *MRc* curve shows the marginal net revenue from the compressors, (see Fig. 20.5). Note that in the absence of the external market, the transfer price of compressors would have been fixed at *OP* 1 = *P* *Q* 2, i.e., the price where *MRc* = *MCc*. At this price the parent company would have bought compressors from its subsidiary only. But, since compressors are to be produced and sold under competitive conditions, the effective marginal cost of the compressor produced by the refrigerator 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 its profit, the firm sets compressor's price at point *P* where *MRc* > *MCc*. Thus, the transfer price of compressor will be fixed at *PQ* 1 and the refrigerator company would buy *OQ* 1 compressors from the compressor division.

The total output of compressors is determined at a level at which *MCc* is equal to *D* (*AR* = *MR*). That is, profit maximizing level of output is determined by point *R*. At point *R*, the total output of compressors is *OQ* 3. Of this, *OQ* 1 is bought by the refrigerator company itself and the remaining output, *Q* 1 *Q* 3 is 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 *MRc* and transfer price

Let us now consider how transfer price is determined when *MRc* shifts upward to the right. The *MRc* may shift upward because of an increase in demand causing an upward shift in *AR* = *MR*. Let the *MRc* in Fig. 20.5 shift to *MR′ c* which intersects with *MCc* at point *B*. In the absence of an external market, the refrigerator company would

Fig. 20.5 *Determination of Transfer Price with*

have set transfer price of

External Market

compressors at *OP* 3—a price

higher than the free market price *OP* 2. But, since there is an external market in which a lower price is given at *OP* 2, the transfer price cannot exceed the market price or else the refrigerator company would not be in a position to maximize its profit. Nor can the transfer price 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 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. 20.5 shows that after the shift in *MRc* curve to *MR′ c*, the demand for compressors by the refrigerator company increases to *OQ* 4 determined by the profit maximizing equilibrium condition that *AR* = *MR* = *MR′ c*. But the subsidiary company cannot produce *OQ* 4 units of compressors, given its *MCc* and the market price. It will, therefore, produce only *OQ* 3 number of compressors, which equalizes *MCc* 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* 3 *Q* 4, will be bought in the external market, at price *OP* 2 = *TQ* 4. Thus, the refrigerator company will buy *OQ* 3 compressors from its compressor division and buy *Q* 3 *Q* 4 in the external market.

20.6.3 Transfer Pricing Under Imperfect External Market

When the refrigerator market is imperfect, the compressor division faces a downward sloping demand curve 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 *ARx* and *MRx* respectively, in Fig. 20.6, and (*ii*) that the 'marginal net revenue' from the internal use of compressors and the marginal cost of producing compressors are represented by *MRc* and

MCc, respectively. With a view to maximizing the overall profit, the refrigerator company will determine the output of compressors where *MCc* = *MRc* + *MRx*, i.e., where marginal cost of compressors equals the composite marginal revenue. The composite marginal revenue is obtained through horizontal summation of the *MRc* and *MRx* curves as shown by *MRt* in Fig. 20.6.

Fig. 20.6 *Transfer Pricing and Imperfect External Market*

As shown in Fig. 20.6, *MCc* intersects *MRt* at point *P* which determines the profit maximizing 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 *MRx* at point *M* and *MRc* 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*, *MCc* = *MRx* and at point *T*, *MCc* = *MRc*. Thus, the refrigerator company (the parent body) will buy *OQ* 1 for internal use 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 company and compressor division maximize their respective profits.

20.7 PEAK LOAD PRICING

There are certain non-storable products, e.g., electricity, telephones, transport and security services, etc., which are demanded in varying measures during the day as well as night. For example, consumption of electricity reaches its peak in day time. It is called 'peak-load' time. It reaches its bottom in the night. This is called 'off-peak' time. Electricity consumption peaks in daytime because all business establishments, offices and factories come into operation. Electricity consumption decreases during nights because most business establishments are closed and household consumption falls to its basic minimum. In Delhi, demand for electricity peaks during summers due to use of *ACs* and coolers, and it declines

to its minimum level during winters. Similarly, consumption of telephone services is at its peak at day time and at its bottom at nights. Another example of 'peak' and 'off-peak' demand is of railway and air services. During festivals, summer holidays, 'Pooja' vacations, etc., the demand for railway and air travel services rises to its peak.

A technical feature of such products is that they cannot be stored. Therefore, their production has to be increased in order to meet the 'peak-load' demand and reduced to 'off-peak' level when demand decreases. Had they been storable, the excess production in 'off-peak' period could be stored and supplied during the 'peak-load' period. But this cannot be done. Besides, given the installed capacity, their production can be increased but at an increasing marginal cost (*MC*).

20.7.1 Problems in Pricing

Pricing of goods like electricity is problematic. The nature of the problem in a short-run setting is depicted in Fig. 20.7. The 'peak-load' and 'off-load' demand curves are shown by *DP* and *DL* curves, respectively. The short-run supply curve is given by the short-run marginal cost curve, *SMC*.

The problem is 'how to price electricity'.

As Fig. 20.6 shows, if electricity price is fixed in accordance with peak-load demand, *OP* 3 will be the price and if it is fixed according to off-load demand, price will be *OP* L. The problem is: what price should be fixed? If a 'peak-load' price (*OP* 3) is charged uniformly in all seasons, it will be unfair because consumers will be charged for what they do not consume. Besides, it may affect

Fig. 20.7 *Peak-Load Pricing of Electricity*

business activities adversely. If electricity production is a public monopoly, the government may not find it advisable to charge a uniform 'peak-load' price.

On the other hand, if a uniform 'off load' price (*OP* 1) is charged, production will fall to *OQ* 2 and there will be acute shortage of electricity during peak hours. It leads to 'breakdowns' and 'load-shedding' during the peak-load periods, which disrupt production and make life miserable. This is a regular feature in Delhi, the capital city of India. This is because electricity rates in Delhi are said to be one of the lowest in the country.

Alternatively, if an average of the two prices, say *P* 2 is charged, it will have the demerits of both 'peak-load' and 'off-load' prices. There will be an excess production to the extent of *AB* during the 'off-load' period, which will go waste as it cannot be stored. If production is restricted to *OQ* 1, price *P* 2 will be unfair. And, during the 'peak-load' period, there will be a shortage to the extent of *BC*, which can be produced only at an extra marginal cost of *CD*.

20.7.2 Double Pricing System

For the above reasons, generally, a double pricing system is adopted. A higher price, called 'peak-load price' (*OP* 3) is charged during the 'peak-load' period and a lower price (*OP* 1) is charged during the 'off-peak' period. During the 'peak-load' period, production is increased to *OQ* 3 at which *DP* intersects *SMC*, and production is reduced to *OQ* 1 during the 'off-peak' period. However, the system has its own advantages and disadvantages.

Advantages Peak-load pricing system has two advantages.

(*i*) It results in an efficient distribution of electricity consumption. Housewives run their dishwashers and washing machines during the 'off-peak' period.

(*ii*) It helps in preventing a loss to the electricity company and ensures regular supply of electricity in the long-run.

Disadvantages This system has two disadvantages as well:

(*i*) The businesses which are by nature day-business pay higher rates than those which can be shifted to 'off-peak' period.

(*ii*) Billing system is the greatest problem. Each consumer will have to install two meters—one for 'peak-load' and another for 'off-load' period with an automatic switch-over system. This can be done.

Alternatively, the problem can be resolved by adopting a progressive tariff rate for the use of electricity. But, in a country like India, all pervasive corruption will make it inefficient. Delhi Vidyut Board (DVB), even after privatization of electricity distribution, is reportedly able to collect only about 48 per cent of its cost of production. The rest goes to the unauthorized users of electricity.

20.8 INTERNATIONAL PRICE DISCRIMINATION: DUMPING

Price discrimination has already been discussed in case of price determination under monopoly. Price discrimination is practiced also in international trade by some countries. In this section, we discuss the method of international price discrimination called *dumping*.

20.8.1 What is Dumping?

Dumping is a practice of exporting goods at a price lower than the domestic price. In other words, when a country exports its product to other countries at a price lower than its domestic price, it is called dumping. Also, when a monopolist or a monopolistically competitive firm sells its product in the foreign market at a price lower than the price it charges in the domestic market, it is called dumping. Dumping amounts to international price discrimination or, more precisely, price discrimination between the domestic market and the foreign markets. There is nothing bad or unethical about dumping—it is a common practice in an imperfectly competitive business world. Dumping is practicable and profitable under the following conditions:

- The seller of a product has some monopolistic control over the domestic market.
- The product price in the foreign markets is higher because of high cost of production.
- The demand for foreign goods is more price elastic than the demand for domestic goods.
- Domestic and foreign markets are separated by distance and high cost of transportation so that reverse export to the domestic market is not possible.

Benefits from dumping

It has been generally proved that dumping benefits the exporting country. The exporting country benefits because dumping provides an opportunity to maintain a higher level of production and employment and also a higher level of exports. That is, the exporting country benefits from its higher level of production and exports.

However, the importing country has both gains and losses from dumping. Whether the importing country has net gain or net loss from dumping depends on several conditions. The importing country benefits because it receives the products at a price lower than its domestic price. However, the importing country loses a great deal due to a decrease in its production of the domestic import substitute and the consequent loss of employment. When the losses exceed the gains, the importing country is a net loser.

20.8.2 Determination of Dumping Price

In case a country decides to opt for dumping, the question that arises is: 'What should be the dumping price'? This question arises because if the dumping price is very low and exports are not very high, the benefits of dumping will be very low and the purpose of dumping will be defeated. And, if dumping price is closer to the domestic price of the product in the importing country, the exports are bound to be low. In that case also, the benefits of dumping may be defeated. So the question arises: What can be the most appropriate dumping price?

Economists have attempted to answer this question theoretically as well as empirically. This question is answered theoretically by replicating the case of price discrimination by a monopoly firm facing two markets with different demand curves and applying the theory to international price discrimination. For the purpose of analysis here, the country is treated to be a monopoly firm. The theoretical framework for the determination of dumping price and price discrimination for the domestic and for foreign markets are illustrated in Fig. 20.8. The curves *DD* = *ARD* and *MRD* represent the demand and marginal revenue curves, respectively, in the domestic market and curves *DF* and *MRF* represent the demand and marginal revenue curves, respectively, in the foreign market. The firm's marginal cost is shown by a straight line drawn under a simplifying assumption that *MC* remainsconstant.

Fig. 20.8 *Pricing under Dumping*

As the theory of price and output determination suggests, a monopoly firm maximizes its profits in the both domestic and foreign markets. The profit maximizing price is determined where $MC = MR$. In this example here, the monopoly firm will maximize its profits by setting prices for the domestic and foreign markets in such a way that $MC = MRD = MRF$. As Fig. 20.8 shows, the *MC* line intersects *MRD* at point *D*. Therefore, the profit maximizing price for the domestic market is fixed at *OPD*. Similarly, *MC* line intersects *MRF* at point *F*. Therefore, profit maximizing price for the foreign market is fixed at *OPF*. The price (*OPF*) set for the foreign market is obviously lower than the price (*OPD*) set for the domestic market. The practice of setting a lower price for the foreign markets than the price set for the domestic market is called 'dumping.'

20.8.3 Kinds of Dumping

There are three kinds of dumping: (*i*) persistent dumping; (*ii*) predatory dumping; and (*iii*) sporadic dumping. A brief description of three kinds of dumping follows.

Persistent dumping Persistent dumping is one that is adopted over a long period of time because of the persistence of the favourable conditions for dumping. The conditions, as mentioned above, are: (*a*) a firm enjoys a monopoly or a near monopoly power in the domestic market in respect of a commodity; (*b*) the foreign market's price of this commodity is higher because of high cost of production; (*c*) foreign demand is more price elastic than the domestic demand; and (*d*) foreign and domestic markets are so separated that re-export of the products is not profitable because markets are separated by such factors as high cost of transportation and high import tariffs. The monopolist adopts dumping persistently under these conditions with the objective of maximizing profits in both domestic and foreign markets.

Predatory dumping A predatory dumping is a temporary dumping that a country adopts for short periods intermittently, if required, with the objective of eliminating competitors in the foreign market. This kind of dumping provides monopoly gains to the dumping country. The predatory dumping is resorted to for the period until the competitors are eliminated. Once the competitors are eliminated, the monopolist country discontinues dumping and raises the price to a higher level and thus exploits the consumers abroad. This kind of dumping is, therefore, regarded as 'an unfair method of competition.'

Sporadic dumping As its name suggests, sporadic dumping is done occasionally and for a very short period. This kind of dumping is generally resorted to under such conditions as (i) there is excess supply for some unpredictable reason; (ii) there is a sudden fall in the domestic demand due to some extraneous factors; and (iii) overproduction due to an erroneous production planning. In order to get rid of excess production, the monopolist sells its product in the foreign market at a price lower than its price in the domestic market. Often, the objective in this case is to prevent a downslide in the domestic price.

20.8.4 Effects of Dumping

Dumping is as an interference with the free trade system. It distorts the price structure and, thus, violates the Pareto optimality conditions. It is, therefore, considered to be injurious to the world's economic welfare. However, all kinds of dumping are not equally injurious to world's economic welfare. The anti-welfare effect of dumping depends on its nature.

Sporadic dumping, for instance, is the least harmful. Instead, it occasionally increases the supply of commodities in the importing countries and increases their welfare without reducing the welfare of the exporting countries.

Predatory dumping, on the other hand, is the most harmful form of dumping. It is most injurious to the world welfare, especially to the welfare of the importing countries as it drives their domestic producers out of business and, thereafter, it results in the exploitation of their consumers.

The welfare effect of *persistent dumping* falls between the two extremes. It adds to the world welfare under the following conditions: (*i*) it increases production and employment in the exporting country; (*ii*) it increases the supply of the dumped product in the importing countries; and (*iii*) it does not throw the domestic producers out of business and does not prevent the growth prospects of import substitute industry in the importing country. If production and employment conditions in exporting and importing countries are different from these conditions, dumping may be injurious to world welfare.

However, it is not easy to distinguish between the different kinds of dumping and measure precisely their adverse effects. Therefore, anti-dumping policy measures are adopted in all cases of dumping.

Some examples of dumping

The history of international trade is replete with the cases of dumping. During the period of the Great Depression of 1930s, dumping was frequently practiced in trade of both manufacturing and primary products, though anti-dumping legislation was passed much earlier, in 1921. During the post-World War II period, the cases of dumping were recorded during the 1960s and 1970s. The cases of dumping of colour TV sets by Sony company of Japan and dumping of cars by the Volkswagen Company of Germany—both in the US market—are two famous cases of dumping. Sony was selling its colour TV sets in the US market at a price of $180 per set while its price in the Japanese market was $333 for the same model. During this period, some European car companies, especially Volkswagen of Germany, dumped their cars in the US market. These cases of dumping led to objections by the US government, which threatened to impose tariffs on their products. In response to this threat, companies like Sony and Volkswagen set up their production units in the US[8] and other dumpers raised their prices. However, some cases of dumping in the US still existed. Therefore, the US government had imposed punitive tariff in 1987 against dumping of Japanese semiconductors.

There were also some other cases of dumping during the 1980s. In the year 1987, the Canadian government found that South Korean automobile companies were dumping their cars in the Canadian market. The Canadian government responded by imposing anti-dumping tariffs to the extent of 35–37 per cent on the South Korean cars. These tariffs were, however, withdrawn the very next year because the Canadian government found dumping unharmful to the Canadian car manufacturers.[9]

8. Peter H Linder, (1991), pp. 183–84.

9. For details, see, M Chacholiades, *op. cit.*, p. 208.

SUMMARY

• This chapter deals with pricing practices adopted by the firms under their own specific conditions whatever the nature of the market structure.

• *Cost-plus pricing* is a very common practice adopted by the firms. Under cost-plus pricing system, price of a product is fixed by adding a reasonable profit margin to average cost of production.

• *Multiproduct pricing* refers to the system of pricing different products of a firm. Each product is so priced that profit from each product is maximized. Given the *AR* and *MR* curves of each product and the common *MC* curve, profit maximizing output of each product is determined following the general rule, i.e., *MC* = *MR*. Once profit maximizing output is determined, given the demand curve for a product, its price is easily determined.

• *Pricing a product over its life cycle* refers to product over its life cycle. With this purpose, the life cycle of a product is classified as (*i*) introduction of the product, (*ii*) period of growth, (*iii*) product maturity period, (*iv*) saturation stage of product, and (*v*) stage of decline. Price of the product is kept low during the introduction period and increased gradually over the period of its growth and maturity. Price is kept constant during the period of saturation and is decreased during the decline period.

• In case of *pricing a new product in competition with established products*, the firms have three options, i.e., pricing product at less than or equal to or at more than the ongoing price. Depending on market conditions, the nature of product, and firm's objectives, firms may opt for any of the three options.

• Problem of *transfer pricing* arises when a firm has to price the product of its own subsidiary unit with a different production cost. The pricing method is different under three market conditions: (*i*) there is no external market for the subsidiary product, (*ii*) there is a perfectly competitive external market, and (*iii*) there is a imperfectly competitive market. In case (*i*), price of the subsidiary product is so determined that product's *MR* = *MC*. In case (*ii*), price for internally consumed part of the product is so determined that internal *MR* = *MC* and price for part of the product sold in the external market is so determined that internal *MC* = external

MR. And in case (*iii*), the firm faces downward sloping *AR* and *MR* curves, the simple market rule is applied.

• The problem of *peak-load* pricing is related to non-storable products, e.g., electricity. It is widely known that demand for electricity in countries like India fluctuates seasonally and very widely. Demand for electricity is low during winters and high during summers. The electricity firms have the option to produce less and charge low price during winters and produce more and charge high price during summers. This is impracticable system. Therefore, firms charge an average price.

• *International price discrimination,* known as *dumping,* is a practice of exporting product at a price lower than the domestic price. Dumping is adopted by a country because it gains in terms of higher level of production and employment and prevention of competing countries. Given the domestic *MC* and foreign *AR* (demand) and foreign *MR* curves, the dumping country sets price of its product for the foreign market and quantity to be exported at the level where its domestic *MC* = foreign *MR*.

REVIEW QUESTIONS

1. What is the controversy on the traditional theories of price determination and empiricists' view on price determination in practice? What is the outcome of the controversy?
2. Discuss the cost-plus method of pricing. How is this method of pricing different from the traditional theories of pricing?
3. What is the method of mark-up pricing? Is this method of pricing different from the marginal rule of pricing?
4. Why does the problem of multiple product pricing arise? How is the price of each product determined by a firm with multiple products?
5. Distinguish between penetrating pricing and skimming pricing? Which of these methods of pricing is beneficial for the firm facing severe competition?
6. What is meant by transfer pricing? How is the price of transfer product determined under the following conditions?

(*i*) There is no market for the transfer product; and

(*ii*) There is a competitive market for the transfer product.

7. Suppose a car company produces cars of two brands – AC-car and non-AC car – under the following conditions: (*i*) average and marginal cost curves for both brands are the same, and (*ii*) the two car brands have demand curves with different elasticities. How will the firm determine the price of each brand with objective of profit maximization?
8. What is the life-cycle of a product? What kind of pricing strategy is adopted over the life-cycle of a product? What do you think will be an appropriate price policy when the demand reaches its saturation and substitute products are likely to enter the market?
9. Discuss the technique of multiple product pricing. Illustrate your answer. Why cannot a single average price be fixed for all products?
10. What is meant by 'peak-load pricing'? Why is sometimes peak-load pricing inevitable? What are its advantages and disadvantages?
11. What is meant by dumping? What are the necessary conditions for successful dumping?
12. Suppose domestic and foreign demand have different levels and demand curves have different elasticities. How are domestic and dumping prices are determined to maximise profits from the domestic and foreign markets?

FURTHER READING

Brigham, EugEnE, F. and JamEs L., PaPPas, *Managerial Economics*, 3rd Edn. (Dryden Press, Hinsdale, III., 1979), Ch. 11.

Douglas, Evan J., *Managerial Economics: Theory, Practice and Problem* (Prentice Hall Inc., N.J., 1979), Ch. 11.

KEating, BErry and holton, J. Wilson, *Managerial Economics An Economic Foundation for Business Decisions* (Atomic Dog Publishing, Indian Imprint by Biztantra, New Delhi, 2003), Ch. 11.

Koutsoyiannis, A., *Modern Microeconomics* (Macmillan Press Ltd., 1979), Ch. 11.

sPEncEr milton h., *Managerial Economics* (Richard D, Irwin, III, 1968), Ch. 11.

salvatorE, DominicK, *Managerial Economics in a Global Economy* (Oxford University Press, New Delhi, 7th Edn. 2012), Ch. 11.

thomas, christoPhEr r. and charlEs, s. mauricE, *Managerial Economics: Concepts and Applications*

(Tata McGraw-Hill Publishing Company, New Delhi, 8th Edn., 2005), Ch. 14.

WEBB, samuEl, c., *Managerial Economics* (Houghton Mifflin, Boston, 1976), Ch. 8.

Part V

CAPITAL BUDGETING AND

INVESTMENT DECISIONS

Capital Budgeting and

CHAPTER

21 Investment Decisions

under Certainty

CHAPTER OBJECTIVES

The objective of this chapter is to discuss the process of capital budgeting and investment criteria under certainty. The content of this chapter includes:

- Meaning of capital budgeting
- Prerequisites of capital budgeting
- Investment criteria under certainty
- Sources of capital
- Cost of Capital

21.1 INTRODUCTION

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*. In a broader sense of the term, investments that generate returns over a number of years can be classified under following categories.[1]

1. *Investment in Financial Assets* including bank deposits, deposits with companies, contribution to provident fund (in excess of compulsory deduction), shares and debentures, government bonds and treasury bills, purchase of NSC, buying units, personal lending, etc.
2. *Investment in Physical Assets* including purchase of land, building, machinery, plants, etc.
3. *Investment in Human Capital* including expenditure on skill formation through education and training that increases productivity and earning capacity of a person.
4. *Miscellaneous Investment* including expenditure on replacement of depreciated and obsolete machinery, product diversification, R&D, installation of safety measures for employees, pollution control for public health and safety, and meeting legal requirements.

1. Some categories of investments (legal and illegal) not included here are investment in political profession (the most lucrative profession in India in recent times) bribing the bureaucratic machinery (from ministers down to menials), bribing the police by *pateriwalas* (the vendors), *hafta* to *gundas* (street bullies) of the area, and so on.

Although investment theories and decisions that we discuss in this part of the book apply to all these categories of investment, we will confine our discussion to the investment in machinery, plant and building. In this chapter, we will discuss the following aspects of capital budgeting and investment decisions in respect to this category of investment.

1. Prerequisites of capital budgeting;
2. Determining the optimum level of capital;
3. Investment criteria and investment decisions under certainty, and
4. Sources and cost of capital.

21.2 PREREQUISITES OF CAPITAL BUDGETING

Capital budgeting is essentially a process of conceiving, analyzing, evaluating and selecting the most profitable project for investment. Capital budgeting is of great significance for at least two reasons.

One, *capital expenditure is generally irreversible*. Once an investment is made in some specialized kind of machinery, plant or equipment, it cannot be converted into cash without a loss because resale value of machinery and equipment is often much lower then their original price.

Two, *the very survival of the firm and its growth over time depends on how well planned is its capital expenditure*. It is, therefore, essential that investment projects are well conceived and evaluated and only gainful projects are selected, given the objective of the firm.

Let us now look at some important prerequisites of capital budgeting.

1. ***Defining Capital Expenditure.*** The term 'capital' assumes a specific meaning in investment decisions. Joel Dean has suggested that 'capital expenditure should be defined in terms of economic behaviour rather than in terms of accounting convention'.2 In terms of 'economic behaviour,' capital expenditure means, as already mentioned, the expenditure on acquiring assets that yield returns over a period of time. For the purpose of capital budgeting, only long-term capital expenditure, that is not adjustable in the short-run, is taken into account. The short-run capital expenditure like 'inventories' and 'receivables' which keep varying and are adjustable in the short-run are ignored. Although long-term capital expenditure varies according to the nature and duration of working life of the project, the capital expenditures involving a commitment for at least one year are generally considered for capital budgeting. Broadly speaking, the long-term capital expenditure includes the following items:

(*a*) expenditure on new capital equipments by a firm in the short-run;

(*b*) expenditure on long-term assets by a new firm;

(*c*) expenditure on expansion or diversification of assets and addition to the existing stock of capital by old firms;

(*d*) expenditure on replacement of depreciated capital;

(*e*) expenditure on advertisement which bears fruit over time, and

(*f*) expenditure on research and development (R&D) and innovation.

2. ***Deciding Planning Period.*** The gestation period and overall success of capital expenditures involve a high degree of 'uncertainty and risk'. The risk arising out of

2. Dean, Joel, *Capital Budgeting* (New york: Columbia University Press, 1951), p. 554.

uncertainty can be reduced considerably by planning capital expenditure for the predictable future. Therefore, a major issue involved in capital budgeting is the determination of span of planning period. A clear vision of plan period is necessary for the following ends.

(*i*) Effective planning, execution and control;

(*ii*) Possible dovetailing of old plan with new ones for future and integrated development of the company;

(*iii*) Assessment of economies of scale and determination of plant-size, and

(*iv*) Financial planning and timely acquisition of necessary finances.

3. ***Choice of Decision Rules.*** One of the essential task in sound capital budgeting is the choice of criteria for accepting or rejecting a project. The criteria must be carefully decided in advance. It must be borne in mind that a particular capital project may be capable of standing the test against several criteria, or in other words, it may be capable of serving several purposes. But the various criteria considered for evaluating a project may not be in conformity with one another or their fulfillment may not be the objective of the firm. It is, therefore, necessary that decision rules for accepting or rejecting a project must be decided beforehand.

The criteria and decision rules are normally chosen on the basis of the objective of the firm, such as profit maximization, asset-building, a regular cash flow or maximization of short or long-term gains, etc. The following steps are generally taken to determine the decision rules for investment.

The ***first step*** in determining the decision rules is to clearly define the objective of investment.

The ***second step*** is to select the criterion for evaluating the projects. The three important criteria for evaluating investment projects are (*a*) Pay-back period; (*b*) Present value criterion or discounted cash flow; and (*c*) Internal rate of return. The criterion chosen will depend on the objective of the firm.

Once an evaluation criterion is selected, the ***third step*** is to decide on the approach for the final selection of projects. There are two approaches for the final selection of projects, viz. (*i*) the accept-reject approach, and (*ii*) the project ranking approach. Any of these two

approaches is adopted to select the investment project depending on the availability of funds.

The ***acc ept-reject approach*** is adopted generally where limited funds are available and the firm has to select one or a few from a number of mutually exclusive and alternative projects.

The ***ranking approach*** is adopted generally when a firm has a large amount of funds to invest in several projects at the same time. The projects are ranked in order of their preferability on the basis of the chosen objective.

4. ***Data Collection.*** An important aspect of capital budgeting is to collect relevant, reliable and adequate data on the following aspects of investment.

(*i*) Alternative avenues of investment,

(*ii*) Cost of investment projects,

(*iii*) The expected returns from the chosen projects,

(*iv*) Period of maturity, fruition, and the productive life of the projects,

(*v*) The market rate of interest, and

(*vi*) Availability of internal and external finances.

Collection of required data on these aspects is necessary to determine where to invest and how much to invest.

21.3 DETERMINING THE OPTIMUM LEVEL OF CAPITAL: A THEORETICAL APPROACH

As noted in Chapter 14, the optimum level of production by a profit maximizing firm is determined where $MC = MR$. The same rule can be applied to determine the *optimum stock of capital*. For a profit maximizing firm, the optimum stock of capital is determined where *marginal cost of capital* (MCk) equals *marginal revenue productivity of capital* [3] ($MRPk$). That is, the optimum level of investment is determined where $MCk = MRPk$. Let us explain these terms briefly.

The marginal cost of capital (MCk) is simply the market rate of interest. In a competitive financial market, the rate of interest is given and, therefore, the MCk is constant for a certain level of investment, as shown by the horizontal part of MCk curve in Fig. 21.1. However when demand for funds increases beyond a certain level, the market interest rate tends to increase as indicated by the upward trend in the MCk curve.

As regards the $MRPk$, it is defind as

$$MRPk = MPPk \cdot P$$

where $MPPk$ = marginal physical productivity of capital, and P = price of the product

In general, capital investment is subject to the *law of diminishing returns*. Therefore, $MPPk$ diminishes as the stock of capital increases. Diminishing $MPPk$ multiplied by constant P gives a diminishing $MRPk$ schedule, as shown by the curve $MRPk$ in Fig. 21.1. The curve $MRPk$ represents the *demand curve for capital* by a firm.

Given the definition of MCk and $MRPk$, the condition for optimum level of capital can be expressed as follows.

$$MCk = MRPk$$

To illustrate the application of this rule of optimum capital determination, let us suppose that a firm is considering 5 projects— *A*, *B*, *C*, *D* and *E* with varying $MPPk$. If we arrange the various projects in the descending order of their $MRPk$ and plot them on a graph, we get a step like demand curve for capital, as shown in Fig. 21.1. A smooth, normal demand curve $MRPk$ can be drawn through the central points of $MRPk$ bartop of the various projects. The overall demand for capital is shown by a demand curve depicted by the $MRPk$ curve.

For drawing the MCk curve, let us suppose that financial market is such that a certain amount of capital,

Fig. 21.1 *Optimum Stock of Capital*

3. The other measures of the marginal revenue productivity of capital are '*marginal efficiency of capital*', 'internal rate of return' and 'marginal returns on capital'. However, we use a more familiar term 'marginal revenue productivity of capital' to explain the determination of the optimum level of investment.

say `20 million, can be raised at fixed cost (interest) of 10 per cent. Beyond this limit, cost of capital begins to rise as shown by the upward trend in the MCk curve. As regards the optimum level of capital, as Fig. 21.1 shows, demand curve (*MRPk*) intersects with *MCk* at point *P*. At piont of intersection (*P*), the total profitable level of demand for capital is determined at `24 million. The firm would accept only the first three projects— *A*, *B*, and *C*. In case of rest of the projects, marginal revenue productivity (*MRPk*) less than the marginal cost of capital (*MCk*) and are hence unprofitable and would, therefore, be rejected.

21.4 THE PROCESS OF PROJECT EVALUATION

The theoretical determination of optimum level of investment, as discussed above, gives only a broad idea about how the optimum level of capital is determined. This part of analysis is, however, not sufficient to meet the requirement of capital budgeting and investment decisions. In actual practice, planning, projection, and analysis of both capital and commodities is required to be able to arrive at an appropriate investment decision. Many big firms spend considerable amount of money on pre-investment research, innovation and market research for the purpose. Besides, the firms collect information on capital requirements of each of their operating units in order to assess the total need for capital. The project proposals submitted by the various units, big and small, are then evaluated with a view to knowing their profitability and viability. The following procedure of estimating the earnings on capital expenditures on different investment projects should be, in general, used.4

(*i*) Earning of each project should be estimated separately.

(*ii*) The two most important sources of earning, viz., cost savings and sales expansion or added profits, must be taken into account.

(*iii*) For estimating future earnings, profit projections must be based on the estimated future prices and costs.

(*iv*) Not only the actual earnings but also the opportunity cost of an investment should be taken into account.

(*v*) The stream of capital earning in the distant future must be appropriately discounted to know its present value, particularly in the case of long-term projects.

(*vi*) For assessing and comparing the earnings, average of invested capital per time unit should be used, instead of initial capital outlay.

(*vii*) Productivity of capital should be estimated on the basis of earnings over the life-time of the asset less cost of the investment.

(*viii*) Estimated earnings must be adjusted on account of the indirect contribution of the proposed investment to the existing production facility. This is, however, a very difficult task and may involve a big margin of error.

(*ix*) In a highly competitive market, abnormal profits create conditions for self-destruction by inviting larger investments from the competitors. This possibility should be thoroughly examined and taken care of in estimating the future earnings.

(*x*) The assessment of the risk factor of projects may involve a varying margin of errors. Some systematic method should be followed to make the necessary adjustment on account of margins of error.

4. Dean, Joel, *Managerial Economics*, *op. cit.*, pp. 562–68.

(*xi*) Where earnings are 'defused and conjectural' (e.g., earnings from expenditure on research and employee's recreation facilities) and quantification thereof is not possible for lack of requisite data, estimating such earnings should be avoided till such data are available.

After each project submitted by the various integrated units is thoroughly examined on the basis of principles mentioned above, the question arises as to whether a particular project is to be accepted, rejected or postponed, i.e., the problem of final selection of projects arises. Let us now discuss the standard investment criteria which are, in practice, applied to evaluate the projects in their final selection, under the condition of certainty. Investment decisions under the condition of *risk and uncertainty* will be discussed in the next chapter.

21.5 INVESTMENT DECISIONS UNDER CERTAINTY

In this section, we discuss the methods of investment decisions under the condition of certainty. In the context of investment, *certainty* means that investors have complete knowledge about the market conditions, especially the investment opportunities, cost of capital and the expected returns on the investment. Specifically, certainty means that there is only one outcome of investment and it is known and assured to the investor. For example, money put in fixed deposit in a bank yields an assured return with full certainty.

In this section, we discuss the various criteria for measuring a certain return from a planned investment. Of the several criteria applied for evaluating the profitability of the various kinds of projects, the three most commonly used criteria under certainty are following:

(*i*) Pay-back (or pay-out) period;

(*ii*) Net discounted present value, and

(*iii*) Internal rate of return or marginal efficiency of capital.

These criteria are equally applicable to a variety of investment decisions regarding new investments and those pertaining to replacement, scrapping, and widening or deeping of capital. Incidentally, from analysis point of view, there is no structural difference between decisions on new investment and those on replacement.

Let us now briefly describe the three criteria mentioned above and look into their applicability. These criteria are discussed here under the condition of *certainty*. Investment decisions under the condition of risk and uncertainety will be discussed in the next chapter.

21.5.1 Pay-Back Period Method

The pay-back period is also known as ‘pay-out’ and ‘pay-off’ period. The *pay-back period* method is the simplest and one of the most widely used methods of project evaluation. *The pay-back period is defined as the time required to recover the total investment outlay from the gross earnings,* i.e., gross of capital wastage or depreciation. If a project is expected to generate a constant flow of income over its life-time, the pay-back period may be calculated as given below.

Total Investment outlays

Pay-back Period = GrossReturnperperiod

For example, if a project costs `40,000 million and is expected to yield an annual income of `8,000 million, then its pay-off period is computed as follows:

`40,000 mil ion

Pay-off Period =

= 5 years

`8,000 million

In case of projects which yield cash in varying amounts, the pay-back period may be obtained through the cumulative total of annual returns until the total equals the investment outlay. The sum of cash inflows gives the pay-back period. For example, suppose that the cost of a project is `10,000 million which yields cash flows over 5 years as given in Col. 3 of Table 21.1. The table provides necessary information for the calculation of pay-back period.

Table 21.1 *Calculation of Pay-Back Period*

Year	*Total fixed outlay*	*Annual Cash-flows*	*Cumulative Total of Col. (3)*
	(` in million)	*(` in million)*	*(` in million)*
(1)	*(2)*	*(3)*	*(4)*
1st	10,000	4,000	4,000
2nd	—	3,500	7,500
3rd	—		

		2,500	10,000
4th	—	1,500	11,500
5th	—	1,000	12,500

As the table shows, the cumulative total of annual cash flows breaks-even with the total outlay of the project (`10,000 million) at the end of the 3rd year. Thus, the pay-back period of the project is 3 years.

In case of projects with different investments yielding different annual returns, the project evaluation procedure can be described as follows. First, pay-back period of each project is calculated and then projects are ranked in increasing order of their pay-back period. Let us suppose, for example, that a firm has to select one out of four riskless projects, viz., *A*, *B*, *C* and *D*. The total cost of each project and their respective annual yields are given in columns (2) and (3), respectively of Table 21.2. The calculation of their respective pay-back period given in column (4) of the table. Project *B* ranks 1st and projects *C*, *D* and *A* rank 2nd, 3rd, and 4th, respectively. The firm will invest in these project in the same order, if it adopts the pay-back period criterion for project evaluation.

Table 21.2 *Ranking of Projects*

Project	*Total outlay* (` *in million)*	*Annual return* (` *in million)*	*Pay-back period* *(Years)*	*Rank*
(1)	*(2)*	*(3)*	*(4)*	*(5)*
A	36,000	6,000	36,000 ÷ 6,000 = 6	4
B	24,000	8,000	24,000 ÷ 8,000 = 3	1
C	20,000	5,000	20,000 ÷ 5,000 = 4	2
D	15,000	3,000	15,000 ÷ 3,000 = 5	3

In case projects *A*, *B*, *C* and *D* yield cash flows at different rates in the subsequent years, the cumulative total method can be adopted to calculate their pay-back periods as shown in Table 21.1 and projects ranked accordingly. After projects are ranked, they are selected in order of their ranking depending on the availability of funds.

All other things being the same, a project with a shorter pay-off period is preferred to those with longer pay-off period. This method of ranking projects or project selection is considered to be simple, realistic and safe. Its simplicity is obvious in the calculation of the pay-off period. It is realistic in the sense that businessmen want their money back as quickly as possible and this method serves their purpose. It is safe since it avoids incalculable risk in the long run.

Drawbacks

Although this method is considered to be a simple, safe and reliable method, some consider this method as 'a crude rule of thumb' and can hardly be defended except on the ground of avoiding risk associated with long pay-back projects. Besides, this method assumes that cash inflows are known with a high degree of certainty, even if there is some uncertainty.

The second and the major drawback of this criterion is that it considers only a short period in which cost of project is recovered. It ignores the period and the subsequent returns, after the pay-off period. This criterion, if applied, may deprive the investor of additional earning in future. For example, suppose that an investor has to make a choice between two Projects *A* and *B*, their costs and returns are given as follows:

(*i*) **Project *A* :** Total cost = `24,000

Annual returns `8,000 over three years

Pay-back period = 3 years.

(*ii*) **Project *B* :** Total cost = `20,000

Annual returns `5,000 over six years.

Pay-back period = 20,000/5,000 = 4 years.

Obviously, according to pay-off period criterion, Project *A* will be preferred to project *B*.

But this will lead to foregoing an additional expected income of `6,000, calculated as follows.

Total yield from Project

B = `5000 × 6 (years) = `30,000

Total yield from Project

A = `8000 × 3 (years) = `24,000

Loss of expected additional income

from Project

B = `30,000 – `24,000 = `6000.

The application of pay-back criterion can be justified only if project *B* involves a high degree of uncertainty and risk. Nevertheless, this criterion can be profitably adopted if terminal year of all projects under consideration is the same.

21.5.2 Net Present Value Method

The Concept of Present Value: The Time Value of Money. The concept of the present value of money is very well reflected in the proverb 'a bird in the hand is worth two in the bush'. In general, money received today is valued more than money receivable tomorrow. Cash in hand is valued more because it gives (*i*) liquidity and (*ii*) an opportunity to invest it and earn return (interest) on it. This is called the **time value of money**. The concept of the time value of money is very often applied to investment decisions. Generally, there is a time-lag between investment and its returns. When an investment is made today, it begins to yield returns at some future date. The time gap between the investment and the first return from the investment is called 'time lag'. During the time lag, the investor

loses interest on the expected incomes. This implies that a rupee received today is worth more than a rupee receivable at some future date. Or conversely, a rupee expected one year hence is worth less than a rupee today. In the context of the time value of money, the present value of a future income is lower than its value if received today.

The concept of present value of money can be explained through an example. Suppose that a sum of `100 held in cash today is deposited in a bank at 10 per cent rate of interest. After one year, `100 today will increase tò110. The amount (principal + interest) is worked out as follows.

Amount = 100 + 100 (10/100)

= 100 + 10 = 110

It follows that `110 expected one year hence is worth only `100 today. This means that `100 is the *present value* of `110 to be earned after a period of one year at the interest rate of 10 per cent. The present value (PV) of `110 can be obtained as follows.

$$PV \text{ of } \grave{}110 = \frac{110}{(1+0.1)} = 100$$

The present value of a future income may thus be defined as its value discounted at the current rate of interest. Alternatively, the present value of an amount expected at a future date is the amount of money which must be invested today to get that amount in future.

The Formula for Computing Present Value. Suppose that an amount X_0 is invested for a period of one year at a compounding interest rate. At the end of the year, the total receipt, say X_1 can be expressed as

$$X_1 = (X_0 + rX_0) = X_0(1 + r) \qquad \text{...(21.1)}$$

Equation (21.1) shows that X_0 increases at the rate of $(1 + r)$ to take the value X_1 after one year. It implies that if X_1 is discounted at the same rate of interest, it gives its present value (PV). The formula for computing the present value is given below.

$$PV \text{ of } X_1 = \frac{X_1}{(1+r)} = X_1\left(\frac{1}{1+r}\right) \qquad \text{...(21.2)}$$

In Eq. (21.2), $1/(1 + r)$ is the *discount rate* for one year. Given the rate of interest (i.e., the numerical value for r), any income receivable after one year can be discounted to its present value. For example, the present value of an income of `500 expected after one year at 10 per cent interest per annum (where $r = 0.10$), can be calclated as

$$PV = 500\left(\frac{1}{1+0.10}\right) = 454.55$$

It means that, at 10 per cent interest rate, the present value of `500 expected after one year is `454.55. Given the formule in Eq. (21.2), the discount rate (d_2) for an income receivable after 2 years will be $1/(1 + r)^2$, and for an income receivable after 3 years, $d_3 = 1/(1 + r)^3$ and so on. The formula for discount rate for the n th yar is given as

$$d_n = \frac{1}{(1+r)^n} \qquad \text{...(21.3)}$$

The present value (PV) formula of an amount receivable in the n th yar is given as

$$PV = X_n\left[\frac{1}{\;}\right] \qquad \text{...(21.4)}$$

⌊(1 *r*) *n* ⌋

+

⌋

X

or

PV =

n

(1+ *r*) *n*

Present Value of an Income Stream. The formula for calculating the total present value (*TPV*) of a stream of annual return (*R*) over *n* year is given as

R

R

R

R

TPV =

1

2

3

+

+

+ ...

n

+

2

3

(1+ *r*) (1+ *r*)

(1+ *r*)

(1+ *r*) *n* ..(21.5a)

n

1

= ∑ *Rn*

n

j 1

=

(1+ *r*)

n

R

or

=

n

∑

n

...(21.5 *b*)

j 1

= (1 + *r*)

Net Present Value and Investment Decision. Having noted the concept of present value (*PV*) and the method of calculating *PV* of a future income, let us now see how investment decisions are taken on the basis of present value. In fact, present value (*PV*) adjusted for the cost of investment provides the basis of investment decisions. The *PV* adjusted for its cost is called 'Net Present Value'. The investment decision—accepting or rejecting a project—is taken on the basis of *net present value*. The *net present value* (*NPV*) may be defined as the difference between the present value (*PV*) of an income stream and the cost of investment (*C*), i.e.,

NV = *PV* – *C*

n

⌈

⌉

or

=

1

$\sum nR$ |

| $- C$

...(21.6)

n

j 1

=

⌊(1+ *r*) ⌋

where *C* is the total cost of investment without any recurring expenditure.

The investment decision rules can be specified as follows.

(*i*) If *NPV* > 0, the project is acceptable;

(*ii*) If *NPV* = 0, the project is accepted or rejected on non-economic considerations;

(*iii*) If *NPV* < 0, the project is rejected.

If investment is a recurring expenditure, the total present cost (*TPC*) for *n* years can be calculated in the same manner as present value of an income stream is calulated, i.e.,

n

C

TPC =

n

∑

n

...(21.7)

j 1

= (1 + *r*)

n

n

And then,

NPV =

n

R

Cn

∑

−

n

∑

n

j 1

= (1 + *r*)

j 1

= (1 + *r*)

n *R* − *C*

=

n

n

∑

n

...(21.8)

j 1

= (1 + *r*)

The investment decision rule in this case is the same as given above. If the *NPV* is positive (i.e., *NPV* > 0), the project is profitable and acceptable. The firm can borrow any amount at the existing interest rate (*r*) and invest in it. When a choice between two projects has to be made, the one with higher *NPV* would be chosen.

21.5.3 Internal Rate of Return (*IRR*) Criterion

The Internal Rate of Return (*IRR*) is also called Marginal Efficiency of Investment (*MEI*), Internal Rate of Project (*IRP*) and Break-even Rate (*BER*). The concept of *IRR* can be illustrated with an example. For example, if a one-year project costing `100 million yields `120 million at the end the year, then its internal rate of return (r) can be btained as follows.

120 million

$(1 + r)$ = `100 million

$= (1 + r)\ 100 = 120$

and

$r = 0.20$

The *IRR* of this project is 0.20 or 20 per cent. No other value of r can equate the *NPV* of the project with its cost.

The *IRR* or *MEI* is defined as ‘ *the rate of interest or return which renders the discounted present value of its expected future marginal yields exactly equal to the investment cost of project*’. In other words, ‘ *IRR* is the rate of return (r) at which the discounted present value of receipts and expenditures are equal’. The *IRR* of a project yielding a stream of returns over n years and involving different investment costs can be obtained by using the formula given in Eq. (21.9).

n

n

n

R

Cn

$\sum$

$=$

n

$\sum$

n

j 1

$= (1 + r)$

j 1

$= (1 + r)$

...(21.9)

n

n

R

C

or

n

n

$\sum$

$-$

n

$\sum$

$n = 0$

...(21.10)

j 1

$= (1 + r)$

j 1

$= (1 + r)$

The *IRR* criterion is basically the same as Keynes’s Marginal Efficiency of Investment (*MEI*).5 This criterion is theoretically superior to other criteria, though it has its own shortcomings. The *IRR* criterion says that so long as internal rate of return is greater than the market rate of interest, it is always profitable to borrow and invest. However, in a perfectly competitive market, a firm’s internal rate of return always equals the market rate of interest.

The Internal Rate of Return vs. Present Value

From Eq. (21.10) it may be inferred that *IRR* and *NPV* criteria lead to the same conclusion or yield the same decision. There are situations, however, where the two criteria give conflicting results. For example, suppose that a firm has to make a choice between project *A* and project *B*, each having a productive life of two years. The stream of net income6 at the end of the year from the two projects and their respective costs are presented in

Table 21.3.

5. However, modern economists treat *IRR* and *MEI* as different concepts. While *IRR* pertains to the capital investment by an individual firm, *MEI* pertains to the investment made by all the firms together. For a detailed discussion, see D.N. Dwivedi, *Macroeconomics: Theory and Policy* (Tata McGraw-Hill, New Delhi, 2003), pp. 160–62.

6. Net Income = Annual receipts less Annual cost.

Table 21.3 *Flow of Net Incomes*

	Cost of project	*Ist year*	*2nd year*
Project *A*	100	0	140
Project *B*	100	130	0

Let us now calculate the *NPV* for both the projects, assuming a 10 per cent expected rate of return, and compare the result with *IRR*. Remember that *NPV* = *PV*– *C*.

0

140

Project *A*:

PV =

+

= 115.70

2

(1+ 0.10) (1+ 0.10)

and

NPV = 115.70 – 100 = 15.70

Since *NPV* is positive (`15.70) at the expected rate of return of 10 per cent, Project *A* is acceptable. But if we raise the expected rate of return to 20 per cent, Project *A* will not be acceptable because at this rate of return, *NPV* is negative (– 2.78), as calulated below:

0

140

NPV =

+

−100

2

(1+ 0.20) (1+ 0.20)

= 97.22 – 100 = –2.78

130

0

Project *B*:

PV =

+

= 118.18

2

(1+ 0.10) (1+ 0.10)

and

NPV = 118.18 – 100 = 18.18.

Project *B* is acceptable at the rate of 10 per cent return since *NPV* which equals 18.18 per cent is positive. It will be acceptable even at the expected return or interest rate of 20 per cent since, in that case, *NPV* will bè8.33 calculated as follows.

130

0

NPV =

+

−100

2

(1+ 0.20) (1+ 0.20)

= 108.33 – 100 = 8.23

Having calculated the *NPVs* for Projects *A* and *B*, let us now calculate the *IRR* for both projects, for comparing the decisions.

By definition, the *IRR* is the rate of return (*r*) which renders the net present value (*NPV*) equal to zero. Using the definition (21.10), *r* for Project *A* may be calculatd as follows.

140

NPV = 0 +

−100 = 0

2

(1+ *r*)

By solving this equation, we can obtain the value of *r* a shown below.

140

NPV = 0 +

−100 = 0

2

(1+ *r*)

140

(1 + *r*)2 =

= 1.40

100

(1 + *r*) = 1.40 = 1.183

r = 0.183 or 18.3 per cent

Likewise, in case of Project *B*, the value of *r* can be obtainded as follows.

130

NPV =

+ 0 = 100

(1+ *r*)

130

(1 + *r*) = 100 = 1.30

r = 0.30 or 30 per cent

We find that *IRR* of Project *A* is 18.3 per cent and for Project *B* it is 30 per cent.

The *NPV* at different interest rates and the *IRRs* of Project *A* and *B* can be tabulated as given in Table 21.4.

Table 21.4 *NPV and IRR of Projects A and B*

Project A

Project B

r

NPV

r

NPV

0.0

40.00

0.00

30.00

10.0

16.70

10.00

18.18

18.3 = *IRR*

0.00

20.00

8.33

20.0

–2.78

30.00 = *IRR*

0.00

The conflict between the two criteria may be shown by plotting the information given in Table 21.4 as shown in Fig. 21.2.

The lines marked by Project *A* and Project *B* show relation between the various rates of return (*r*) and the corresponding *NPV* for Projects *A* and *B*. The two lines intersect at point *P*. The value of *r* at point *P* is 7.7 per cent.7 It shows that only at 7.7 per cent rate of return, both projects are equally acceptable. Below a rate of 7.7 per cent return, Project *A* is preferable because its *NPV* is higher than that of Project *B*. But above 7.7 per cent return, Project *B* is preferable because its *NPV* is higher than that of Project *A*. It follows that if a firm opts for Project *A* with higher

Fig. 21.2 *NPV and IRR*

7. The value of common *r* can be calculated as follows: Let *r* be such that *NPVA* = *NPVB*. We know that

140

130

NPVA = 0 +

2 – 100 and *NPV*

+ 0 – 100

(1 + *r*)

B = (1 + *r*)

In order to find the common *r*, let us equate the two equations.

140

130

0 +

2

(1 + *r*) – 100 =

+ 0 – 100

(1 + *r*)

Then

140

130

=

2

(1 + *r*)

(1 + *r*)

2

(1+ *r*)

140

=

(1 + *r*) 130 = ; or (1 + *r*) = 1.077

Thus,

r = 0.077, or 7.7 per cent.

NPV, it will earn a return less than 7.7 per cent and will have a longer pay-back period. Thus, the choice between the two projects will be based on the pay-off-period.

Furthermore, if firms evaluate the two projects on the basis of their *IRR*, Project *B* should be preferable since its *IRR* = *r* = 30 per cent is greater than that of Project *A* (with its *IRR* = 18.3 per cent). Obviously, the two criteria (*NPV* and *IRR*) produce conflicting conclusions in regard to the choice of projects. In actual practice, however, the firms are guided by their objective.

We have discussed so far the investment decisions under the condition of certainty.

In reality, however, businessmen are often required to take investment decisions under the condition of risk and uncertainty which will be discussed in the next chapter.

21.6 SOURCES AND COST OF CAPITAL

In the preceding section, we have discussed the demand for capital and investment decision criteria under the condition of certainty. These factors make the *demand side* of the capital market. *Supply side* of the capital market is an equally important factor that plays an important role in capital budgeting. The two important aspects of the supply side of the capital market are (*i*) the sources of capital, and (*ii*) the cost of capital. These aspects are discussed in this section.

21.6.1 Sources of Capital

Business firms meet their demand for capital from various sources which can be grouped under (1) internal sources, and (2) external sources. Let us now discuss the implications of capital raised from these sources and also the cost thereof.

(1) Internal Sources of Capital

The firm's own saving is its internal source of capital. Internal savings are normally generated in two ways: (*i*) by creating depreciation funds, and (*ii*) by ploughing back the profit or through the retained earnings. **Depreciation reserves** are created out of firm's earnings with a view to maintaining capital intact or for replacement of the worn-out capital. Also, firms raise their internal finance by retaining a part or total profit8 for reinvestment. Many large corporations rely mainly on their internal savings as a matter of policy. In the industrially advanced countries, the greatest proportion of corporate investment is financed out of internally acquired funds. Preponderance of this practice may be ascribed to the following factors.

(*i*) Firm's desire to display their assets and efficient management;

(*ii*) To avoid heavy cost of raising funds through issue of bonds and stocks, etc.;

(*iii*) To avoid uncertainty in raising funds from the market, and

(*iv*) To enhance capital gains by increasing the market value of company's stock.

Where ploughing-back profit is a permanent feature of investment policy, it becomes an integral part of capital budgeting.

The main managerial task in regard to raising internal funds are (*a*) forecasting the availability of internal funds; (*b*) determining the depreciation reserve and plough-back profit; and (*c*) deciding the amount for long-term investment.

8. The proportion of the total divisible profit retained for the purpose of reinvestment depends on the decision of shareholders.

(2) External Sources of Capital

The external source of capital is the capital market. The following are the ways and means of raising funds from the capital market.

(*a*) Sale of bonds. The firms may borrow funds directly from the capital market by selling some kind of bonds, e.g., mortgage and debenture bonds. Raising funds through the sale of bonds has certain advantages: (*i*) the total interest on bonds and debentures is permissible as business cost under corporate income-tax laws, and (*ii*) it increases the share of stockholders in case the company is making higher profits as bondholders get only a fixed rate of interest. However, heavy financing through bonds weakens the position of company's stock. For this reason, conservative managerial groups avoid heavy bond-financing.

(*b*) Issue of new common stock (or equity shares). Another common method of raising funds from the market is to issue common stocks. The extent to which a company can raise finances by issuing new shares depends, among other things, on the choice and preference of old stockholders. They may object to new issues if they apprehend loss or 'dilution' of their power to control the company. Besides, issuing new stocks involves heavy transaction cost, mainly on account of fulfillment of legal technicalities. It has disadvantage also in regard to tax liabilities. From a stockholder's point of view, small stockholders hold a position inferior to bondholders.

(*c*) Issue of preferred stocks. The third method of capital accumulation from the market is the sale of preference stock. The main distinction between preference and common stock is that preferred stockholders get preference over common or equity holders in the payment of dividend. But, bondholders get preference over the preferred stock-holders.

(*d*) Convertible securities, direct loans, etc. The other methods of borrowing from the market are (*a*) direct borrowing from the financial institutions such as commercial banks, insurance companies, industrial finance corporations, Industrial Credit and Investment Corporation of India (*ICICI*), Industrial Development Bank of India (*IDBI*) and the Unit Trust of India (*UTI*), etc. and (*b*) issuing hybrid securities such as convertible bonds which can be traded in the market like equity shares, under certain restrictions; this is an advantage too for the security holders. Companies borrow funds though convertible securities, in case they prefer somewhat safer type of securities. A relatively recent method of capital accumulation is raising capital through public deposits.

21.6.2 Cost of Capital

As mentioned above, capital is a scarce and productive commodity. Since every scarce and useful commodity has a price, capital has a price too, termed as *cost of capital*. The cost of capital consists of explicit and implicit cost. **Explicit cost** of capital is the interest paid on it whereas **implicit cost** is the *opportunity cost*, expected return from the second best use of money capital. In the strict sense of the term, implicit cost of capital is the *opportunity cost* of money capital.

The cost of capital plays an important role in capital budgeting decisions. Since capital is available at a cost, it impels the managers to make best possible use of available funds. Besides, cost of capital can be an important decision variable in the selection or rejection of projects: if cost of capital is less than the return on its investment, it is an acceptable proposition.

For estimating the cost of capital, different concepts of cost are used. In this section, we discuss briefly the various concepts of cost of capital used in the analysis of capital budgeting and also the methods of computing cost of different kinds of capital.

21.6.3 Concepts and Measures of Cost of Capital

(*i*) Cost of Debt-Capital. Debt capital refers to the funds directly borrowed from the market through public deposits, bonds and debentures. The cost of debt capital may be defined as the rate of return that must be earned on the borrowed capital to keep the earning of common stockholders unchanged. Since payment of corporate income tax affects the cost of capital, a distinction is made between before-tax and after-tax cost of capital. Before-tax cost of debt capital is the same as interest rate on the borrowed amount. It implies that if a firm borrows a certain sum and lends the same, its before-tax interest earning will be equal to its interest payments.

Since interest paid on borrowed capital is deductible from the taxable income, cost of debt capital is reduced to the extent of tax saving. After-tax cost of debt capital = (interest rate) × (1 – tax rate). For example, if tax rate is 40 per cent and interest rate is 10 per cent, the after-tax cost of capital can be worked ou as follows.

$$\text{Cost of capital} = 10\left(1 - \frac{40}{100}\right) = 6 \text{ per cent}$$

(*ii*) Cost of Preferred Stock. Conceptually, the cost of preferred stock is similar to the cost of debt-capital. It may thus be defined as the rate of return on preferred stock that must be earned to keep the earning available to the common stockholders unchanged. The cost of preference stock or the required rate of return can be obtained by dividing the dividend per preferred stock by market price of the new preferred sock issues.

$$\text{Cost of Preferred Stock} = \frac{\text{Dividend per Share}}{\text{Market price of New Issue}}$$

Dividend paid on preferred stock is not deductible from income assessable for taxation. Therefore, after-tax and before-tax cost of preferred stock are the same.

(*iii*) Cost of Common Stock or Equity Capital. The cost of equity capital may be defined as the minimum rate of return on the projects financed through the sale of common stocks that can keep the market value of issues unchanged. That is, the opportunity cost of old issues must be equal to the rate of return on them, the degree of risk remaining the same. The cost of equity capital (*CEC*) is estimated by the following formula:

Dividend per Stock

CEC =

+ Risk premium

Market price of Stock

For example, suppose common stock of a firm has a market price of `25 per share and the current dividend is `2.50 per common stock. Suppose also that the risk premium or what is also known as expected growth rate of the firm is 5 per cent. Then the cost of equity (*CEC*), i.e., the *required minimum rate of rturn*, can be calculated as follows.

CEC = 2.5 + 0.05 = 0.1 + 0.05

25

= 0.15 or 15 per cent

It means cost of equity is 15 per cent. There is another method of estimating cost of equity capital, known as the *Dividend Valuation Method*. This method is similar to the *Present Value* method and cost of equity (*CEC*) is similar to *IRR*. In simple words, *CEC*

is the rate of discount which equalizes the discounted present value of expected dividends (*D*) to the present value of share. If equity-holders expect a constant dividend (*D*) per share over *n* years, then the present valu of share (*PVS*) can be obtained as

D

PVS = (1+ *r*) *n*

Here, *r* is the cost of equity, i.e., *r* = *CEC.*

(*iv*) Cost of Retained Earning. The capital raised through the retained earnings is not free of cost. The cost of retained earnings is simply the expected return from the re-investment of dividends. In other words, the cost of retained earnings is its **opportunity cost**, i.e., the income forgone by the shareholder which might have been obtained through the next alternative investment of dividend. An important modification, however, has to be made with respect to tax on dividend, because it is only the dividend *minus* tax that can be reinvested not the total dividend. The cost of retained earning or its opportunity cost may thus be estimated as

Dividend – Tax

Cost of Retained Earnings = Price of Share × 100

where dividend and tax are calculated on a per share basis.

(*v*) *Weighted Cost of Capital* We have discussed the cost of different sources of capital separately. In general, however, firms acquire their finances from different sources for at least two reasons: (*i*) firms acquire cheaper external finance with a view to preserving their internal potentials for financing future projects, especially if cost of external capital is expected to increase, and (*ii*) they might like to maintain a balance between merits and demerits of different sources of finance. Firms may, therefore, decide to acquire capital from different sources. Under this pattern of financing, the cost of any single source of capital is not an appropriate decision variable. What matters is the **weighted average cost of capital**. The weighted average cost of capital may be calculated as shown in Table 21.5. Let us assume that the percentage distribution of the total stock of capital financed through different sources of capital is given in column 2 and the corresponding interest rate is given in column 3. Percentage share of each type of capital is its weightage (*W*) in the total stock of capital. As calculated in the Table 21.5, the weighted (average) cost of capital comes to 7.0 per cent.

Table 21.5 *Weighted Average of Cost of Capital*

Form of capital	*Percentage share in the total capital (W)*	*Cost of capital (C) (%)*	*Total cost of capital = C × W*
(1)	*(2)*	*(3)*	*(4)*
Debt Capital	30		

6.00

180

Preferred Stock

20

7.50

150

Equity Stock

25

8.00

200

Retained Earnings

25

6.80

170

Σ *W* = 00

Σ *CW* = 700

CW

Σ

Weighted Cost of Capital =

700

=

= 7.0%.

W

Σ

100

Concluding Remarks

Capital budgeting is not only one of the most important tasks of business management, but also a complicated procedure. Manager's skill, experience, intuition and foresight are perhaps needed more in taking appropriate investment decisions than in any other aspect of business management. The managers should, therefore, clearly define the objective and targets, find out the most suitable projects of investment relevant to the firm's objective, examine the feasibility and profitability of the projects, select the projects on the basis of appropriate investment criteria and determine their demand for capital. They should next examine the cost of various sources of capital, accumulate the necessary funds and then, finally, take up the investment projects.

SUMMARY

- Capital budgeting is a process of conceiving, analysing, evaluating and selecting the most profitable project for investment. In economic sense, investment means acquiring assets that yield returns over a period of time.
- Optimum level of capital is determined at the level at which marginal revenue productivity *(MRP)* of capital is equal to marginal cost (*MC)* of capital.
- As regards the choice of investment projects, under the condition of certainty, investment project is selected on the basis three important criteria: (*i*) pay-back period, (*ii*) present value method, and (*iii*) internal rate of return.
- *Pay-back period* is determined by dividing total investment by annual gross return. If there are options for investing in more than one project, then projects are listed in order of their pay-back period and selected accordingly.
- *Present value* of future return from an investment is estimated by discounting the return by the market rate of interest and compared with original investment. If discounted value of return is equal to or greater than the investment, then investment project is acceptable.
- *Internal rate of return (IRR)* is the rate of return (or interest) which renders the discounted present value of an expected annual return exactly equal to the marginal cost of investment. In case of a one-year project, the *IRR* is the rate return at which the discounted present value of receipts and payments are equal.
- Sources of investment funds are (*i*) internal sources plough back profits, and (*ii*) external sources including bonds, shares, and loans.

• Cost of internal capital is opportunity cost and cost of external capital is generally the market rate of interest.

REVIEW QUESTIONS

1. What is meant by capital budgeting? What is the importance of capital budgeting in long-term investment decisions? What are the prerequisites of capital budgeting?

2. What is meant by optimum level of capital? How is the optimum level of capital determined?

3. Suppose a firm has five investment projects—A, B, C, D and E—and their costs and expected return in the following table.

Table: *Project Cost and Expected Returns*

Project	*Cost of Project (`)*	*Expected returns (%)*
A	5,000	12
B	10,000	8
C	6,000	10
D	8,000	6
E	4,000	4

From the data given in the table (*i*) derive the demand curve for capital and (*ii*) find the demand for capital at the fixed interest rate of 8% per annum.

4. What criteria are used for making choice of investment projects? Explain the pay-back period criterion and its limitations as a choice criterion.

5. Explain the concept of the present value of a future income. Explain why it is necessary in an investment decision to discount the future income stream.

6. Suppose an investment yields an income of `500 in the first year, `1000 in the second year and `1500 in the third year. If rate of interest is 10%, what is the present value of this income stream?

7. Define and explain the concept of marginal efficiency of capital. What is its significance in investment decisions? Explain and illustrate the use of marginal efficiency of capital in determining the optimum level of investment.

8. Suppose an investment project costs `5,000 and yields an annual income of `2,500 for a period of three years. Find the marginal efficiency of capital.

9. What is meant by time value of money? How is the present value of an income expected in future calculated?

10. What is meant by the net present value of future returns? Explain the use of net present value in making investment decisions.

11. What is internal rate of return? What is the method of working out internal rate of return? Suppose an investment project costs `5,000 and yields an annual income of `2,500 over three years, find the internal rate of return.

12. Suppose a company is considering an investment of `50 million. The market rate of interest is 10 per cent and anticipated marginal efficiency of investment is 12 per cent. How will the company react to each of the following changes in the conditions?

(*a*) Market rate of interest increases to 14 per cent.

(*b*) Market rate of interest increases to 12 per cent.

(*c*) *MEC* decreases to 9 per cent due to increase in operational cost.

13. Suppose an investment costing `500 million is expected to yield returns of `300 million in the first year, `200 million in the second year, and `100 million in the third year. It has no scrap value after the third year. Find (*a*) the present value of the total return at 10% interest rate and

(*b*) whether project is acceptable at 10% interest. Also, compare the internal rate of return and the market rate of interest and find whether the project is acceptable.

14. What is meant by cost of capital? What are the different sources of capital available to a corporate firm?

15. How is the cost of equity and debt capital calculated? What factors are taken into account while determining an appropriate combination of equity and debt capital?

16. Why do the firms use different sources to acquire capital from the capital market? How is the weighted cost of capital calculated? Illustrate your answer through an example.

FURTHER READING

Baumol, W.J., *Economic Theory and Operations Analysis* (Prentice-Hall of India, New Delhi, 1985, 4th Edn.), Ch. 25.

Bierman, Harold Jr. and SmitH Seymour, *The Capital Budgeting Decisions* (Macmillan Company, New York, 1975).

Bierman, Harold and S. Smidt, *Capital Budgeting Decision* (Macmillan, New York, 1984).

dean, Joel, *Capital Budgeting* (Columbia University Press, New York, 1951).

Jean William, H., *Capital Budgeting: The Economic Evaluation of Investment Projects* (International Textbook Co., Scanton, Pa., 1969).

manSfield, e., *Managerial Economics*, (W.W. Norton and Co., Scanton, Pa., 1969.)

Salvatore, d., *Managerial Economics* Oxford University Press, New Delhi, 2012.

Soloman, ezra, *The Theory of Financial Management* (Columbia University Press, 1964).

CHAPTER

22 Investment Decisions under Risk and Uncertainty

CHAPTER OBJECTIVES

Having discussed the investment decisions under certainty in the preceding chapter, we proceed in this Chapter to discuss investment criteria under the conditions of risk and uncertainty. By going through this chapter, you will learn the following aspects.

- What is meant by risk and uncertainty
- How risk and uncertainty are measured
- What investment criteria are used for investment under the condition of risk
- What is probability theory approach to investment decisions
- How decision tree map is prepared and decisions taken
- What are different approaches for investment decisions under uncertainty

22.1 INTRODUCTION

In Chapter 22, we have discussed investment decisions under the condition of certainty, i.e., most of the parameters and market conditions are known to the investor or can be predicted. In reality, however, 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 chapter, however, we concentrate on the widely used methods of investment decision-making. The investment decision techniques are discussed under (*i*) investment decisions under risk, and (*ii*) investment decisions under uncertainty. Let us begin with the concepts of and distinction between risk and uncertainty as applied to business decision-making.

22.2 CONCEPTS OF RISK AND UNCERTAINTY

The concept of risk and uncertainty can be better explained and understood in contrast to the concept of certainty. The concept of certainty has already been defined in Chapter 21. Here, let us recall the concept of certainty as it will help in understanding 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

MC

outcome of their investment decisions. For example, if you deposit your savings in 'fixed deposit' bearing 10% interest, you know for certain that the return on your investment in time deposit is 10%, and *FDR* can be converted into cash any day. Or, if you buy government bonds, treasury bills, etc. bearing an interest of 11%, you know for sure that the return on your investment is 11% per annum, your principal remaining safe. In either case, you 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*.

In reality, 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 and 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 or otherway 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 the 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.

22.2.1 Meaning of 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 it can be reliably estimated. For example, if a company doubles its advertisement expenditure, there are four probable outcomes: (*i*) its sales may more-than-double, (*ii*) they may just double, (*iii*) increase in sales may be less than double, and (*iv*) sales do not increase at all. The company has estimated the probabilities of the four possible outcomes on the basis of its past experience as (*i*) more-than double – 20% (or 0.2), (*ii*) almost double – 40% (or 0.4), (*iii*) less-than double – 50% (or 0.5) and (*iv*) no increase – 10% (or 0.1). It means that there is 80% risk in expecting more-than-double increase in sales, and 60% risk in expecting doule increase in sales, and so on.

There are two approaches to estimating 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. In case of *a priori probability*, we know that when a coin is tossed, the probabilities of 'head' or 'tail' are 50:50, and when a dice is 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'.[1]

22.2.2 Meaning of 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 nor can it be reliably 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, rate of buildings catching fire, and so on, but it is not possible to predict the death of a particular insured individual, a particular car meeting an accident or a particular house catching fire, etc.

The long-term investment decisions involve a great deal of uncertainty with unpredictable outcomes. 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:

(*a*) complete ignorance, and

(*b*) partial ignorance.

In case of *complete ignorance*, investment decisions are taken by the investor 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

1. For these statistical techniques, refer to a standard text on statistics.

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.

22.3 INVESTMENT DECISIONS UNDER RISK

The concept of risk has already been defined in the preceding section. In this section, we discuss the methods of making investment decisions under the condition of risk.

22.3.1 The Pay-Off Matrix Method

A pay-off matrix2 is 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, let us understand the terms— *strategy, states of nature* and *pay- offs*—used in pay-off matrix method of investmnet decision-making.

Strategy. Strategy means making strategic choice of one from several alternative actions that can be taken to achieve a certain goal. For example, given the profit maximization objective, firm conceives the following actions:

(*i*) reducing cost of production through scale expansion,

(*ii*) making aggressive advertisement for increasing sales,

(*iii*) product diversification,

(*iv*) product innovation,

(*v*) cutting down the price for 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 market conditions.

State of Nature. The ***states of nature*** refer to the future market conditions on which the firm has no control. The future market conditions might arise on their own or may be the result of actions taken by the firm. The market 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 firms 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.

2. 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.

Estimation of Pay-offs. **Estimating the pay-off of different strategies objectively** 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 22.1. 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 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 22.1.

Table 22.1 *The Pay-off Matrix*

	States of Nature			
Strategy	*N1*	*N2*	*N3*	*N4*
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 conditions of uncertainty, 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 are very large, there will be a large number of probabilities. Then, according to the Bayesian rule,

$$P_1 + P_2 + P_3 + \dots + P_n = 1.$$

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.

$$E(R) = R_1 P_1 + R_2 P_2 + R_3 P_3 + \dots + R_n P_n = \sum_{i=1}^{n} R_i P_i \qquad \dots(22.1)$$

where $E(R)$ = expected value or weighted mean of pay-offs; $R_1, R_2 \dots R_n$ are expected pay-offs; $P_1, P_2 \dots P_n$ are the probabilities.

Let us suppose that subjective probability distribution corresponding to the pay-offs given in Table 22.2 is given in Table 22.2.

Table 22.2 *Subjective Probability Distribution*

	States of Nature			
Strategy	*N1*	*N2*	*N3*	*N4*
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 ob tained as follows.

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) + 16 (0.20) + 4 (0.10) – 2 (0.10)	= 12.40

S_3 $16 (0.40) + 8 (0.40) + 6 (0.15) – 1 (0.05) = 10.15$

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 (14.80) of the alternative strategic investment options available to the deci sion-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.

22.3.2 Risk Adjusted Discount Rate Method

The use of *risk-adjusted discount rate* is one of the methods used for *adjusting the valuation models* 3 *for risk*. A common, though crude, method of accounting for the risk factor in invest ment decisions is to use a *risk-adjusted discount rate* in the formula for calculating the *net present value* (*NPV*). The risk-adjusted dis count rate (*d*) is defined as

$$d = \frac{1}{1 + r + \mu} \quad \ldots(22.2)$$

where *r* is risk free discount rate, and *u* 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 calculate the *NPV*. The formula for calculating the *risk adjusted NPV* for *nth* year is given as

$$\text{Risk-adjusted } NPV = \sum_{t=1}^{n} R_n d^n - C_0 \quad \ldots(22.3)$$

where R_n = return in the *n* th year, and C_0 = *origi nal* cost of capital.

3. The *valuation models* refer generally to *present value* and *net present value* models.

By substituting Eq. (22.2) for *d* in Eq. (22.3), we get,

$$\text{Risk-adjusted } NPV = \sum_{t=1}^{n} \frac{R_n}{(1 + r + \mu)^n} - C_0 \quad \ldots(22.4)$$

Given the model for computing risk-adjusted *NPV* in Eq. (22.4), it 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 cashflow or return of `132 million after one year. Assuming a discount rate of 8% (or 0.08) and risk probability of 0.12, the risk-adjusted *NPV* can be easily calculated as shown below.

Risk-adjusted *NPV* = 132

`

million − 100

`

million

1 + 0.08 + 0.12

= `10 million

This calculation shows that the project gives a risk-free *NPV* = `10 million, which is, incidentally, a 10% risk-free return. If we work out *NPV* without the risk-factor of 0.12, we find that the *NPV* comes tò22.22 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.

Merits 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 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 automati cally 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 20th year will be

1

d =

20

$(1+ r + \mu)$

This method reduces the *NPV* of a sum expected after 20 years to zero.

Third, this method is extremely straightforward and easy to handle.

However, a major **weakness** of this method is that it does not provide any method of estimating the risk factor. Therefore, this method has to be supplemented with a technique of estimating the risk-factor. Or else, the decision-makers will have to use their own intuitive discretion.

22.3.3 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 proce dural 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 ap proach*, on the other hand, it is the expected return (*R*) which is adjusted for the risk factor. In the

procedure of estimating certainty equivalent *NPV*, 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 amount of return. The formula for *certain ty-equivalent approach* is given as follows.

α *R*

Certainty-equivalent *NPV* =

n

− *C*

...(22.5)

n

0

$(1+ r)$

where α is the risk-equivalent factor.

In fact, the co-efficient a in Eq. (22.5) is called *certainty- equivalent coeffi cient*. The use of the coefficient (a) makes an expected risky-return (*R* n*) equivalent to an expected certain return (*Rn*). The *certain ty- equivalent coefficient* (a) is worked out as the ratio of *Rn* to *R* n*. That is,

Expected certain return

R

a =

n

=

...(22.6)

Expected risky return

R * *n*

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 one year. Project *B* is risk-free and is expected to yield a sum of `40 million after one year. The investor treats both the projects as equally good. It means that he treats a risk-freè40 million as an equivalent of risky `50 million. In that case,

`40 mm

a =

= 0.8

`50 mm

Here the, *certainty equivalent co-efficient* equals 0.8 and risk factor is 0.2. The value of a varies between 0 and 1, and degree of risk moves in the reverse direction, i.e., the smaller the value of a, the greater the degree of risk. For example, if a = 0, it means that the expected sum is totally risky or the degree of risk is 100 per cent. And, if a = 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 consid ering 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 (a) = 0.8. By certainty-equivalent approach, the choice between the two pro jects can be made by comparing their certainty-equivalent *NPV*.

Project *A*

Rn

44

Certainty-equivalent *NPV* =

− *C* =

− 30

0

(1 + *r*)

(1+ 0.1)

= `10 million

Project *B*

Certainty-equivalent *NPV* =

R

α *n* – *C*

(1 + *r*) *n*

0

(0.8)55

= (1+0.1) – 28 = `12 million

The calculation of certainty equivalent *NPV* of projects *A* and *B* 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*.

22.3.4 Probability Theory Approach

In contrast with the methods discussed above, the *probabili ty theory method* considers a whole range of possible risk-return combinations which represent adequately the full range of alternative outcomes of a risky undertaking. 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. 22.3.

Table 22.3 *Alternative Cash-Flows and Associated Probabilities*

Alternative

1st Year

			2nd Year		*3rd Year*	
Cash flows (`)	*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 arè1850 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 in surance 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.

Probability Theory Approach and Portfolio Selection:

Markowitz 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 instru ments. The application of this method to the problem of portfolio selection is an exemplary case. In his approach, Morkowitz has utilized 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.4 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

4. Expected earnings can be obtained by extrapolating the returns on the basis of past earnings of similar securities or on the basis of expert opinion.

curve, showing the various rates of return and the associated risk, as shown by the curve *RP* in Fig. 22.1. This curve shows the various combinations of return rates and the associated risk.

The curve *RP* is called the ***risk-earning possibility curve***. The various points on the curve *RP* represent, in a sense, the opti mal 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

Fig. 22.1

the lower part of the curve, e.g., point *L*, and

Risk-Earning Combination

those who prefer higher risk and higher return would choose point *H*. This pattern of portfolio selection does not provide an optimal combi nation of stocks and securities.

The optimal combination of portfolio can, however, be deter mined 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. Con trary 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 curve. The optimal combination is shown by point *T* in Fig. 22.1. 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.

22.3.5 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.

First, the decision-makers are required to make a choice (or series of choices) from the alternative investment avenues avai lable to them. They are not supposed to leave the matter undecid ed.

Second, the decision-makers know for sure that all the decisions will yield a positive outcome, but they cannot predict 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 are the outcomes.

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

Decision Tree for Investment Decision

Fig. 22.2

under these conditions is called decision tree. *A **decision tree** is a graphi cal 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 Fig. 22.2, presents the entire gamut of decision options and possible outcomes in the form of a diagram and, thereby, guides the decision-makers to a ra tional decision.

The decision tree method is illustrated by a simple hypothetical example in Fig. 22.2. Suppose an investor visualizes two viable Projects *A* and *B*. Project *A* costs `500 million and Project *B* costs `400 million. The prospective yields (cashflows) 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 or low, de pending on the market conditions and consumers' perception and prefer ence. The demand conditions determine the expected returns, the cashflows from both the projects. The investor has the informa tion about the expected cashflows under the different states of economy (i.e., high, medium or low growth of the economy) and the demand prospects. All the possibilities are presented graphically in Fig. 22.2.

The decision tree 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 econo my. 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 cashflows 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 cashflows. 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 Fig. 22.2, 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*.

22.3.6 Simulation

We have discussed so far the methods of making investment deci sions under a given probability of risk. In real life situa tions, however, business decision-making is much more complex. A large number of unpredictable factors have to be taken into account, especially where investment involves a huge sum and the project has very long life. For example, suppose a car manufacturing company, announces its plan to set up another production plant to produce a smaller car for low-income group of people. The number of considerations that may have figured in its decision-making process include size and cost of the plant, life of the plant, future demand pros pects, market saturation point, competition by the rival firms, optimum level of production, probable price, input prices (labour and material costs), expected rate of re turn, labour problem, selling costs, prospects for increasing price, price elasticity, profitable market share, government policy, sales tax, entry of new competing brands of car, 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 variables 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 *probabil ity distribution* of each factor involved in decision-making. Probability distribution can be generated by company's own market research department, economic research centres and other published and non-published information. 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 cas h 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 distribu tion 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 generates a rate of return for every variable. In this process, the computer builds a frequency distribution of the rates of return on different set of condi tions. The decision makers can choose one or many from the options generated by the simulation technique.

Although simulation is a versatile and very powerful tech nique, a full-scale simulation is very expensive and time consum ing technique. Therefore, this technique can be used only by large corporations where a huge investment is involved and divi sion making is beset with complexities. This technique is, however, not feasible for general use.

22.3.7 Sensitivity Analysis

Sensitivity analysis is a simple version of a full-fledged simu lation technique and is also inexpensive. While simulation technique uses the entire range of probability distribution 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 cashflow) asso ciated with the highest probability of each variable. After estimating the target variables, chang es in the decision variables are introduced to find their effect on the target variable, i.e., 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 deci sion-makers in eliminating unimportant variables and in concen trating on the most important ones, i.e., the variables to which the rate of return is highly sensitive. This is called *sensitivity analysis*.

22.4 INVESTMENT DECISIONS UNDER UNCERTAINTY

In the preceding sections, we have discussed the techniques that are used in invest ment decision-making under the condition of *risk*. In this sec tion, we discuss the techniques of investment decisions under *uncertainty*. As defined in section 22.2, *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 neces sarily *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.

22.4.1 Wald's Maximin Decision Criterion

Wald's *maximin decision criterion* says that the decision-makers should first specify 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 apply ing it to our example given in Table 22.1 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 22.4. row-wise. The maximin column presents the worst outcome of each strategy. The best or the highest outcome out of the worst outcomes is 5 of strategy *S* 1. Going by the maximin criterion, the decision-makers would accept strategy *S* 1.

Table 22.4 *Application of Maximin Criterion*

	States of Nature				
Strategy	*N1*	*N2*	*N3*	*N4*	*Maximin*
S 1	20	12	6	5	5
S 2	15	16	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.

22.4.2 Minimax Regret Criterion: The Savage Decision Criterion

Minimax regret criterion is another decision rule under the condition of 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 measured by the di fference between the pay-off of a given strategy and the pay-off of the best strategy *under the same state of nature*. " [5] Thus, regret is the cost of a wrong decision. Suppose an investor has three strategies for investment, *S* 1, *S* 2 and *S* 3, giving re turns 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 we have used in section 22.2, Table 22.1. 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, ... *Sn*) and all the states of nature (*N* 1, *N* 2, ... *Nn*), we get a regret matrix as shown in Table 22.5. From the regret matrix, we can find 'maximin regret' by listing the maximum regret for each strategy, as shown in the last column. The column 'maximin regret' shows that maxi mum regret is minimum (3) in case of strategy *S* 4. According to maximin criterion, therefore, strategy *S* 4 should be selected for investment.

Table 22.5 *Pay-off Matrix and Regret Matrix*

	States of Nature				*Regret Matrix*				*Maximin*
Strategy	*N1*	*N2*	*N3*	*N4*	*N1*	*N2*	*N3*	*N4*	*Regret*
S 1	20	12	6	5	0	0	0	0	0
S 2	15	10	4	– 2	5	2	2	7	7
S 3	16	8	6	– 1	4	4	0	6	6
S 4	5	12	3	2	15	0	3		

3

3*

22.4.3 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 Hur wicz 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_i) is given below.

$D_i = \alpha \text{ Max } i + (1 - \alpha) \text{ Min } i$

where D_i = decision index of the ith strategy; and a = probability of maximum pay-off.

5. Salvatore, D., *Managerial Economics* (Oxford University press, 2012). p. 582.

The construction of Hurwicz *decision index* is illustrated in Table 22.6. 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 a (where a is subjective probability of pay-off). Note that the same probability applies to all the strategies. Columns (5), (6) and (7) give similar values of minimum pay-offs of the four strat egies. The last column (8) gives the *decision index*.

Table 22.6 *Hurwicz Decision Index*

Strategy	*Max*	a	a *Max*	*Min*	*(1* – a *)*	*(1* – a *) 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.

22.4.4 Laplace Decision Criterion: The Beyesian Criterion

The Laplace criterion uses the Bayesian rule to calculate the *expected value* of each strategy. As mentioned earlier, 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' as defined in the case of pay-off matrix in section 22.2.1. Once the expected value of each strategy is worked out, then the strategy with the highest expected value is selected.

The Laplace decision rule avoids the problem that arises due to subjectivity in assuming a probability of pay-offs. This criterion is, there fore, regarded as the criterion of rationality because it is free from a decision-maker's attitude towards risk.

To sum up, uncertainty is an important factor in investment decisions but there is no unique method of dealing with uncer tainty. There are several ways of making investment decisions under the condition of uncertainty. None of the methods, as described above, lead to a flawless decision. However, 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.

SUMMARY

• Risk refers to a situation in which a business decision is expected to yield more than one outcome and probability of each outcome is known or can be estimated.

• Uncertainty refers to a situation in which there is more than one outcome of a business decision and probability of no outcome is known nor can be estimated.

• Under the *condition of risk*, there are five methods of taking investment decisions: (*i*) pay-off matrix method, (*ii*) risk-adjusted discount rate method, (*iii*) certainty equivalent approach, (*iv*) the probability theory approach, and (*v*) decision-tree method.

• Under pay-off matrix method, the investor uses a strategy of working out a matrix of returns from different options under different degrees of probability and selects the best option for investment.

• Under risk-adjusted discount rate method, the investor has to work out the risk-adjusted returns from different investment options by using a risk-adjusting formula and has to find the best option for making investment.

• Under certainty equivalence method, a certainty-equivalence coefficient is measures by using a formula, i.e., by dividing expected certain return by expected risky return. The estimated certainty-equivalence coefficients from different investment avenues are compared to take the investment decision.

• The probability theory approach takes into account the whole range of risk-return combinations of various investment options. Under Morkowitz method, based on estimates of risk-return combination, a *risk-return possibility curve* and *risk-return indifference map* are drawn and the best option is selected.

• Another method of taking investment decision is *decision tree method*. Under this method, the investor works out the possible returns from all available possible options and presents it graphically, called decision tree. The decision tree disclosed the best option for investment decision.

• Under the *condition of uncertainty*, three criteria are applied for investment decisions: (*i*) minimax regret criterion, (*ii*) Hurwicz decision criterion, and (*iii*) Laplace decision criterion. (Since these methods are technical, need to be studied from the text).

REVIEW QUESTIONS

1. Define the concepts of risk and uncertainty. How does uncertainty create a different situation for investment decision-making compared to risk?

2. What is a pay-off matrix? What kind of data is required to construct a pay-off matrix? How does it help in investment decision-making?

3. Suppose an investor has a certain amount to invest and considers four strategic actions for investment, *S* 1, *S* 2, *S* 3 and *S* 4.

The expected pay-offs of the

Strategy

States of economy

four strategies under three

High growth

Low growth

Recession

different states of the economy

S 1

30

18

8

(high growth, low growth and

S 2

20

25

2

recession) are given in the

S 3

25

15

–5

following Table.

S 4

10

15

6

4. Explain the risk adjusted

discount rate method of Construct a pay-off matrix and find the most acceptable strategy.

investment decision. What is the purpose of using risk-adjusted discount rate? Suppose a return of `1000 is expected two years hence; *IRR* = 0.1; and risk probability = 0.08. Find the risk-adjusted present value of the expected return.

5. Suppose an investment project costing `100 million is expected to yield a return of `115 million after a period of one year. If *IRR* = 0.08 and risk probability = 0.12, find the risk-adjusted *NPV*. Is the project acceptable on these conditions? Suppose expected return increases tò125 million, will it affect the investment decision?

6. What is a decision tree? What are the basic elements and steps in the construction of the decision tree? How does a decision tree help in investment decision-making?

7. Explain probability theory approach to investment decision-making. What are the limitations of this approach?

8. What role does a decision tree play in business decision-making? Illustrate the choice between two investment projects with the help of a decision tree assuming hypothetical conditions about the states of nature, probability distribution and corresponding pay-offs.

9. Define risk-return possibility curve and risk-return indifference curve. Illustrate graphically investment decisions with the help of these curves.

10. What is certainty equivalent approach to investment decision-making? How is this method different from risk-adjusted discount rate approach?

11. How is investment decision under the condition of uncertainty different from that under the condition of certainty? Explain the minimax regret criterion of investment decision.

12. Write short notes on the following.

(*a*) Hurwicz Decision Criterion

(*b*) Laplace Decision Criterion

(*c*) Wald's Maximum criterion.

FURTHER READING

Baumol, W.J., *Economic Theory and Operations Analysis* (Prentice-Hall of India, New Delhi, 1985), Ch. 25.

Brigham, E.F., and J.L., PaPPas, *Managerial Economics* (Dryden Press, 1976), Ch. 3.

hErtz, DaviD, B., "Risk Analysis in Capital Investment", *Harvard Business Review*, Jan.-Feb. 1964.

JEan, William h., *Capital Budgeting: The Economic Evaluation of Investment Projects* (International Text Book Co., Scanton, Pa., 1969).

magEE, John, F., "How to Use Decision Trees in Capital Invest ment", *Harvard Business Review*, Vol. 42, 1964.

mansFiElD, E., *Managerial Economics: Theory, Applications, and Cases*, 1996 (W.W. Norton & Co., New York), Chs. 14 and 15.

massE, PiErcE, *Optimal Investment Decision, Rules for Action and Criteria for Choice* (Prentice-Hall, Englewood Cliffs, N.J., 1962).

salvatorE, D., *Managerial Economics*, 2012 (Oxford University Press, New Delhi), Chs. 12–14.

Part VI

MACROECONOMIC ASPECTS OF MANAGERIAL ECONOMICS

CHAPTER

23 Macroeconomics: Definition, Nature, Scope and Importance

CHAPTER OBJECTIVES

The objective of this chapter is to introduce macroeconomics and highlight its relevance to managerial economics. This chapter makes you understand the following aspects of macroeconomics.

- Definition of macroeconomics
- Nature of macroeconomics—theoretical and policy orientation
- Scope of macroeconomics—the subject matter of macroeconomics
- Importance of macroeconomics from business decisions point of view

23.1 INTRODUCTION

We have so far discussed the microeconomic aspects of managerial economics. What we have discussed so for is essentially *managerial microeconomics* 1. The subject matter of managerial microeconomics consists of theories of demand and supply, theory of production, cost-output relationships, market mechanism, theory of price determination under different kinds of market structure, pricing principles and prices, and capital budgeting and investment decisions. Managerial decisions on these issues, except on consumers' demand, fall within the purview of the decision powers of the managers. Most micro level managerial decisions are made generally with short-run perspective assuming that the business environment of the economy would continue to remain the same and changes therein, if any, are predictable and can be taken into account while making business decisions.

However, managerial decision-making, especially in case of large and medium scale business corporations, does not remain always confined to short-run and internal microeconomic issues. In case of large and medium scale business firms, managerial decisions are taken with a long-run perspective of economic conditions and business environment of the country. For, economic conditions and business environment in most developed and developing countries do not remain the static. Economic environment continues to change—some times positively, some times negatively; some times slowly, some times rapidly. For instance, Indian economy had grown at 8 per cent in 2003–04, but growth rate declined in the subsequent years and

1. In fact, the textbooks on Managerial Economics by the foreign authors indicate that, in foreign universities, managerial economics is confined to microeconomics.

it is predicted to be 4.9 per cent in 2013–14. According to a survey conducted by National Council of Applied Economic Research and Master Card, firms' business confidence had declined in the early 2012–13 due to expected decrease in domestic sales, production, exports and imports. It is anticipated to decline further in 2013–14.

The trend in the general price level—inflationary or deflationary—is another important factor shaping the economic environment. While inflation tends to create an optimistic trend, deplation creates a pessimistic trend in the economy. Besides, economic policies of the government are yet another important determinant of economic environment and business environment. Changing economic conditions and business environment have significant impact on the business decisions. As a result, business decisions on internal matters are likely to be affected by the deteriorating economic environment of the country. Macroeconomics is the study of the factors that determine the economic environment of the country and business prospects. Therefore, macroeconomics constitute a subject matter of managerial economics.

23.2 DEFINITION OF MACROECONOMICS[2]

Defining economics has been a difficult proposition. So is the case with defining the macroeconomics. In general, *macroeconomics is the study of economy as a whole.* However, economists have defined macroeconomics differently. For instance, look at the definition of macroeconomics offered by some famous economists.

Gardner Ackley Macroeconomics "concerns the over-all dimensions of economic life. ... More specifically, macroeconomics concerns itself with such variables as aggregate volume of an ecconomy, with the extent to which its resources are employed, with size of the national income, with the 'general price level.[3]

Kenneth E Boulding "Macroeconomics is the study of the nature, relationships and behaviour of aggregates of economic quantities.... Macroeconomics ... deals not with individual quantities as such, but with aggregates of these quantities ... not with individual incomes, but with the national income, not with individual prices, but with the price levels, not with individual output, but with the national output' ".[4]

J. M. Culbertson "Macroeconomic theory is the theory of income, employment, prices and money".[5]

P. A. Samuelson "Macroeconomics is the study of the behaviour of the economy as a whole. It examines the overall level of a nation's output, employment, prices, and foreign trade".[6]

Although these definitions are fairly comprehensive, they do not reveal the exact nature and scope of modern macroeconomics, nor do they fully capture its subject matter. Since "macroeconomics is [still] a young and imperfect science" (Mankiw, *Macroeconomics,* 2003, p.3), it is difficult to define it precisely. However, the definitions quoted above do

2. The terms 'microeconomics' and 'macroeconomics' were coined and used by a Norwegian economist, Ragnar Fricsh in 1933 in his paper "Propagation Problems and Impulse Problems in Dynamic Economics" in *Economic Essays in the Honour of Gustav Cassel* (London, 1933). The prefixes 'micros' and 'macros' meaning, 'small' and 'large' respectively, have been derived from the Greek language.

3. Gardner Ackley, *Macroeconomic Theory,* (Macmillan, NY, 1961) p. 4.

4. Kenneth E. Boulding, *A Reconstruction of Economics,* (John Wiley & Sons, Inc., NY, and Chapman & Hall, London, 1950). p. 171.

5. J. M. Culburtson, *Macroeconomic Theory and Stabilization Policy,* (McGraw-Hill, NY, 1968). p. 7.

6. Paul A. Samuelson and William D. Nordhaus, *Economics,* McGraw-Hill, 1995, 15th Edn., p. 381.

give an idea of the central theme of theoretical macroeconomics, and this is what matters in economics. The central theme that emerges from the above definitions may be stated as follows. *Macroeconomics is essentially the study of the behaviour and performance of the economy as a whole. More importantly, it studies the relationship and interaction between the 'factors and forces' that determine the level and growth of national output and employment, general price level, and the balance of payments positions of an economy.* This definition too should be treated only as a working definition of macroeconomics.

To comprehend better the **subject matter of macroeconomics**, let us look at the kinds of questions that macroeconomics seeks to answer.

• What determines the levels of economic activities, national income, total output, the general price level, and the overall employment in a country?

• How is the equilibrium level of national income determined?

• How product and monetary markets interact to bring about a general equilibrium in the economy?

• What causes fluctuations in the national output and employment?

• What determines the general level of prices in a country?

• What determines the level of foreign trade and trade balance?

• What causes disequilibrium in the balance of payments of a country?

• How do the monetary and fiscal policies of the government affect the economy?

• What economic policies can steer the economy on the path of growth?

These are some major theoretical questions that macroeconomics seeks to answer.

Besides, there are various areas and kinds of business decisions which have a long-run perspective and long-run financial implications, e.g., business decisions related to setting up a new production unit like Reliance planning to set up a new power plant in Uttar Pradesh; innovation and introduction of a new product like Tatas planning to launch a low-price car; adopting a new production technology, setting up a new industry or retail chains or shopping malls; going for franchise of a multinational company; and so on. All such business decisions have to be taken in view of the long-term perspective of business environment of the country. Business environment of a country is constituted of economic, social and political environment of the country. Therefore, business decision-makers have to take into account the economic, social and political environment of the country. For example, establishm ent of Special Economic Zones (SEZs) and the problem of land acquisition for SEZs has created both social and political problems—social problem because farmers lose the source of their livelihood and this often leads to their resistance and loss of human lives as it happened in Singrur and Nandigram of West Bengal; political problem because this issue has to be resolved at the political level. The social and political issues fall outside the purview of the macroeconomic aspects of managerial decisions. However, macroeconomic aspects of business environment are very much a part of managerial economics. The need for the study of economic aspects of business environment extends the scope of managerial economics to include also the study of macroeconomics.

In this part of the book, we present a brief discussion on some major macroeconomic concepts and theories that are generally used to assess the overall economic conditions and business environment for the purpose of business decision-making. In this introductory chapter, we present a brief discussion on the nature and scope of macroeconomics, major macroeconomic issues, and the scope of macroeconomics.

23.3 NATURE OF MACROECONOMICS

The nature of a subject is determined on the basis of whether the subject has a purely *positive* (theoretical) or *normative* (policy) orientation or both. A positive science has theoretical orientation in the sense that it aims at formulation of theories and laws. And a science having *normative* orientation aims at setting norms for finding solution to practical problems and provides policy guidelines. Macroeconomics has both positive (theoretical) and policy orientations.

Macroeconomics as a *positive science* has theoretical orientation. As a positive science, it uses macroeconomic models to explain the behaviour of macroeconomic variables—national income, aggregate consumption, savings and investment, price level, and employment. This results in formulation of macroeconomic theories. An important aspect of macroeconomic theories is that they provide a framework and analytical models to analyze the macroeconomic phenomena. The theories of income determination, consumption, investment, employment, price-level determination, product and money market equilibrium, exchange rate and balance-of-payments constitute the main body of the macroeconomic theories.

Macroeconomics as a *policy* or *normative* science provides guidelines for the formulation of economic policies. And also, it studies the working of macroeconomic policies of the government and investigates into the effects of public policies on the economy. It provides a framework for policy evaluation and guidelines for the formulation of appropriate macroeconomic policies.

In fact, the ultimate aim of macroeconomic studies is to formulate macroeconomic policies for macroeconomic management. It is perhaps for this reason that some economists consider macroeconomics as a purely *policy science*. According to some economists, “Macroeconomics is first and foremost a policy science”.7 In fact, it analyzes the various macro aspects of the economy with the aim of providing the base and logic for policy formulation with the objective of controlling and guiding the economy on the path of stable growth of output and employment. Thus, macroeconomics as a policy science provides a framework and instruments for managing the economy and guiding it on the path of growth and stability.

23.4 ORIGIN AND GROWTH OF MACROECONOMICS

Before we proceed to take up the study of macroeconomic theories, let us have a quick look at the origin and growth of the discipline as a separate branch of study. The evolution of macroeconomics can be divided into three stages.

(*i*) Pre-Keynesian era;

(*ii*) Keynesian revolution;

(*iii*) Post-Keynesian developments; and

(*iv*) Neo-Keynesianism.

The basic macroeconomic thoughts in these stages are briefly discussed here.

(*i*) Pre-Keynesian era. The pre-Keynesian era refers to the period of economic thoughts of *classical* and *neo-classical* economists. The classical and neo-classical economists had not formulated any unified macroeconomic theory though some macroeconomic thoughts can be traced to their writings. Their macroeconomic thoughts were in the forms of certain ‘postulates’ which can be summarized in the following words. If market forces of demand and supply are allowed to have free play, i.e., the *laissez-faire* system, then (*i*) there will always be full

7. Dernburg, Thomas, E., *Macroeconomics: Concepts, Theories and Policies* (McGraw-Hill Book Company, New York, 7th Edn. 1985), p. 4.

employment in the long-run, or there will be no involuntary unemployment; (*ii*) there will be neither over-production nor under-production, and (*iii*) the economy will always be in equilibrium in the long-run. Until the Great Depression of 1930s the classical postulates were not tested. However, the Great Depression proved all the classical postulates wrong. It exposed the inadequacy of the *theoretical* foundations of the ‘*laissez-faire* doctrine’, and its failure to explain the persistent violent business cycles. The classical economics had no solution to the problems of economic fluctuations. This marked the collapse of the classical postulates on macroeconomics.

(*ii*) The Keynesian Revolution. The collapse of the classical economics necessitated a fresh look at the working of the economic system and devising corrective measures and safeguards against the failures of the market economy. It was in this background that Keynes published a revolutionary book *The General Theory of Employment Interest and Money* 8 in 1936, which laid the foundation of macroeconomics.

The Keynesian macroeconomics deals mainly with determination of national income, employment, growth and economic stabilization. According to Keynes, the level of output and employment in an economy is determined by the aggregate demand, given the resources. The unemployment in any country is caused by lack of aggregate demand and economic fluctuations are caused by demand deficiency. The demand deficiency can be removed through compensatory government spending. Keynesian economics stresses the role of *demand management* by the government for the stable growth of the economy. “Perhaps the most fundamental achievement of the Keynesian revolution was the reorientation of the way economists view the influence of government activity on the private economy.”9

The period between the late 1930s and mid-1960s is called the period of “Keynesian Revolution” or the “Keynesian era”. During this period most economists were Keynesian and most governments, especially developed ones, had adopted Keynesian policies. The impact of Keynesian thoughts had also spread over the underdeveloped countries like India. However, the real economic world has never conformed to any particular economic thoughts or principles, ideas or ideology. It goes through a continuous process of evolution. It passes from one system to another, rendering prevailing thoughts, ideas and laws redundant and paving the way for new ones. The same happened with Keynesian revolution.

(*iii*) The Post-Keynesian Developments. Until the 1970s, the Keynesian thoughts and policies had global appeal and application. However, Keynesian economics started showing signs of failures during the 1970s. This raised the doubt about the relevance and applicability of Keynesian economics. Consequently, several other schools of macroeconomic thoughts emerged, viz., **Monetarism, Supply-side macroeconomics**,

New Classical macroeconomics and **New Keynesianism.**

Monetarism refers to economic thoughts of a school of economists called monetarists. The monetarists believe that 'only money matters' in the working and management of economy. ***Monetarists*,** a group of economists led by Milton Friedman claimed that Keynesian theory had failed to predict national output, rates of employment and unemployment, price levels, and interest and foreign exchange rates. The monetarists came out with a new revolutionary thought that the *role of money* is central to the growth and stability of national output, not the role of aggregate demand for real output as Keynesians believe. The

8. John Maynard Keynes, ' *The General Theory of Employment, Interest and Money* (Macmillan, London, 1936).

9. Blinder, Alan S. and Robert M. Solow, "Does Fiscal Policy Matter?" in *Macroeconomics Under Debate*, ed. by Alan S. Blinder, 1989 (Harvester Wheatsheaf, New York), p. 1.

monetarists shifted the emphasis in theoretics from the role of aggregate demand for real output to money demand and supply, and its policy orientation from demand management to monetary management. The *monetarists* emphasized that money supply, not aggregate demand, is the main determinant of output and inflation. This view is called *monetarism*.

The emergence of *Monetarism* led to a prolonged debate between the monetarists and the Keynesians, the central theme of debate being 'what determines the aggregate demand'. "While mainstream theories point to a number of different forces that influence aggregate demand—monetary and fiscal policies, investment spending, net exports and so forth—monetarists hold that changes in the money supply are far more important than all other forces in affecting nominal *GNP* in the short-run and prices in the long-run."10 The debate remains inconclusive.

New Classical Macroeconomics. During the 1980s, the Keynesian view was attacked by another group of economists, called **radicalists** led by Robert E. Lucus, the Nobel Laureate of 1995, their views known as *new classical macroeconomics*. In the opinion of Lucas, Keynesian orthodoxy has turned redundant not only from economic policy point of view but also from theoretical and methodological point of view. The new classicalist school emphasizes the role of individual's *rational expectations* about the future economic events, especially those on the supply side of the economy and about the future government policies. The core of the radicalist thought is that people's expectations about government monetary and fiscal policies determine the behaviour of aggregate supply and demand curves. The anticipated changes in monetary and fiscal policies cause a shift in aggregate demand curve, which causes an immediate and equal shift in the aggregate supply curve. Therefore, shifts in aggregate demand and supply curves do not cause any change in the real output. The new classical macroeconomics remains a matter of debate.

Supply-Side Economics. Another school of macroeconomics, that emerged during the 1980s is called "Supply-Side Economists". While Keynesians and monetarists have both built their arguments for 'what determines the aggregate demand' on the basis of the factors operating on the demand side of the market, "supply-side economists", led by Arthur Laffer, emphasized the role of the factors operating on the supply-side of the market. While Keynesian economists emphasize the role of shift in aggregate demand in changing employment and output, supply-siders stress the role of shift in the aggregate supply curve. Arthur Laffer, widely known for his famous "Laffer curve," argued that a cut in tax rate shifts aggregate supply curve rightward and leads to a rise in output and employment.

(*iv*) The New Keynesianism. While several path-breaking contributions were made to macroeconomic thoughts over the past 30 years, Keynesian economics prevailed as the focal point of reference for all the schools of economic thoughts in their reconstruction of macroeconomics. An attempt was made in the subsequent years to revive Keynesianism. As a result, there emerged yet another school called "New Keynesians." Contrary to the new classical group, the new Keynesians argue that market does not clear always, in spite of individuals working for their own interest. According to them, 'information problem and cost of changing prices lead to some price rigidities', which cause fluctuations in output and employment.

It is clear from the foregoing discussion that macroeconomics is still in the process of developing as a science. Given the purpose and scope of this book, however, we will confine our discussion to the *Keynesian macroeconomics*.

10. Vercelli, A., *Methodological Foundations of Macroeconomics*, (Cambridge, 1991), p. 4.

23.5 IMPORTANCE OF MACROECONOMICS

The importance of macroeconomics can be viewed from two angles: (*i*) how it contributes to government's economic policy formulation for the management of the economy, and (*ii*) how it guides business managers for taking appropriate business decisions and business planning. It has been proved over time that macroeconomics has contributed a great deal to both these areas. In this section, we discuss briefly how macroeconomics contributes to these two areas.

23.5.1 Application of Macroeconomic Management

In modern times, most countries are confronted with three major macroeconomic problems: (*i*) problems in achieving a sustainable growth rate, (*ii*) problem of unemployment, and (*iii*) problem of inflation. Given these problems, the basic economic functions of the government are (*i*) ensuring a sustainable economic growth rate; (*ii*) ensuring full employment, (*iii*) ensuring price stability. Macroeconomic problems of the country have to be resolved effectively as the economic fate of the country and its people depends on its economic strength. Macroeconomics contributes a great deal in formulating economic policies to achieve these ends.

As regards the ***problem of economic growth,*** in modern times, both developed and developing economies are striving hard to achieve and maintain a reasonably high growth rate. The developed countries had succeeded in achieving high growth rate after the World War II. But from the 1980s and till today, they have been facing the problem of decline in their economic growth rate. The worst of it, the 21st century started with the second worst global recession in 2008. The recent global economic recession beginning in the US has impacted the global economy. It is in this context that macroeconomics contributes a great deal. In fact, the importance of macroeconomics lies in fact that macroeconomics provides a theoretical framework to analyse the working system of the economy, to identify the factors responsible for decline in economic growth rate, and to find reasonable solution to the problem of decline in growth rate.

Similarly, ***unemployment problem*** has been a persistent problem of the developing and underdeveloped countries. For instance, in India the unemployment rate11 estimated on the basis of 'current daily status (CDS) was 8.2 per cent in 2004–05, though it declined to 6.6 per cent in 2009–10. Yet this is considered to a very high rate of unemployment. The recent global recession of 2008–2010 has created unemployment even in developed countries. For example, unemployment rate had risen from 2 per cent to about 6 per cent in current years due to economic recession.

Inflation, i.e., a persistent increase in the general price level, is another problem which most countries have faced over time and failed to control inflation. In India, the inflation rate varied between 6 and 10 per cent after 2008–09 and 2013–14, which is considered to be the biggest economic and political problem in the country. There is rather a paradoxical situation: the unemployment problem persists to continue despite economic growth and inflation whereas these factors might have mitigated the problem of unemployment. This adds complexity to basic macroeconomic problems of the country.

Furthermore, both developed and developing countries are constantly confronted with some or the other kind of macroeconomic problems, e.g., **recession and depression, stagflation, balance-of-payment deficits, outflow of capital, mounting debt burden or a country falling into debt trap,** and so on. These problems have to be solved if eventual economic

11. Economic Survey: 2012–13, p. 275.

collapse like the Great Depression of 1930s has to be averted. Even if economic catastrophes of this magnitude do not take place, macroeconomic problems like recession, unemployment, inflation and increasing external debt burden need to be solved because they have serious socio-political implications for the country in general and the government in particular.

The macroeconomic issues of a country need to be resolved effectively as they determine the economic fate of a country and its place in the world economy of today. As Samuelson has put it, "The political, social, and military fate of the nations depends greatly upon their economic success."12 The internal security, law and order situation, social harmony also depend to a great extent on the economic condition of the common man of a country. The macroeconomic issues have received increasing attention of the economists, politicians, governments, and international bodies (IMF and WB) alike. This is perhaps the most important reason why macroeconomics has gained so much importance in the recent years.

However, finding a lasting solution to three macroeconomic problems is becoming more and more difficult now because of growing complexity of the modern economic system. Modern economic system is becoming complex due to the following reasons.

(*i*) Growing insatiable human desire to consume more and better goods and services, building demand pressure on scarce resources;

(*ii*) Increasing economic interaction between the nations due to globalization of economic activities, causing international conflict of interest;

(*iii*) Increasing international flows of capital and technology, which help economic growth but cause outflow of domestic income; and

(*iv*) Growth of international economic unions which causes group-war on economic issues.

With growing macroeconomic complexities and macroeconomic challenges, macroeconomics has emerged as the most challenging and fascinating branch of economic science. As Samuelson puts it, "... no area of economics is today more vital and controversial than macroeconomics."[13] The importance of and interest in macroeconomics has increased tremendously over the past 25 years for both theoretical and practical reasons. Macroeconomics provides a sound analytical framework and guidelines for finding a reasonable solution to these problems.

23.5.2 Application of Macroeconomics to Business Decision-Making

Like macroeconomics contributes a great deal in the formulation of economic policies by the government for the management of the economy, macroeconomics helps to a great extent in business management. The importance of macroeconomics in relation to business management lies in fact that application of macroeconomic concepts and theories guide the business managers in understanding the economic environment of the country. Having a clear understanding of business environment of the country is a crucial factor in making business decisions especially in relations to those matters which have long-run implications and managerial decisions regarding *future business plan*, and those having long-run implications. All kinds of business decisions regarding future business plans are taken in view of the current and future *business environment* of the country. Business environment of a country is constituted of the economic, political and social conditions prevailing in the country. Given the political and social conditions, macroeconomic conditions of the country play a very important role

12. Samuelson, P.A., *Economics,* 1989 (McGraw-Hill, London), p. 76.

13. *Ibid*.

in business decisions, especially those pertaining to future business plans. Macroeconomics provides the basis for assessing the business environment of the country and judging its likely effect on planned business. Future business plans include conceiving new business ventures, giving them a concrete shape, planning future business actions, and their implementation. Whatever the area of business decisions, the current and the future business environment of the country are very important considerations in all kinds of business decisions. The factors that determine the business environment of the country are discussed below.

(i) The current and future trend in GDP/GNP High sustainable growth in *GDP/GNP* offers a promising prospect for all kinds of modern goods and services and, therefore, a good business environment. On the contrary low growth rate or decline in *GDP* reduces business prospects. For example, a 9 per cent growth in India's *GDP* in 2007–08 had provided a high business prospect and scope for all kinds of modern industries and a recession-like situation in the later half of 2008 in the economy had led to deterioration in the business enviornment in the country. But the fast revival of the economy in 2009 had brightened the business prospects. However, the downside in the economy in 2013–14 has once again downgraded the economic environment of the country.

(ii) The trend in the aggregate demand for consumer and capital goods Increasing demand for consumer and capital goods indicates expansion in the economy and a good business prospect. Stagnated or declining aggregate demand, even with increasing *GDP*, reduces business prospects.

(iii) Trend in the rate of savings and investment Rising rate of total savings indicates high availability of business finance and investible funds. Low rate of savings creates financial scarcity leading to rise in the interest rate. Rising interest rate is a big constraint on the business prospect in the country. Note that even a low rate of interest in a country facing recession, as in the US, fails to promote the business prospect and, therefore, investment.

(iv) The general price level and expected future trends The general level of price in a country may show three kinds of trends: (*i*) rising at a high rate resulting in *inflation*, (*ii*) declining rapidly showing *deflation*, and (*iii*) remaining stable, may be with minor fluctuations. Both inflation and deflation at high rates affect business prospects adversely—recall the effects of 12–13% inflation in 2007–08 and deflationary trend in 2008–09 when inflation rate turned out to be negative in June- July 2009. A *moderate rate of inflation* and also *price stability* provide a good business environment and business prospects.

(v) The level of employment and the likely trend Although trend in overall employment does not appear to be a direct concern of the business managers, it has an important and serious implication for business management. Growing employment indicates not only a better social environment in the country but also a better business prospect, because the rise in employment shows rise in wage incomes which leads to increase in demand for consumer goods, as a major part of wage incomes is spent on consumer goods and services. Rising unemployment reduces demand for consumer goods, on the one hand, and creates social problems and social crimes like kidnapping of businessmen for ransom, on the other, which affect business environment of the country adversely.

(vi) International aspects of the economy In today's world of globalisation, economic transactions across the border have serious implications and repercussions for domestic business.

International economic transactions include (*i*) international flows of goods and services—exports and imports, (*ii*) international flows of capital—inflows and outflows of capital, and (*iii*) international flows of labour—in and out migration of labour and income remittances. All these aspects of international economic transaction have their effects on the economy, both favourable and unfavourable. For example, large scale foreign investment has increased the level of investment in India, but it has lead to appreciation of the rupee, which is adversely affecting the exports of some kinds of goods and services, especially textiles and leather goods. So, the business managers, especially those involved in foreign trade and investment, have to take into account international trends and their impact on the domestic economy and on their business.

(vii) Government's macroeconomic policies Government's macroeconomics policies, especially *monetary* and *fiscal* policies, play a very important role in creating a favourable or an unfavourable environment in the country. A liberal monetary policy indicated by lower rate of interest and easy availability of funds promotes business environment and *vice versa*. The fiscal policy, i.e., government's taxation and expenditure policy, is devised to meet the need of the country to control its business cycle, high inflation and deflation, restructuring of industrial sector in favour of the society as whole, etc. For example, RBI adopted a stringent monetary policy to control double-digit inflation in India in October-November 2008, but it adopted a liberal, easy money policy in December 2008 when both growth rate and inflation rate started sliding down. The Finance Minister announced an expansionary fiscal policy including reduction in VAT and CST rates and additional public expenditure of `17,000 crore. The anticipated economic effects of monetary and fiscal policies that are adopted by the government have to be taken into consideration by the business managers while making their future plans.

Given the important elements of business environment, business managers have to make their future plans in conformity with the likely business conditions in the country. However, understanding and crystallising business environment is not an easy task. The knowledge of macroeconomic concepts, theories, and their relevance to business decisions contribute towards appropriate decision making.

SUMMARY

• Macroeconomics is the study of behaviour and performance of the economy as a whole. Macroeconomics studies the relationship and interaction between the 'factors and forces' that determine the level and growth of national income, employment, general price level, and balance of payments of the country.

• By nature, macroeconomics is both a positive and a normative science. In other words, it is both a theoretical and a policy science.

• Macroeconomics emerged as a new branch of economics in 1936. The foundation of macroeconomics was laid down by J. M. Keynes in his book *The general Theory of Employment, Interest and Money* published in 1936.

• Many other developments were made in macroeconomics by other economists in the later years. Post-Keynesian developments in macroeconomics are classified as Monetarism, Supply-Side Macroeconomics, New Classical Macroeconomics, and New Keynesianism.

• Macroeconomics has a great deal application to the formulation economic policies of the government for economic management and to business decision-making as it gives the picture of the economic environment of the country.

REVIEW QUESTIONS

1. What is macroeconomics? How is it different from microeconomics? What are the uses and limitations of macroeconomics?
2. Macroeconomics does not study which of the following?
(*a*) Performance of the entire economy
(*b*) Determination of the levels of economic activities
(*c*) Price and output determination of a commodity
(*d*) Factors and forces of economic fluctuations
(*e*) Interrelationships between macro variables.
3. Is macroeconomics a positive science or a normative science? Explain in this regard the importance of macroeconomics.
4. What is the scope of macroeconomics? How is it different from the scope of microeconomics?
5. What are the major macroeconomic issues confronting the economies of the world? How do you think macroeconomics can help in solving related problems?
6. Write a short note on the development in the field of macroeconomics in the post-Keynesian period.
7. How does macroeconomics contribute to business decision-making?
8. What factors led to emergence of macroeconomics? Did macroeconomics resolve economic problems of post Great Depression era?
9. Suppose a firm is planning to expand the production of its product. How do you think macroeconomics will be helpful in determining the scale of production.

FURTHER READING

ackley, G., *Macroeconomic Theory* (NY, Macmillan, 1961).
Baumal, William J., *Economic Dynamics* (NY, Macmillan, 1962).
Harrod, R. F., *Towards a dynamic Economics* (London, Macmillan, 1960).
Hicks, J. R., *Capital and Growth* (London, Clarendon, 1965).
macHlup, F., “Statics and Dynamics: Kaleidoscopic Words” in his *Essay in Economic Semantics* (Prentice-Hall, N.J., 1963).
macHlup, F., “Micro and Macroeconomics: Contested Boundaries and Claims of Superiority in his Essays on Economic Dynamics (W.W. Norton, NY, 1977).

CHAPTER **24** Circular Flow Model of the Economy

CHAPTER OBJECTIVES

The objective of this chapter is to describe how an economy works. Working of the economy has been shown and analysed by using the circular flows of production factors, products and money under the following three models.

• Simple economy model – two-sector model including households and firms
• Closed economy model – three-sector model consisting of households, firms and government
• Open economy model – Four-sector model consisting of households, firms, government and foreign sector

24.1 INTRODUCTION

We begin our study of macroeconomics by having a look at how an economy works. An economy is a system of interrelated economic activities and economic transactions. Basic economic activities include production, exchange and consumption. The economic activities are carried out in an integrated manner that creates a continuous process of economic transactions – buying and selling. Economic transactions generate two kinds of flows:

(*i*) Product or goods flow, and
(*ii*) Money flow.

The product flow consists of factor flows and product flows, i.e., flow of factors of production and of goods and services. In a monetized economy, factor and product flows generate money flows in the form of factor payments and payments for goods and services. The two kinds of flows go in opposite direction in a circular manner and make two kinds of *circular flows*. An economy keeps working so long as the two flows go on uninterrupted. The working of an economy can, therefore, be viewed as circular flows of product and money and the size of the economy as the volume of goods flow.

This chapter presents a brief description of how goods and factor flows are generated and how an economy works in a systematic manner. To begin with, we will first give a description of circular flows in a simple economy consisting of only two sectors: (*i*) **households** and (*ii*) **firms**. The **households** have two characteristics: (*a*) they are owners of all factors of production and (*b*) they are consumers of all final goods and services. **Firms**, on the other

hand, have two characteristics too: (*a*) they hire factors of production from the households and (*b*) they produce and sell their final products to the households. This model is then extended to include the **government sector** making it a 3-sector model. Finally, the model is extended further to include also the **foreign sector** (comprising only exports and imports goods and services) to make it a complete circular flow model consisting of households, firms, government and foreign sector.

24.2 CIRCULAR FLOWS IN A SIMPLE ECONOMY MODEL

We begin with the description of circular flows in a simple economy consisting of only two sectors, viz., households and firms, and there is no government and no foreign trade. In our simple economy model, *households* are assumed (*i*) to own all the factors of production, (*ii*) to consume all final goods and services and (*iii*) their income consists of wages, rent, interest and profits. The *business firms,* on the other hand, are assumed (*i*) to hire factors of production from the households;1 (*ii*) to produce and sell goods and services to the households; and (*iii*) they do not save, i.e., there is no corporate saving.

Fig. 24.1 *The Circular Flows of Goods and Money in a Simple Economy*

The working of and circular flows of incomes and expenditure in the two-sector model are illustrated in Fig. 24.1. As the figure shows, factors of production flow from the households to the factor market and from the factor market to the firms. As shown in the lower half of the figure, goods and services produced by the firms flow from the firms to the households. The arrows showing factor and product flows make the *product flows* or *real flows.* Note that real flows take a circular path.

1. The households also produce and consume certain goods and services. In their capacity as producers, they belong, functionally, to the category of firms.

Product flow generates *money flow.* As the figure shows, money flows from the firms to the households in the form of factor payments as wages, interest, rent and dividends. Factor payments take the form of household incomes. Households spend their incomes on goods and services produced by the firms. As a result, money incomes flow from the households to the firms in the form of payments for goods and services. Thus, money paid by the firms as factor payments flows back to the firms as payments made by the households for goods and services. This makes the circular flow of money.

Note that product and money flows make the circular flows in the economy and that products and money flow in opposite directions. These flows represent the working of the simple economy. An important feature of product and money flows is that the *value* of real flow equals the money flow. This equality results from the fact that *factor payments* are equal to *household incomes* and since households spend their total income on consumer goods and services, *household expenditure* equals the total receipts of the firms, which equals *the value of the output.* These equalities can be summarized in the form of *identities* as follows. In the final analysis, household incomes ≡ factor payments ≡ the money value of output. This identity holds so long as households spend their total income, i.e., households do not hoard any part of their income, and firms spend their total receipts on hiring factors of production from the households.

The Effect of Withdrawals and Injections

The product and money flows shown in Fig. 24.1 assume that there are no *withdrawals* from and *injections* into the economy. *Withdrawals* means withholding money incomes as idle cash balance. This withholding is not 'saving' because savings are returned to the circular flows in the form of purchase of capital goods (investment). Withdrawals are also called ***leakages***. *Injection,* on the other hand, means money expenditure in addition to factor incomes. In reality, however, there are *withdrawals from* and *additions to* the circular flows. Let us look at the forms and nature of withdrawals and injections. In our two-sector model, a *withdrawal is an amount set aside by the households and/or by the firms, not to be spent on the goods and services over a period of time*. If households set aside a part of their current income as a provision for old age or as security against the loss of job, etc., it is called a *withdrawal*. It is important to note that *a withdrawal is not a saving*. For, savings are ultimately returned to the circular flows in the form of investment expenditure. Likewise, firms may withhold a part of their sales revenue and not return it to the circular flows if they anticipate depression. *Withdrawals reduce the volume of the circular flow.*

Injections, on the other hand, are *the amount that is spent by the households and/or firms in addition to their current incomes generated within the regular economy.* Injections may be made by the households in the form of spending past savings or hoardings. Injection by the firms may take the form of spending their accumulated savings. Firms may inject money into the economy by borrowing from households. *Injections increase the size of the flow.*

24.3 CIRCULAR FLOWS OF PRODUCT AND MONEY IN A THREE-SECTOR MODEL

This section presents circular flows in a three-sector model. A three-sector model is created by adding the government sector to the two-sector model. The inclusion of the government into the model brings in the government's tax revenue and expenditure and the effect of

its fiscal operations (taxation and spending) on the circular flows. For simplicity sake, however, we will consider only three kinds of monetary flows between the government and the rest of the economy, viz.,

(*i*) direct taxes on both households and firms,

(*ii*) government expenditure,

(*iii*) transfer payments and subsidies.

These fiscal operations of the government have different kinds of effects on the circular flows of goods and money flows.

The real and money flows in a three-sector model are shown in Fig. 24.2. Note that in this modified figure, 'factor market' is placed in the centre to make pl ace for the 'government sector' at the top. In Fig. 24.2, real and money flows between the households and firms (or business sector) are the same as shown in Fig. 24.1. Let us now look at the product and money flows between the government on one hand and households and firms on the other.

Services

Fig. 24.2 *Real and Money Flows in a Three-Sector Model*

As Fig. 24.2 shows, a part of the household incomes flows to the government in payment of taxes. The government spends a part of its tax revenue as 'factor payments' to the households, i.e., on purchase of factor services (labour and private property) and a part in the form of transfer payments, as pension and food subsidy, etc. These flows make *money flow* between the households and the government. As regards the *real flows,* factors of production move from the households to the government and social services (schools, hospitals, police, roads, etc.) flow from the government sector to the households. These flows make the *real flows.* Thus, a part of household resources (real) and money incomes keep circulating between the house-holds and the government. Note that the two flows need not be equal.

Similar flows take place between the government and the firms. Firms pay a part of their incomes as taxes to the government. In return, the government pays back a part

of its tax revenue in the form of payments for purchases from the firms and a part as subsidies. This makes *money flow* between households and firms. The flow of goods and services from firms to government makes the *real flow.*

It may be noted at the end that *taxes are withdrawals* from the circular flows and *government expenditure is an injection* into the income stream. The *transfer payments* by government (e.g., old age pension, subsidies, unemployment allowance, etc.) are *injections* to the circular flows.

24.4 CIRCULAR FLOWS IN A FOUR-SECTOR MODEL: A MODEL WITH FOREIGN SECTOR

In this section, we describe circular flows of goods and money in a four-sector model. Four-sector model is formed by adding foreign sector to the three-sector model. Foreign sector consists of two kinds of international economic transactions:

(*i*) Foreign trade, i.e., export and import of goods and services, and

(*ii*) Inflow and outflow of capital.

Fig. 24.3 *Circular Flows in Four-Sector Model*

International transactions and the consequent flows of goods and money make a complex system. For simplicity sake, however, we assume that foreign sector consists only of exports (*X*) and imports (*M*) of goods and services and that households export only labour but import goods and non-labour services.

The circular flows of goods and money in a four-sector model are illustrated in Fig. 24.3. In this figure, the internal flows of goods and money (*i*) between households and firms, (*ii*) between households and government and (*iii*) between firms and government are the same as shown in Fig. 24.2. We will, therefore, concentrate on the flows of goods and money (*i*) between households and the rest of the world, (*ii*) between domestic firms and the rest of the world and (*iii*) between the government and the rest of the world.

In Fig. 24.3, 'foreign sector' is shown at the bottom of the figure. As mentioned above, foreign sector consists of exports from and imports to the domestic economy by (*a*) households, (*b*) firms and (*c*) government. As shown in the figure, households export only manpower (labour). In return, they receive foreign remittances. But they import goods and services for which they make payments. The inflows and outflows pertaining to households need not be equal. Firms, on the other hand, are shown to import and export both goods and services. So is the case with the government. The government sector exports and imports both goods and services.

Let us look at the consequences of exports and imports on the volume of circular flows. Exports (*X*) from any sector make goods and services flow outside the domestic economy and make money (foreign exchange) flow into the domestic economy in the form of 'receipts from export'. Exports make foreign incomes flow into the domestic economy. Similarly, imports (*M*) cause inflow of goods and services and outflow of money converted in foreign exchange. This means outflow of domestic income to foreign countries. Another flow is generated by the 'export of manpower' by the households. The export of manpower brings in 'foreign remittances' in terms of foreign exchange. Foreign exchange converted in domestic currency makes another inflow of income. These inflows and outflows go on continuously so long as there is foreign trade and export of manpower.

So far as the effect of foreign trade on the magnitude of the overall circular flows is concerned, it depends on the *trade balance,* defined as $X - M$. If $X > M$, it means inflow of foreign income is greater than its outflow or there is net gain from foreign trade. This increases the magnitude of circular flows of income and expenditure. By the same logic, if $X < M$, it decreases the magnitude of circular flows. And, if $X = M$, inflow and outflows of incomes are equal. This leaves the volume of circular flows unaffected.

SUMMARY

- A monetised economy works through the circular flow of factors of production, products and money. Money flows in reverse direction of production factors and goods and services.
- Three kinds of models are used to explain the circular flows of products and money: (*i*) two-sector model, (*ii*) three-sector model, and (*iii*) four-sector model.
- The two-sector model, known also as simple economy model, consists of only two sectors – households and firms. The two-sector model shows the flow of factors of production (labour and capital) from households to firms and flow of money in the form of factor payments from the firms to household (the factor owners). Products

created by the firms flow from the firms to the household and money from the households to firms.

- Three-sector model is constructed by adding government sector to the two-sector model. Addition of government sector brings about two kinds of other flows: (*i*) flow of money in the form of tax from households and firms to the government, and (*ii*) flow of goods and services from the government to the households and firms.
- Four-sector model is constructed by adding foreign sector to the three-sector model. Addition of foreign sector created international flow of goods and services in the form of exports and imports and inflow and outflow of money. Export of goods and

services generates inflow of money (foreign exchange) and imports generate outflow of money (foreign exchange).

REVIEW QUESTIONS

1. What are the two main flows in an economy? How do they arise? What do they signify?
2. Describe an economy in terms of circular flows of income and expenditure. What determines the magnitude of the circular flows?
3. What is meant by withdrawals and injections? How do they affect the size of the circular flows of income and expenditure in an economy?
4. How does the addition of the government sector to the two-sector model change the structure of the model and of the circular flows?
5. What is the effect of change in personal taxes and the government expenditure on the circular flows of income and expenditure? Does a balanced budget policy result in expansion or reduction in the circular flows?
6. Illustrate graphically the circular flows of income and expenditure in a four-sector model. Explain also the effect of adverse and favourable balance of trade on the size of the circular flows.

FURTHER READING

Baumol, W.J., *Economic Dynamics: An Introduction* (The Macmillan Company, New York, 2nd Edn., 1959).

BenJamin, Higgins, *Economic Development* (Central Book Depot, Allahabad, 1961).

Domar, e., "Capital Expansion, Rate of Growth and Employment." *Econometrica*, Vol. 14, 1946.

HarroD, r.F., "An Essay in Dynamic Theory", *Economic Journal*, Vol. 49, 1939.

HarroD, r.F., *Towards a Dynamic Economics* (Macmillan, 1948).

KinDleBerger, C.P. and B. HerriCK, *Economic Development* (McGraw Hill Kogakusha, International Edn., 1977).

meaDe, J.e., *Neo-Classical Theory of Economic Growth* (Oxford University Press, 1961).

sen, a.K., (ed.), *Growth Economics* (Penguin Books, England, 1970) 'Introduction' by the Editor.

soloW, r.m., "A Contribution to the Theory of Economic Growth," *Quarterly Journal of Economics,* 7 Feb., 1955.

sWan, T.W., "Economic Growth and Capital Accumulation," *Economic Record*, November 1956.

ToBin James, "A Dynamic Aggregate Model," *Journal of Political Economy*, 63, April, 1955.

CHAPTER

25 National Income: Concept and Measurement

CHAPTER OBJECTIVES

The objective of this chapter is to discuss the concepts and measurement of the national income. The content of this chapter consists of the following aspects.

• Definition of national income
• Some basic concepts related to national income
• Methods of measuring national income
• Problems in measuring national income
• Measurement of national income in India

25.1 INTRODUCTION

In the preceding chapter, we have discussed the working of the national economy through the circular flow of products and money. The volume of the circular flow determines the national income of country. National income is the final outcome of the economic activities of the people of the nation. National income is the most important macroeconomic variable. The level of national income of a country determines its economic strength to sustain the economic, political, social and international problems. It determines the level of employment, per capita income, aggregate demand and supply of goods and services, and the level of production.

What is more important form the view point of business management is the fact that change in national income determines the business environment of the country. Business environment matters a great deal in making business decisions. We begin our discussion by defining the term 'national income' and some basic concepts related to national income measurement.

25.2 DEFINITION OF NATIONAL INCOME

National income can be defined as *the sum of money value of all final goods and services produced in a country over a period of one year*. Some terms used in the definition of national income need to be clarified. The term 'money value' means the value estimated at the current price of the goods and services. As regards the term 'final goods and services', productive activities create many goods and services. While some goods and services are final, some are used again in the process of production, and some take the form of capital. In all cases, while measuring national income only *final goods and services* are taken into account.

National income can be defined also as the *sum of all factors incomes*. Factors of production, viz., land, labour, capital and entrepreneurship, are used to produce goods and services. Use of factors of production generates factor incomes in the form of rent, wages, interest and profit. These factor incomes constitute the national income of a country. The sum of all these factor-incomes gives the estimate of national income.

25.3 NATIONAL INCOME: SOME BASIC CONCEPTS

Measuring national income is an extremely complex and complicated task as it involves both conceptual and practical problems. Therefore, before we discuss the methods of estimating national income, it is useful to understand the concepts used in estimating national income. In this section, we take the view of some concepts used in estimation of national income.

1. Gross Domestic Product (***GDP*****).** Gross domestic product (*GDP*) is the measure of the total market value of all final goods and services produced in the domestic economy during a period of one year *plus* income earned by the foreigners in the country *minus* income earned by countrymen from abroad.

2. Gross National Product (***GNP*****).** *GNP* is the measure of the total market value of all final goods and services produced in the domestic economy during a period of one year *plus* incomes earned abroad by the citizens *minus* income earned by the foreigners in the country.

3. Economic and Non-economic Products. In estimating national income only economic products are included. Goods and services that are produced to be sold at market price and goods and services that are produced by the government and public organizations[1] are treated as ***economic products***. Thus, economic products include both marketable and non-marketable goods. And, ***non-economic products*** include services rendered to self, to family, to relations and to neighbours. Non-economic products are not included in national income.

4. Intermediate and Final Products. Products (goods and services) that are used in the process of further production are considered to be *intermediate product* and products that are consumed by the final consumers are considered to be *final products*. The same product may be an intermediate or a final product. For examples, when wheat produced by farmers are consumed by themselves, it is treated as *final good*, but when it is sold to bread companies, it is treated as intermediate product. Likewise, when services provided by the government, e.g., transport, telephonic, postal, railway services, etc. are used by the consumers, they are treated as *final goods* and when used in the process of production, then these services are treated as *intermediate goods*. In estimating national income, only final products are taken into account.

5. Transfer Payments. Payments made by the people to other people, organizations or to the government without any equal transfer in return are treated as *transfer payment*. For example, gifts paid to relatives and friends, donations given to social organization, and taxes paid to government authorities, etc., are transfer payments. Such payments are not taken into account in national income estimates.

1. Goods and services produced by the government and public organizations are paid for by the people in the form of tax or fees.

National Income as Money Flow

We have defined national income from the angle of *product flows*. The same can be defined in terms of *money flows*. While economic activities generate flow of goods and services, on the one hand, they generate money flows, on the other, in the form of factor payments—wages, interest, rent, profits, and earnings of self-employed. Thus, national income may also be estimated by adding the factor earnings and adjusting the sum for indirect taxes and subsidies. The national income thus obtained is known as *national income at factor cost*. It is related to money income flows.

The concept of national income refers to the income of the society as a whole. It differs fundamentally from the concept of *private income*. Conceptually, national income refers to the money value of the entire volume of final goods and services resulting from all economic activities of the country. This is not true of private income. Also from the calculation point of view, there are certain receipts of money or of services and goods that are not ordinarily included in private incomes but are included in the national incomes, and *vice versa*. National income includes, for example, employer's contribution to the social security and welfare funds for the benefit of employees, profits of public enterprises and services of owner occupied houses. But it excludes the interest on warloans, social security benefits and pensions. These items are, however, included in the private incomes. The national income is, therefore, not merely an aggregation of the private incomes.

25.4 MEASURES OF NATIONAL INCOME

25.4.1 Gross National Product (GNP)

Of the various measures of national income used in national income analysis, *GNP* is the most important and widely used measure of national income. It is the most comprehensive measure of the national productive activities in an open economy. *The GNP is defined as the value of all final goods and services produced during a specific period, usually one year, plus incomes earned abroad by the nationals minus incomes earned locally by the foreigners*. The *GNP* so defined is identical to the concept of gross national income (*GNI*). Thus, *GNP* = *GNI*. The difference between the two is only of procedural nature. While *GNP* is estimated on the basis of product-flows, *GNI* is estimated on the basis of money income flows, (i.e., wages, profits, rent, interest, etc.).

25.4.2 Gross Domestic Product (GDP)

The Gross Domestic Product (*GDP*) is defined as the market value of all final goods and services produced in the domestic economy during a period of one year, *plus* income earned locally by the foreigners *minus* incomes earned abroad by the nationals. The concept of *GDP* is similar to that of *GNP* with a significant procedural difference. In case of *GNP*, incomes earned by the nationals in foreign countries are added and incomes earned locally by the foreigners are deducted from the market value of domestically produced goods and services. But, in case of *GDP*, the process is reversed—incomes earned locally by foreigners are added and incomes earned abroad by the nationals are deducted from the total value of domestically produced goods and services.

25.4.3 Net National Product (NNP)

NNP is defined as *GNP less* depreciation, i.e.,

NNP = *GNP* – Depreciation

Depreciation is that part of total productive assets which is used to replace the capital worn out in the process of creating *GNP*. Briefly speaking, in the process of producing goods and services (including capital goods), a part of total stock of capital is used up. 'Depreciation' is the term used to denote the worn out or used up capital in the process of production. An estimated value of depreciation is deducted from the *GNP* to arrive at *NNP*. The *NNP*, as defined above, gives the measure of net output available for consumption and investment by the society (including consumers, producers and the government). *NNP* is the real measure of the national income. *NNP* = *NNI* (net national income). In other words, *NNP* is the same as the national income at *factor cost*. It should be noted that *NNP* is measured at market prices including direct taxes. Indirect taxes are, however, not a part of actual cost of production. Therefore, to obtain real national income, indirect taxes are deducted from the *NNP*. Thus, *NNP less* indirect taxes = National Income.

25.4.4 Some Accounting Definitions

(*a*) **Accounting Identities at Market Price**

GNP ≡ *GNI* (Gross National Income)

GDP ≡ *GNP less* net income from abroad

NNP ≡ *GNP less* depreciation

NDP (Net Domestic Product) ≡ *NNP less* net income from abroad

(*b*) **Some Accounting Indentities at Factor Cost**

GNP at factor cost ≡ *GNP* at market price *less* net indirect taxes

NNP at factor cost ≡ *NNP* at market price *less* net indirect taxes

NDP at factor cost ≡ *NNP* at market price *less* net income from abroad

NDP at factor cost ≡ *NDP* at market price *less* net indirect taxes

NDP at factor cost ≡ *GDP* at market price *less* depreciation

25.5 METHODS OF MEASURING NATIONAL INCOME2

National income of a country is created by its people participating in different kinds of economic activities and producing goods and services. For measuring national income, an economy is viewed from three different angles.

1. The national economy is considered as an *aggregate of productive units* of different sectors such as agriculture, mining, manufacturing, trade and commerce, services, etc.

2. The whole national economy is viewed as a *combination of individuals and households owning different kinds of factors of production* which they use themselves or sell factor-services to make their livelihood.

2. This section is based mostly on Paul Studenski's, *The Income of Nations (Part Two), Theory and Methodology* (University Press, New York, 1958).

3. The national economy may also be viewed as a collection of consuming, saving and investing units (individuals, households, firms and government).

Following these notions of a national economy, national income may be measured by three different corresponding methods:

1. *Net product method*—when the entire national economy is considered as an aggregate of producing units;

2. *Factor-income method*—when national economy is considered as combination of factor-owners and users;

3. *Expenditure method*—when national economy is viewed as a collection of spending units.

The procedures which are followed in measuring the national income in a *closed economy*—an economy which has no economic transactions with the rest of the world—are briefly described here. The measurement of national income in an open economy and adjustment with regard to income from abroad will be discussed subsequently.

25.5.1 Net Output or Value Added Method

The net output method is also called the *value added method*. In its standard form, this method consists of three stages: "(*i*) estimating the gross value of domestic output in the various branches of production; (*ii*) determining the cost of material and services used and also the depreciation of physical assets; and (*iii*) deducting these costs and depreciation from gross value to obtain the net value of domestic output..."[3]. The net value of domestic product thus obtained is often called the *value added* or *income product* which is equal to the sum of wages, salaries, supplementary labour incomes, interest, profits, and net rent paid or accrued. Let us now describe the stages (*i*) and (*ii*) in some detail.

(a) Measuring Gross Value. For measuring the gross value of domestic product, output is classified under various categories on the basis of the nature of activities from which they originate. The output classification varies from country to country depending on (*i*) the nature of domestic activities; (*ii*) their significance in aggregate economic activities, and (*iii*) availability of requisite data. For example, in the US, about seventy-one divisions and subdivisions are used to classify the national output; in Canada and the Netherlands, classification ranges from a dozen to a score; and in Russia, only half a dozen divisions are used. According to the CSO publication, fifteen sub-categories are currently used in India.

After the output is classified under the various categories, the value of gross output is computed in two alternative ways: (*i*) by multiplying the output of each category of sector by its respective market price and adding them together or (*ii*) by collective data about the gross sales and changes in inventories from the account of the manufacturing enterprises and computing the value of *GDP* on the basis thereof. If there are gaps in data, some estimates are made thereof and gaps are filled.

(b) Estimating Cost of Production. The next step in calculating the net national product is to estimate the cost of production including depreciation. Estimating cost of production is, however, a relatively more complicated and difficult task because of non-availability of adequate and requisite data. Much more difficult is the task of estimating depreciation

3. Studenski, Paul, *op. cit.*

since it involves both conceptual and statistical problems. For this reason, many countries adopt factor-income method for estimating their national income.

However, countries adopting net-product method find some ways and means to calculate the deductible cost. The costs are estimated either in absolute terms (where input data are adequately available) or as an overall ratio of input to the total output.

The general practice in estimating depreciation is to follow the usual business practice of depreciation accounting. Traditionally, depreciation is estimated at some percentage of capital, permissible under the tax-laws. In some estimates of national income, the estimators deviate from the traditional practice and estimate depreciation as some ratio of the current output of final goods.

Following a suitable method, deductible costs including depreciation are estimated for each sector. The cost estimates are then deducted from the sectoral gross output to obtain the net sectoral products. The net sectoral products are then added together. The total thus obtained is taken to be the measure of net national product or national income by net product method.

25.5.2 Factor-Income Method

This method is also known as *income method* and *factor-share method*. Under this method, the national income is calculated by adding up all the "incomes accruing to the basic factors of production used in producing the national product". Factors of production are conventionally classified as land, labour, capital and organization. Accordingly, the national income equals the sum of the corresponding factor earnings. Thus,

National income = Rent + Wages + Interest + Profit

However, in a modern economy, it is conceptually very difficult to make a distinction between earnings from land and capital, on the one hand, and between the earnings from ordinary labour and entrepreneurial functions, on the other. For the purpose of estimating national income, therefore, factors of production are broadly grouped as labour and capital. Accordingly, national income is supposed to originate from two primary factors, viz., labour and capital. In some activities, however, labour and capital are jointly supplied and it is difficult to separate the labour and capital contents from the total earnings of the supplier. Such incomes are termed as *mixed incomes*. Thus, the total factor-incomes are grouped under three categories; (*i*) labour incomes, (*ii*) capital incomes and (*iii*) mixed incomes.

(a) Labour Incomes. Labour incomes included in the national income have three components: (*a*) wages and salaries paid to the residents of the country including bonus and commission, and social security payments; (*b*) supplementary labour incomes including employer's contribution to social security and employee's welfare funds, and direct pension payments to retired employees;4 (*c*) supplementary labour incomes in kind, e.g., free health and education, food and clothing, and accommodation, etc. Compensations in kind (food and clothes) to domestic servants and such other free-of-cost services provided to the employees are included in labour income. War bonuses, pensions, service grants are not included in labour income as they are regarded as 'transfer payments'. Certain

4. Conventionally, pension to the retired employees is considered to be a 'transfer payment' and is excluded from the labour income and national income accounting. In the US, however, this item is included in national income (See Studenski, *op. cit.*, pp. 11 and 118–20).

other categories of income, e.g., incomes from incidental jobs, gratuities, tips, etc., are ignored for lack of data.

(b) Capital Incomes. According to Studenski, capital incomes include the following kinds of earnings:

(*a*) dividends excluding inter-corporate dividends;

(*b*) undistributed before-tax-profits of corporations;

(*c*) interest on bonds, mortgages, and saving deposits (excluding interests on war bonds, and on consumer-credit);

(*d*) interest earned by insurance companies and credited to the insurance policy reserves;

(*e*) net interest paid out by commercial banks;

(*f*) net rents from land, buildings, etc., including imputed net rents on owner-occupied dwellings;

(*g*) royalties and

(*h*) profits of government enterprises.

The data for the first two items are obtained mostly from the firms' books of accounts submitted for taxation purposes. But the definition of profit for national accounting purposes differs from that employed by taxation authorities. Some adjustments in income tax data therefore, become, necessary. The data adjustments generally pertain to (*i*) excessive allowance of depreciation made by the firms; (*ii*) elimination of capital gains and losses since these items do not reflect the changes in current income and (*iii*) elimination of under or over-valuation of inventories on book-value.

(c) Mixed Income. *Mixed incomes* include earnings from (*a*) farming enterprises, (*b*) sole proprietorship (not included under profit or capital income) and (*c*) other professions, e.g., legal and medical practices, consultancy services, trading and transporting, etc. This category also includes the incomes of those who earn their living through various sources as wages, rent on own property, interest on own capital, etc.

All the three kinds of incomes, viz., labour incomes, capital incomes and mixed incomes, added together give the measure of national income by *factor-income method*.

25.5.3 Expenditure Method

The expenditure method, also known as **final product method,** measures national income at the final expenditure stages. In estimating the total national expenditure, any one of the two following methods are used: (*i*) all the money expenditures at market price are computed and added up together and (*ii*) the value of all the products finally disposed off are computed and added up, to arrive at the total national expenditure. The items of expenditure which are taken into account under the *first method* are (*a*) private consumption expenditure; (*b*) direct tax payments; (*c*) payments to the non-profit-making institutions and charitable organizations like schools, hospitals, orphanages, etc. and (*d*) private savings. Under the *second method*, the following items are considered: (*a*) private consumer goods and services, (*b*) private investment goods, (*c*) public goods and services and (*d*) net investment abroad. The second method is more extensively used because the data required in this method can be collected with greater ease and accuracy.

Treatment of Net Income from Abroad. We have so far discussed methods of measuring national income of a 'closed economy'. But most economies are *open* in the sense that they carry out foreign trade in goods and services and financial transactions with the rest of the world. In the process, some nations make net income through foreign trade while some lose their income to foreigners. The net earnings or losses from foreign trade change the national income. In measuring the national income, therefore, the net result of external transactions is adjusted to the total. Net incomes from abroad are added to, and net losses from the foreign transactions are deducted from the total national income arrived at through any of the above three methods.

Briefly speaking, all exports of merchandise and of services like shipping, insurance, banking, tourism, and gifts are added to the national income. And, all the imports of the corresponding items are deducted from the value of national output to arrive at the approximate measure of national income. To this is added the net income from foreign investment. These adjustments for international transactions are based on the international balance of payments of the nations.

25.6 CHOICE OF METHODS

As discussed above, there are three standard methods of measuring national income, viz., net product (or value added) method, factor-income or factor cost method and expenditure method. All the three methods give the same measure of national income, provided requisite data for each method is adequately available. Therefore, any of the three methods may be adopted to measure the national income. But all the three methods are not suitable for all the economies simply for non-availability of necessary data and for all purposes of estimating national income. Hence, the question of choice of method arises.

Two main considerations on the basis of which a particular method is chosen are: (*i*) the purpose of national income analysis and (*ii*) availability of necessary data. If the objective is to analyze the *net output* or *value added*, the net output method is more suitable. In case the objective is to analyze the factor-income distribution, the suitable method for measuring national income is the *income method*. If the objective at hand is to find out the expenditure pattern of the national income, the expenditure or final products method should be applied. However, availability of adequate and appropriate data is a relatively more important consideration is selecting a method of estimating national income.

Nevertheless, **the most common method is the net product method** because (*i*) this method requires classification of the economic activities and output thereof which is much easier than classifying income or expenditure; and (*ii*) the most common practice is to collect and organize the national income data by the division of economic activities.

Thus, the easy availability of data on economic activities is the main reason for the popularity of the output method.

It should be, however, borne in mind that no single method can give an accurate measure of national income since the statistical system of no country provides the total data requirements for a particular method. The usual practice is, therefore, to combine two or more methods to measure the national income. The combination of methods again depends on the nature of data required and sectoral break-up of the available data.

25.7 MEASUREMENT OF NATIONAL INCOME IN INDIA

In India, a systematic measurement of national income was first attempted in 1949. Earlier, many attempts were made by some individuals and institutions. The first attempt to estimate India's national income was made by Dadabhai Naoroji in 1867–68. The first systematic attempt was made by Prof. V.K.R.V. Rao to estimate India's national income for the year 1931–32. Many other attempts were subsequently made, mostly by the economists and the government authorities, to estimate India's national income.5 These estimates differ in coverage, concepts and methodology and are not comparable. Besides, earlier estimates were mostly for one year, only some estimates covered a period of 3 to 4 years. It was therefore not possible to construct a consistent series of national income and assess the performance of the economy over a period of time.

In 1949, a National Income Committee (*NIC*) was appointed with P.C. Mahalanobis as its Chairman, and D.R. Gadgil and V.K.R.V. Rao as members. The *NIC* not only highlighted the limitations of the statistical system of that time but also suggested ways and means to improve data collection systems. On the recommendation of the Committee, the Directorate of National Sample Survey was set up to collect additional data required for estimating national income. The *NIC* estimated the country's national income for the period from 1948–49 to 1950–52. In its estimates, the *NIC* also provided the methodology for estimating national income, which was followed till 1967.

In 1967, the task of estimating national income was assigned to the Central Statistical Organization (*CSO*). Till 1967, the *CSO* had followed the methodology laid down by the *NIC*. Thereafter, the *CSO* adopted a relatively improved methodology and procedure which had become possible due to increased availability of data. The improvements pertain mainly to the industrial classification of the activities. The *CSO* publishes its estimates in its publication, *Estimates of National Income*.

Methodology. Currently, output and income methods are used by the *CSO* to estimate the national income of the country. The output method is used for agriculture and manufacturing sectors, i.e., the commodity producing sectors. For these sectors, the value added method is adopted. Income method is used for the service sectors including trade, commerce, transport and government services. In its conventional series of national income statistics from 1950–51 to 1966–67, the *CSO* had categorized the income in 13 sectors. But, in the revised series, it had adopted the following 15 break-ups of the national economy for estimating the national income; (*i*) Agriculture; (*ii*) Forestry and Logging; (*iii*) Fishing; (*iv*) Mining and Quarrying; (*v*) Large-scale Manufacturing; (*vi*) Small-scale Manufacturing; (*vii*) Construction; (*viii*) Electricity, Gas and Water Supply; (*ix*) Transport and Communication; (*x*) Real estate and Dwellings, (*xi*) Public Administration and Defence; (*xii*) Other Services and (*xiii*) External Transactions. The national income is estimated at both constant and current prices.

5. Some often quoted estimates were made by Atkinson, F.J. (1875 and 1895); Major Baring (1881); W. Digby, (1898–99); Curzon (1901); E.A. Home, (1911); C.N. Vakil and S.K. Muranjan (1891–94 and 1911–14); Findlay Shirras (1911 and 1921); K.T. Shah and K.J. Khambata (1900–14 annual and 1921–22); V.K.R.V. Rao (1925–29 and 1931–32); *Commerce, Journal* (1938–39, 1942–43 and 1947–48). (Year in the parentheses are the reference years).

Estimates of India's GNP and NNP

The estimates of India's *GNP* and *NNP*, both at factor cost at current prices and at constant prices of 1999–2000 are given in Table 25.1. Growth rates of *GNP* and *NNP* for the period from 1999–00 to 2005–06 are given in Table 25.2.

Table 25.1 *Estimates of India's GNP, NNP and Per capita Income at Factor Cost*

(At current prices and at 2004–2005 prices)

Year	Gross National Product (` in Crore) At Current Prices	Gross National Product (` in Crore) At Constant Prices	Net National Product (` in Crore) At Current Prices	Net National Product (` in Crore) At Constant Prices	Per Capita NNP (`) At Current Prices	Per Capita NNP (`) At Constant Prices
2000–01	1969249	2318974	1762358	2074858	17295	20362
2001–02	2147677	2450681	1918827	2190737	18450	21065
2002–03	2321510	2551730	2075329	2278363	19653	21575
2003–04	2601508	2757120	2329354	2466093	21729	23005
2004–05	2949089	2949089	2629198	2629198	24143	24143
2005–06						

3364387

3228177

3000666

2877284

27131

26015

2006–07

39220042

3534849

3501313

3149149

31206

28067

2007–08

4561574

3879457

4076878

3451829

35825

30332

2008–09

5270644

4133292

4705447

3664388

40775

31754

2009–10

6070903

4488314

5411104

3966480

46249

33901

2010–11(3R) 7167053

4863886

6406834

4293585

54021

36202

(3R) = 3rd Revision.

Source: *Economic Survey*: 2012-13, Table 17, p A3, and *Economic Survey*: 2013–14, Statistical Appendix, Table 1.1, p.2.

Table 25.2 *Annual Growth Rate of India's GNP and NNP*

(At current prices and at 2004–05 prices)

Gross National Product

Net National Product

Per Capita NNP

(` in Crore)

(` in Crore)

(`)

Year

At Current

At

At Current

At

At Current

At

Prices

Constant

Prices

Constant

Prices

Constant

Prices

Prices

Prices

1999–2000

10.8

7.6

15.8

6.7

13.5

4.6

2000–01

7.5

4.0

10.4

7.5

8.4

5.6

2001–02

9.1

5.7

7.1

3.7

5.2

1.8

2002–03

8.1

4.1

8.9

5.6

6.7

3.5

2003–04

12.1

8.0

8.2

4.0

6.5

2.4

2004–05

14.1

7.0

12.2

8.2

10.6

6.6

2005–06

16.5

9.5

12.9

6.6

11.1

4.9

2006–07

16.4

9.5

14.1

9.4

12.4

7.8

2007–08

15.5

9.7

16.7

9.4

15.0

7.9

2008–09

15.2

6.5

16.4

9.6

14.8

8.1

2009–10

18.4

8.6

15.4

6.2

13.8

4.7

2010–11 (3R)

(3R) = 3rd Revision

Source: *Economic Survey*: 2012–13, Table 1.2, p. A4, and *Economic Survey*: 2013–14, Statistical Appendix, Table 1.2, p.3.

SUMMARY

• National income is defined as the money value of all final gods and services produced over a period of one year. Alternatively, national income is also defined as the sum of factors income (wages, rent, interest and profit) generated over a period of one year.

• National income is measured in terms of (*i*) Gross Domestic Product (*GDP*), and (*ii*) Gross National Product (*GNP*).

• *GDP* is the measure of the total market value of all final goods and services produced in the domestic economy during a period of one year *plus* income earned by the foreigners in the country *minus* income earned by countrymen from abroad.

• *GNP* is the measure of the total market value of all final goods and services produced in the domestic economy during a period of one year *plus* incomes earned abroad by the citizens *minus* income earned by the foreigners in the country.

• Three methods are used to measure the national income: (*i*) Net product or Value Added Method, (*ii*) Factor-income method, and (*iii*) Expenditure method.

• Under *Net Product or value added method*, national income is estimated in three stages: (*i*) estimating gross value of domestic output, (*ii*) estimating cost of production, and (*iii*) estimating depreciation of capital. National income is worked out by deducting cost of production and depreciation from the gross value.

• Under *factor income method,* national income is estimated by adding up all factor incomes. Thus, National Income = Rent + Wages + Interest + Profits. In general, national income is estimated by working out and adding up *labour income, capital income* and *mixed income.*

• Under *expenditure method,* national income is estimated by adding up (*i*) private consumption expenditure, (*ii*) direct tax payments, (*iii*) payments as donation, and (*iv*) private savings.

• In India, national income is estimated by Central Statistical Organization (CSO) by combining the different methods for different sectors of the economy depending on the availability required data.

REVIEW QUESTIONS

1. What is the relevance of national income statistics to business decisions? What kind of business decisions have to be taken by taking national income status?
2. Distinguish between (*i*) economic and non-economic production, and (*ii*) final products and intermediate products. What is their relevance in measurement of national income?
3. Distinguish between (*a*) *GDP* and *GNP*, (*b*) *NNP* and *NDP*, (*c*) *Nominal GNP* and *Real GNP*.
4. Explain different methods of measuring national income. How is the method of estimating national income chosen?
5. Distinguish between net product method and factor income method. Under what conditions are these methods used to estimate national income?
6. Explain the value added method of estimating national income? Why value is added method applied to estimate the national income?
7. What method is used for estimating national income in India? Explain the method of income estimation in India.

FURTHER READING

AbrAhAm, W.I., *National Income and Economic Accounting* (NJ: Prentice-Hall, 1969).
beckermAn, W., *An Introduction to National Income Analysis* (London: English Language Book Society, 1968).
GurGles, R. and ruGGles, N., *National income Accounts and Income Analysis* (New York: McGraw-Hill, 1956).
studensky, P., *The Income Nations: Part II – Theory and Methodology* (Delhi: Khosla & Co., 1969).

CHAPTER

26 Theory of National Income Determination

CHAPTER OBJECTIVES

The objective of this chapter is to describe the framework used for the discussion of the theory of national income determination. The aspects discussed in this chapter are following:

• Basic concepts and economic functions used in theory of income determination
• Derivation of aggregate demand and supply, consumption, and saving functions
• Determination of national income in a simple economy – the two-sector model
• The concept and application of investment multiplier – how change in investment changes the level of national income

26.1 INTRODUCTION

In the preceding chapter, we have discussed the methods of estimating national income. The method of estimating income does not reveal how the level of national income is determined. In this chapter, we proceed to discuss the *theory of national income determination*. The theory of national income determination explains how equilibrium level of national income is determined. This theory was propounded by a British economist, John Maynard Keynes, in 1936 in his book *The General Theory of Employment, Interest and Money*. The Keynesian theory of income determination has been discussed in this chapter. For sake of analytical convenience, the Keynesian theory of income determination is presented in the framework of three models[1]:

1. Two-sector model: households and business sector,
2. Three-sector model: households, business and government sectors, and
3. Four-sector model: households, business, government and foreign sectors.

The fundamental principle of the Keynesian theory of national income determination is that the national income is determined by *aggregate demand* and *aggregate supply* and equilibrium level of national income is determined at the level where aggregate demand equals aggregate supply. The Keynesian theory of income determination has been

1. As has already been shown in 'Circular Flow Models' in Chapter 24, the economy is divided in four sectors: (*i*) households, (*ii*) firms, (*iii*) government, and (*iv*) foreign sector.

discussed in this chapter under the above three models. We begin our discussion on the theory of income determination with two-sector model.

26.2 DETERMINATION OF NATIONAL INCOME: TWO-SECTOR MODEL

26.2.1 Assumptions of Two-sector Model

First*,* there are only two sectors in the economy: households and firms. Households own the factors of production and they sell their factor services or provide them to the firms to earn their living in the form of factor payments—wages, rent, interest and profits. Also, households are the consumers of all final goods and services. The firms, on the other hand, hire factor services from the households and produce goods and services which they sell to the households.

Secondly, there is no government. Or, if government is there, it does not perform any economic function; it does not tax, it does not spend and it does not consume.

Thirdly, the economy is a closed one: there is no foreign trade. It implies that there is no outflow or inflow of goods and services to and from foreign countries.

Fourthly, there are no corporate savings or undistributed (or retained) corporate profits, i.e., the total corporate profit is distributed as dividends.

Finally, prices of all goods and services, supply of labour and capital, and the state of production technology are given and remain constant.

As noted above, according to Keyens, national income of a country is determined by two factors: (*i*) aggregate demand (*AD*) and (*ii*) aggregate supply (*AS*) of goods and services. And, the equilibrium level of national income is determined where *AD* equals *AS*. Before we illustrate graphically the determination of national income, let us explain the concepts of aggregate demand and aggregate supply.

26.2.2 Aggregate Supply

The *aggregate supply* (*AS*) refers to the total value of goods and services produced and supplied in an economy per unit of time. Aggregate supply includes both consumer goods and producer goods. The goods and services produced per time unit multiplied by their respective (constant) prices give the total value of the national output. This is the aggregate supply in terms of money value.

Aggregate Supply Schedule. The Keynesian aggregate supply schedule or aggregate supply curve is drawn on the assumption that total income is always spent – no part of it is retained or withheld. By Keynesian assumption, total expenditure is always equal to the total income. If income increases, expenditure increases by the same amount. This relationship between income and expenditure is shown by a 45° line in Fig. 26.1. This line is also called *aggregate supply schedule*. In the Keynesian theory

Fig. 26.1 *The Aggregate Supply Curve*

of income determination, aggregate income equals consumption (*C*) plus savings (*S*). Therefore, *AS* schedule is generally named as *C* + *S* schedule. The aggregate supply (*AS*) curve is also sometimes called 'aggregate expenditure' (*AE*) curve.

26.2.3 Aggregate Demand

The aggregate demand is an *ex-post* concept. It implies effective demand which equals actual expenditure. The aggregate effective demand is equal to the aggregate expenditure made by the society per unit of time. Aggregate demand (*AD*) consists of two components:

(*i*) aggregate demand for consumer goods (*C*) and

(*ii*) aggregate demand for capital goods (*I*).

Thus,

$$AD = C + I \quad \text{...(26.1)}$$

Aggregate Demand Schedule

The aggregate demand *AD* schedule is also called *C* + *I* schedule. In the Keynesian framework, investment (*I*) is assumed to remain constant in the short-run. But, consumption (*C*) is treated to be a constant function of income (*Y*). Pending detailed discussion on the consumption function till the next section, let us let us note here that the Keynesian consumption function is given as

$$C = a + bY \quad \text{...(26.2)}$$

where *a* denotes a constant denoting *C* when *Y* = 0 and *b* denotes a constant ratio of income consumed, i.e., $b = \Delta C / \Delta Y$.

By substituting Eq. (26.2) in Eq. (26.1), *AD* function can be expressed as

$$AD = a + bY + I \qquad \ldots(26.3)$$

Let us now illustrate the construction of the $C + I$ schedule by assuming:

(i) $C = 50 + 0.5\,Y$, and

(ii) I = `50 billion

By substituting consumption function and I = `50, *AD* function given in Eq. (26.3) can now be written as

$$AD = 50 + 0.5\,Y + 50 = 100 + 0.5\,Y$$

An aggregate demand schedule based on the above assumptions is given in Table 26.1. It shows the relationship between income and aggregate demand ($C + I$). The $C + I$ schedule is plotted in Fig. 26.2.

Table 26.1 *Aggregate Demand Schedule*

(` in billion)

Income (Y)	*C = 50 + 0.5Y*	*I = 50*	*C + I Schedule*
0	50 + 0 = 50	50	100
50	50 + 25 = 75	50	125
100	50 + 50 = 100	50	150
150	50 + 75 = 125	50	175
200	**50 + 100 = 150**	**50**	**200**
250	50 + 125 = 175		

50

225

300

50

+ 150 =

200

50

250

350

50

+ 175 =

225

50

275

400

50

+ 200 =

250

50

300

26.2.4 National Income Determination: Graphical Presentation

Having explained the concept and derivation of aggregate supply and aggregate demand curves, we now turn to present the theory of income determination. The last column of Table 26.1 represents the aggregate demand and the first column represents the aggregate supply. It can be seen in the table that *AS* and *AD* are equal only at one level of income and expenditure, i.e., at `200 billion. The equilibrium level of the national income is therefore determined at `200 billion. If for some reason, *AD* exceeds *AS* or *AS* exceeds *AD*, an adjustment process will bring them back in balance at `200 billion. The information contained in Table 26.1 can be used to illustrate income determination graphically.

The data contained in Table 26.1 is presented graphically in Fig. 26.2. The *AS* schedule is drawn on the assumption that total income (*Y*) is always equal to total expenditure (*E*). The *AS* schedule has, therefore, a constant slope of 1. The *C* + *I* schedule is the vertical summation of the *C* and *I* schedules.

As Fig. 26.2 shows, *C* + *I* and *C* + *S* schedules intersect at point *E* determining the equilibrium level of income at `200 billion. Note that at point *E*,

$AD = AS$

$C + I = C + S$

$150 + 50 = 200$

Thus, the equilibrium level of national income is determined at `200 billion.

Why Not Equilibrium at Any Other Point?

Note that beyond the equilibrium point, *E*, *AD* < *AS* or (*C* + *I*) < (*C* + *S*). It means that at any point on the *AS* schedule beyond point *E*, the firms would be producing more than what households demand. If firms produce goods and services worth more than `200 billion, they will find that they have produced in excess of aggregate demand and their unsold stocks are piling up. For example, suppose firms produce goods and services worth `250 billion. As Table 26.1 shows, this level of output (*AS*) exceeds the aggregate demand (*AD*) by `25 billion. Note that at output or *Y* = `250 billion, *AD* equals `225 billion (see Table 26.1). Therefore, firms' unsold stock equals goods and services worth `25 billion. Hence, they reduce their production and cut down their expenditure on inputs. As a result, the demand for factors of production decreases. This reduces household incomes and, thererfore, their expenditure on goods and services. This process continues until the equilibrium level of income

Fig. 26.2 *National Income Determination*

reaches `200 billion.

Similarly, below `200 billion level of national income, aggregate demand exceeds aggregate supply. The firms, therefore, find that their output is less than what the society is willing to consume. They realize that they could make a greater income by producing and selling a larger output. For example, if firms produce goods worth only `150 billion, they find *AD* exceeding *AS* by `25 billion. That is, demand worth `25 billion remains unsupplied. Firms are, therefore, encouraged to produce more and generate more income to the society. The society, in its turn, spends more as its income increases. The process continues until the equilibrium level of national income is reached. Once the equilibrium level of national income is determined, it is supposed to remain stable.

Having described the theory of national income determination in two-sector model, let us now discuss in detail the relationship between *C* and *Y* and between *S* and *Y* with a view to understanding the process of national income determination. Let us first look into the relationship between income and consumption, generally expressed through consumption function.

26.3 CONSUMPTION FUNCTION

Consumption function is a mathematical expression of the relationship between aggregate consumption expenditure and aggregate disposable income, expressed as $C = f(Y)$. The private demand for goods and services accounts for the largest proportion of the aggregate demand in an economy and plays a crucial role in the determination of national income.

The total volume of private expenditure in an economy depends, according to Keynes, on the total current disposable income of the people and the proportion of income which they decide to spend on consumer goods and services. As mentioned above, this relationship between aggregate consumption demand and the aggregate disposable income is expressed through a 'consumption function' expressed as

$$C = a + bY \quad \text{...(26.4)}$$

where C = aggregate consumption expenditure; Y = total disposable income; a = a constant consumption term; and $b = \Delta C/\Delta Y$, i.e., the proportion of marginal income spent on consumption.

The consumption function given in Eq. (26.4) is based on the assumption that there is a constant relationship between consumption and income, as denoted by constant 'b' denoting 'marginal propensity consumer'.

It may be added here that the original Keynesian function assumes a decreasing ratio between consumption and income. According to Keynes, the consumption function stems from a 'fundamental psychological law'. The law states that marginal propensity to consume ($\Delta C/\Delta Y$) decreases with the increase in income in the short-run. This law implies that total consumption increases but not by an equal amount of increase in income. This Keynesian hypothesis of income-consumption relationship was later termed as the *absolute income hypothesis*. Some early empirical studies[2] based on cross-section and time-series data have supported the hypothesis.

2. Tobin, James, 'Relative Income, Absolute Income and Saving," in *Money, Trade and Economic Growth*, (Macmillan, 1951); and Arthur Smithies, "Forecasting Postwar Demand: I," *Econometrica*, January 1945.

However, the economists have found empirically that Keynesian consumption function may be applicable to individual consumption behaviour but not for the aggregate consumption expenditure. They have found empirically that, at the aggregate level, consumption function is a linear function. It is now a convention to use a linear consumption function[3] at the aggregate level, as given in Eq. (26.4) in the analysis of income determination.

26.3.1 Propensity to Consume

The propensity to consume refers to the proportion of the total and the marginal incomes which people spend on consumer goods and services. The proportion of the marginal income consumed is called 'Marginal Propensity to Consume' (*MPC*), and the proportion of the total income consumed is called 'Average Propensity to Consume' (*APC*). Let us now discuss these concepts in detail.

(a) The Marginal Propensity to Consume (MPC). The concept of *MPC* is related to the marginal consumption-income relationship. In other words, *MPC* refers to the relationship between change in consumption (ΔC) and the change in income (ΔY). Symbolically, $MPC = \Delta C/\Delta Y$.

As mentioned above, according to the consumption function envisaged by Keynes, marginal propensity to consume ($\Delta C/\Delta Y$) decreases with increase in income. In the theory of income determination, however, a ***constant*** marginal propensity to consume is assumed. For example, suppose that income increases from `200 tò300, and as a result, consumption increases from `250 to

Fig. 26.3 *Income Consumption Relationship*

`325, as shown in Fig. 26.3. Thus, the change in income $\Delta Y = 300 - 200 = 100$, and change in consumption, $\Delta C = 325 - 250 = 75$. Thus,

$MPC = \Delta C/\Delta Y = 75/100 = 0.75$

Similarly, if income increases from `300 tò400, and consumption expenditure rises from `325 tò400, the $MPC = 75/100 = 0.75$. This kind of relationship between income and consumption is expressed through a ***linear consumption function***, as shown by the line marked *C* in Fig. 26.3.

The *MPC* can be derived from the consumption function as follows. Given the consumption function in Eq. (26.4),

$C = a + bY,$

3. We shall also assume a linear consumption function, i.e., a constant marginal propensity to consume.

Let *Y* increase by ΔY so that

$C + \Delta C = a + b(Y + \Delta Y) = a + bY + b\Delta Y$

and

$\Delta C = -C + a + bY + b\Delta Y$

Since $C = a + bY$, by substituting $a + bY$ for *C*, we get

$\Delta C = -(a + bY) + a + bY + b\Delta Y$

$\Delta C = b\Delta Y$ ...(26.5)

By dividing both sides of Eq. (26.5) by *Y*, we get

ΔC

$MPC = \Delta Y = b$

According to the Keynesian theory of aggregate consumption, $\Delta C/\Delta Y = b$ is always less than unity, but greater than zero, i.e., $0 < b < 1$. This fundamental relationship between income and consumption plays a crucial role in the Keynesian theory of income determination.

(b) Average Propensity to Consume (APC). Average Propensity to Consume is defined as the proportion of total income spent on consumer goods and services, i.e.,

$APC = CY$

where *C* is total consumption expenditure and *Y* is total disposable income. Given the consumption function, $C = a + bY$, *APC* can be obtained as

$C\ a + bY$

$APC =$

=

Y

Y

If consumption function is given as

bY

$C = bY$ then $APC =$

$= b$

Y

Note that if consumption function is given as $C = bY$ (i.e., without constant term ' *a*'),

then $APC = b = MPC$.

26.3.2 Properties of Consumption Function

The Keynesian consumption function has the following properties.

1. It states the relationship between consumption expenditure and disposable income.

If consumption function is empirically estimated for a country, total consumption

expenditure can be predicted if growth rate of income is known and income

distribution is given.

2. It states that income-consumption relation is given by the *MPC* (denoted by *b*),

under the assumption that $0 < b < 1$.

3. Consumption function of the form, $C = a + bY$ or $C = bY$ implies a linear

relationship between consumption and income, i.e., a constant *MPC*.

4. Consumption function implies a saving function. That is, if consumption function

is known, the saving function can easily be obtained.

26.4 DERIVATION OF SAVING FUNCTION

Having explained the Keynesian aggregate consumption function, we turn to derive the

Keynesian saving function in this section. Like consumption, saving (*S*) is also the function

of income (*Y*), i.e.,

$S = f(Y)$

Since $Y = C + S$, consumption and saving functions are counterparts of each other.

Therefore, if one of these functions is known, the other can be easily obtained. For

example, if consumption function is given as $C = a + bY$, then saving function can be

derived as follows.

We know that

$S = Y – C$

...(26.6)

By substituting consumption function, $C = a + bY$ for *C* in Eq. (26.6), we get

$S = Y – (a + bY)$

$= – a + (1 – b) Y$

...(26.7)

Equation (26.7) gives the saving function in which '1 – *b*' is **marginal propensity**

to save (*MPS*). It can be proved as follows:

Since

$Y = C + S \therefore \Delta Y = \Delta C + \Delta S$

Dividing both sides by *Y*, we get

ΔC ΔS

+

ΔS

ΔC

=

ΔY ΔY = 1 or

1–

ΔY

ΔY

ΔC

Since

$= b$, by substitution, we get

ΔY

$MPS = \Delta S$ or $MPS = 1 – b$

ΔY

Numerical Example. Let us now show the derivation of saving function through a numerical example. Let consumption function be given as

$$C = 100 + 0.75\,Y \qquad ...(26.8)$$

Given the Eq. (26.8), Eq. (26.6) can be written as

$$S = Y - (100 + 0.75\,Y)$$
$$= Y - 100 - 0.75\,Y$$
$$= -100 + (1 - 0.75)\,Y$$
$$= -100 + 0.25\,Y \qquad ...(26.9)$$

The consumption and saving functions are graphed in Fig. 26.4. The 45° line shows income-consumption relation with $Y = C$ at all levels of income. In the analysis of national income determination, it also shows the total sales proceeds, i.e., the value of the total planned output. The schedule $C = 100 + 0.75\,Y$ gives the income-consumption relationship—consumption being a linear function of income. The schedule $S = -100 + 0.25\,Y$ is the saving schedule derived from the consumption schedule. The saving schedule shows the income-saving relationship.

Fig. 26.4 *Income, Consumption and Savings Schedules*

26.5 TWO-SECTOR MODEL OF NATIONAL INCOME DETERMINATION: ALGEBRAIC TREATMENT

In preceding sections, we have presented the Keynesian theory of income determination in its simplest form and have explained the derivation of the consumption and saving functions. In this section, we present the two-sector model of income determination in its formal form.

As stated above, equilibrium level of national income is determined at the level of income at which aggregate demand for output ($C + I$) is equal to aggregate supply of incomes ($C + S$). Thus, equilibrium condition of national income is given as

Aggregate Demand = Aggregate Supply,

or

$$C + I = C + S \qquad ...(26.10)$$

Since C is common to both the sides, the equilibrium conditions can also be stated as

$$I = S \qquad ...(26.11)$$

Given these conditions of equilibrium, there are two alternative ways to show the determination of national income:

(*i*) by using aggregate demand ($C + I$) and aggregate supply ($C + S$) schedules, and

(*ii*) by using only saving (S) and investment (I) schedules.

The two approaches are known as **income-expenditure approach** and **saving-investment approach**, respectively. Let us now explain in detail the determination of national income by the two approaches.

26.5.1 Income-Expenditure Approach

According to the income-expenditure approach or, what is also called 'aggregate demand and aggregate supply approach', the equilibrium of national income is determined by

$$C + I = C + S$$

Since $C + S = Y$, the national income equilibrium condition can also be restated as

$$Y = C + I$$

Since at equilibrium, $C = a + bY$, by substitution, we get national income equilibrium condition as

$$Y = a + bY + I \text{ or } Y(1 - b) = a + I$$

1

Therefore,

$$Y = \frac{(a + I)}{1 - b}$$

...(26.12)

Suppose empirical consumption function is given as

C = 100 + 0.75 Y and

I = 100

1

Then

Y = 100 + 0.75 Y + 100 =

(100 +100)

1– 0.75

1

=

(200)= 800

0.25

Thus, given the consumption function, as in Eq. (26.8) and investment at 100, the national income equilibrium is determined at `800.

Determination of equilibrium level of national income by aggregate demand and aggregate supply approach is also presented graphically in Fig. 26.5. The $C + S$ schedule represents the aggregate supply of income and C and I schedules represent, consumption and investment functions, respectively. The $C + I$ schedule, i.e., the aggregate demand schedule, is formed by vertical summation of C and I schedules. The $C + I$ and $C + S$ schedules intersect at point E which is the equilibrium point. At this point,

$Y = C + I$

800 = 700 + 100

Once national income is determined, it will remain stable in the short-run. Any production in excess of or below the equilibrium output will create conditions for the income and expenditure to return to the equilibrium position, E. For, the expectations of businessmen are realized only when aggregate expenditure equals aggregate income. While aggregate supply ($C + S$) represents the aggregate value expected by business firms, aggregate demand ($C + I$) represents their realized value. At equilibrium, *expected value* equals *realized value*.

Fig. 26.5 *Determination of National Income: Income-Expenditure Approach*

As mentioned above, production in excess of equilibrium output will result in undesired accumulation of inventories which reduces profits. For example, if goods and services worth `1,000 are produced, the unsold stock will equal `50, because, at this level of income, society plans to spend only `950.4 This will force the business firms to cut down their output and, return to the point of equilibrium output through the process of *reverse multiplier*.5 Similarly, when production is below the equilibrium level, realized value exceeds the expected value. This gives incentive to produce more and make larger profit, and to reach the equilibrium level through the process of multiplier.

26.5.2 Saving-Investment Approach

The determination of national income can also be explained by saving-investment approach, i.e., by using only saving (S) and investment (I) schedules. We have noted that national income equilibrium is determined where $I = S$. By our earlier assumptions, I = 100, and consumption function as,

C = 100 + 0.75 Y

Given the consumption function and I = 100, the saving function can be written as

S = – 100 + 0.25 Y

Given the saving function and investment, equilibrium of national income gets determined where $I = S$, i.e., where

100 = – 100 + 0.25 Y

...(26.13)

4. When Y = `1,000, $C + I$ = 100 + 0.75 (1000) + 100 = 950.

5. The concept and theory of multiplier is discussed in the next section.

Solving Eq. (26.13) for Y, we get national income equilibrium at

Y = 800

Obviously, the saving-investment approach determines the same equilibrium level of national income (`800) as the income-expenditure approach.

Determination of national income by saving-investment approach is illustrated in Fig. 26.6. Here, S-schedule represents the saving function, $S = -100 + 0.25\,Y$, and I-schedule is drawn as, $I = 100$.

Fig. 26.6 *Determination of National Income: Saving and Investment Approach*

The S and I schedules intersect at point E where *planned saving* equals *planned investment* and equilibrium of national income is determined at `800 which is the same as one determined by income-expenditure approach.

26.6 INCREASE IN INVESTMENT AND INVESTMENT MULTIPLIER

In the preceding section, we have explained the determination of national income equilibrium under the condition of a given aggregate demand schedule, $C + I$, assuming consumption function and investment to be given. In this section, we explain the effect of shifts in the aggregate demand schedule on the equilibrium level of national income confining our analysis only to a *two-sector model*. A shift in the aggregate demand schedule, in a two-sector economy may be caused by a shift in consumption schedule or in investment schedule or both. Consumption expenditure is, however, found to be a stable function of income whereas investment may change due to autonomous factors. It is, therefore, generally assumed that the shift in the aggregate demand schedule takes place due to a shift in the investment schedule. Let us assume that aggregate demand schedule shifts upward due to a permanent upward shift in the investment schedule. The increase in investment may be the result of an *autonomous investment* in some business adventures.

Fig. 26.7 *Shift in Aggregate Demand Function and Increase in National Income*

The economy being in equilibrium, an increase in investment causes an upward permanent shift in aggregate demand schedule. An upward permanent shift in the aggregate demand schedule leads to an increase in national income, as shown in Fig. 26.7. The initial aggregate demand schedule is shown by $C + I$ schedule. It intersects aggregate supply schedule ($C + S$) at point E_1 where the equilibrium level of national income is determined at Y_1. Let us suppose now that investment (I) increases causing an upward shift in investment schedule from I to $I + \Delta I$. This causes an upward shift in aggregate demand schedule from $C + I$ to $C + I + \Delta I$. With the shift in aggregate demand schedule, the equilibrium point of national income shifts from E_1 to E_2 and the equilibrium level of national income increases from Y_1 to Y_2. The increase in national income (ΔY) may be obtained as

$\Delta Y = Y_2 - Y_1$

The increase in the national income, ΔY, is the result of ΔI. Note that in Fig. 26.7, ΔY is much greater than ΔI. A question arises here: 'Is there any definite relationship between ΔY and ΔI?' If yes, what determines that relationship? These questions take us to the theory of ***investment multiplier.***

26.6.1 Investment Multiplier

To workout the investment multiplier, let us first look at the relationship between ΔY and ΔI. This can be done by comparing the two equilibrium levels of national income in Fig. 26.7.

At equilibrium point E_1, $Y_1 = C + I$

Since $C = a + bY$, by substitution, we get

$Y_1 = a + bY_1 + I$

$= 1\ (a + I)$

...(26.14)

$1 - b$

Similarly, at equilibrium E_2,

$Y_2 = C + I + \Delta I = a + bY_2 + I + \Delta I$

$= 1\ (a + I + \Delta I)$

...(26.15)

$1 - b$

By subtracting Eq. (26.14) from Eq. (26.15), we get

1

1

$\Delta Y =$

(a + I + Δ I)−

(a + I)

1− b

1− b

Δ Y = 1 Δ I

...(26.16)

1− b

Equation (26.16) gives the relationship between Δ Y and Δ I. It reveals that Δ Y is 1/(1 – b) times Δ I. Therefore, 1/(1 – b) is the investment *multiplier* (*Im*). The value of multiplier can be obtained by dividing both sides of Eq. (26.16) by Δ I. That is,

Im = Y

Δ

1

=

...(26.17)

Δ I 1 – b

Thus, multiplier

(*Im*) = 1

...(26.18)

1– b

The *multiplier* may thus be defined as *the ratio of the change in national income due to change in investment.* Since Δ Y is the result of Δ I, the *multiplier* as measured in Eq. 26.18 is called ***investment multiplier.***

Determinant of Investment Multiplier

Note that in Eq. (26.18) ' b' stands for the *MPC* (i.e., b = Δ C/Δ Y). It may, therefore, be concluded that *MPC* is the determinant of the value of the multiplier. The higher the *MPC*, the greater the value of multiplier. This relationship can be tabluted as shown below.

MPS and the Multiplier

The value of multiplier can also be obtained through the *marginal propensity to save*, i.e., *MPS*. In Eq. (26.18), 1 – b is the same as 1 – *MPC*.

We know that 1 – *MPC* = *MPS*. Therefore,

1

1

Im =

=

...(26.19)

1− *MPC* *MPS*

MPC

m

Numerically, if *MPC* = 0.75, *MPS* = 0.25, then

0.00

1.00

multiplier,

0.10

1.11

0.50

2.00

1

1

Im =

=

= 4

0.75

4.00

1− *MPC* 0.25

0.80

5.00

0.90

10.00

The multiplier may, therefore, be defined also as the

1.00

∞

reciprocal of *MPS*. If *MPS* is known, investment multiplier

(*Im*) can be easily obtained.

26.6.2 Static and Dynamic Multiplier

Sometimes, from application point of view, a distinction is made between *static multiplier* and *dynamic multiplier*. In this section, we explain the difference between the static and dynamic multipliers and also describe the process of dynamic multiplier.

Static multiplier is also known as 'comparative static multiplier', 'simultaneous multiplier', 'logical multiplier', 'timeless multiplier' or 'lagless multiplier'. The concept of static multiplier assumes that the change in investment and the resulting change in income are simultaneous. There is no time lag between the change in investment and the resulting change in income. In other words, the shift of national income equilibrium from point E 1 (in Fig. 26.7) to point E 2 due to change in investment (ΔI) has no time-lag. Static multiplier also assumes that there is no change in *MPC* at the aggregate level as the economy moves from one equilibrium position to another. It ignores the process by which changes in income and consumption expenditure lead to a new equilibrium. Also, static multiplier assumes income distribution and consumers' preferences to remain unchanged.

The concept of ***dynamic multiplier*,** or what is also known as **period multiplier** or 'sequence' multiplier, does not make the assumptions of the static multiplier. Dynamic multiplier traces the process by which equilibrium of national income shifts from one position to another. In real life, income level does not increase instantly when an autonomous investment is made because there is a time-lag between increase in income and consumption expenditure.

The process of dynamic multiplier may be described as follows. Suppose that autonomous investment increases by `100, i.e., $\Delta I$ = `100. Assume also that *MPC* = 0.8, and there is no expenditure other than the consumption expenditure.

When autonomous investment increases by `100, it subsequently increases the income of the recipients by `100, i.e., in the *first* round of expenditure-income process, $\Delta I = 100 = \Delta y$ 1. The recipients of `100 spend `80 (= 100 × 0.8) on consumer goods and services. In the *second* round, those who supply goods and services worth `80, receive an additional income of `80. That is, Δy 2 = 80. Of this, they spend `64 (= 80 × 0.8). This results in an additional income ($\Delta y$ 3) of `64 to those who supply consumer goods and services. This process continues till the value of $\Delta y \to 0$. Note that the value of Δy decreases in the subsequent rounds of income and expenditure so that Δy 1 > y 2 > y 3.... The whole series of Δy generated by $\Delta I = 100$ may be written as

$\Delta Y = \Delta y$ 1 + Δy 2 + Δy 3 ... Δyn–1

ΔY = 100 + 100(0.8) + 100(0.8)2 + 100(0.8)3... + 100(0.8) n–1

= 100 + 80 + 64 + 51.20......→ 0 = 499.999 = 500

Given the income (ΔY) generated over time, the investment multiplier (*Im*) can be obtained as

$Im = \Delta Y$ 500

=

= 5

ΔI

100

The process of ***dynamic multiplier*** may be generalized as follows. The whole series of additional incomes caused by ΔI over time may be written as

$\Delta Y = \Delta y + \Delta y(b) + \Delta y(b)2 + \Delta y(b)3 ... \Delta y(b)$ n–1

= Δy(1 + b + b 2 + b 3 . . . bn–1)...

The sum of these values is obtained6 as

⎛ 1

= y

⎞

Δ | 1– *b* |

⎝

⎠

...(26.20)

Since Δ *y* = Δ *I*, we may rewrite Eq. (26.20) as

1

= Δ *I* 1– *b*

...(26.21)

From this equation, the investment multiplier (*Im*) may be obtained as

Δ *Y*

1

Im =

=

Δ *I*

1– *b*

...(26.22)

Note that dynamic multiplier is the same as static multiplier.

26.6.3 Limitations of Multiplier

Despite its important uses in macroeconomic analysis, the concept of multiplier has certain *limitations*, which should be borne in mind while using this concept.

First, the limitation of the multiplier theory is related to the rate of *MPC*. If the rate of *MPC* is lower in an economy, the rate of multiplier will also be lower. As a corollary of this, since *MPC* in less developed country is comparatively higher, the multiplier there must be higher than in the developed countries. This may, however, not be true in real practice because of other limitations of multiplier.

Secondly, the working of multiplier assumes that those who earn income firm autonomous investment would continue to spend a certain percentage of their newly earned income on consumption and that there are no leakages in the expenditure process. This assumption may not hold in real practice since people may like to spend a part or whole of their additional income on

6. **The process of estimation:** The series of incomes is given as

Δ *Y* = Δ *y* + Δ *y*(*b*) + Δ *y*(*b*)2 + Δ *y*(*b*)3 ... Δ *y*(*b*) *n*–1

...(*i*)

Now let the terms inside the bracket be summed up as

S = 1 + *b* + *b* 2 + *b* 3 + ... *bn*–1

...(*ii*)

By multiplying both sides of this equation by *b*, we get

S.b = *b* + *b* 2 + *b* 3 + *b* 4 + ... *bn*

...(*iii*)

If Eq. (*iii*) is substrated from Eq. (*ii*), all terms, except 1 and *bn*, on the right hand side of Eq. (*iii*) cancel out.

Then

S – *S.b* = 1 – *bn* or *S*(1 – *b*) = 1 – *bn*

1 – *n*

b

S = 1– *b*

...(*iv*)

1

When *n* → ∞, *bn* → 0. It can, therefore, be omitted. Then *S* = 1– *b*

...(*v*)

By substituting Eq. (*iv*) into Eq. (*i*) for the terms in the bracket, we get

⎛ 1

Y

y

⎞

Δ = Δ | 1− *b* |

⎝

⎠

(*i*) payment of past debts;
(*ii*) purchase of existing durable consumer goods and other assets, like old houses, second hand cars, etc.
(*iii*) purchase of shares and bonds from the shareholders or bond-holders, and
(*iv*) purchase of imported goods.

These are known as **leakages** in the consumption flows, which reduce the rate of multiplier. For example, let us suppose that a building contractor earns `50,000 from a contract, which he pays to his creditor. His creditor buys an old house. The person who sells the house buys of second hand imported car. The money thus keeps circulating but is never spent in the manner that can generate demand for new consumer goods. In that case, multiplier will be 1. The other leakages are holding idle cash, deposits in foreign banks, etc.

Thirdly, the working of multiplier is based on the assumption that the goods and services are always available in adequate supply. But, if goods and services are in scarcity, the real consumption expenditure will be reduced due to inflation whatever the rate of *MPC*. Consequently, the multiplier will be reduced. If expenditure continues to increase in face of scarcity, it generates inflation, while real income does not increase.

Finally, under the condition of full-employment, the theory of multiplier will not work because additional goods and services cannot be produced or additional real income cannot be generated.

Despite its limitations, the concept of multiplier is an important tool of analyzing the process of multiple increase in income caused by increase in investment. The concept multiplier is meaningfully applied to measure the impact of public expenditure, taxation and foreign trade.

26.7 INCOME DETERMINATION IN THREE-SECTOR MODEL: A MODEL WITH GOVERNMENT

We have explained above income determination in a simple, two-sector economy. A real economy, however, also contains two other major sectors, viz., the government sector and the foreign sector. In this section, we explain income determination in a **three-sector model** including household, business and government sectors. The **four-sector model** including three domestic sectors and the foreign sector will be discussed in the next section.

Let us assume an economy consisting of three sectors, viz., households, firms and government. Inclusion of the government into the model brings in two variables: (*i*) *government expenditures*, and (*ii*) taxation. While governmnet expedntiures are injections to the income stream, *taxation* is *withdrawal* from the income stream. Government expenditure adds to the aggregate demand, and taxation, on the other hand, reduces the aggregate demand. How the level of national income is affected by the inclusion of government expenditure and taxation into the model is discussed below.

To begin with, let us incorporate government sector into our two-sector model, assuming that the government follows a *balanced budget* policy, i.e., government expenditure (G) equals the amount of taxes (T), and both G and T are exogenously determined. Let us now redefine the parameters of the aggregate demand and aggregate supply for the three-sector model.

Aggregate demand (AD) in the three-sector model can be expressed as

$$AD = C + I + G$$

and aggregate supply (AS) as

$$AS = C + S + T$$

The equilibrium of national income is determined where

$$AS = AD$$

Thus, at equilibrium,

$$Y = C + S + T = C + I + G$$

Since $S = I$ and $T = G$, at equilibrium, $Y = C + I + G$

...(26.23)

In Eq. (26.23), $C = a + bYd$ where Yd is disposable income, i.e., $Yd = Y – T$ and

T = lump-sum tax

By substituting consumption function $a + bYd$ for C in Eq. (26.23) and $Y – T$ for Yd in the resulting equation, the quilibrium equation can be written as

$Y = a + b(Y – T) + I + G$

...(26.24)

Solving for *Y*, we get equilibrium level of national income as

$Y = a + bY – bT + I + G$

$Y(1 – b) = a – bT + I + G$

$Y = 1 (a - bT + I + G)$

...(26.25)

$1- b$

Alternatively, the equilibrium condition of national income can be derived by *saving-investment approach* as follows. Consider the equilibrium condition given as

$C + S + T = C + I + G$

Since *C* on each side gets cancelled out, equilibrium equation can be rewritten as

$S + T = I + G$

For a numerical example, recall our earlier consumption function and constant investment from section 26.5.1.

$C = 100 + 0.75\ Y$ and $I = 100$

Let us also assume that in the balanced budget of the government

$G = T = 50$

By substituting these values in Eq. (26.25), we get equilibrium level of national income as

1

$Y =$

$(100 - 0.75\times 50 + 100 + 50) = 850$

$1- 0.75$

The three-sector model of income determination is presented graphically in Fig. 26.8. As the figure shows, the aggregate demand schedule (*C* + *I* + *G*) intersects with the aggregate supply schedule (*C* + *S* + *T*) at point *E* determining the equilibrium of national income at `850. The same level of equilibrium income is determined by the *I* + *G* and *S* + *T* schedules intersecting at point *E*′. Let us now see what happens if the government decides to increase its expenditure.

26.7.1 Change in Government Expenditure and the Government Expenditure Multiplier

Let us now analyze the impact of *change* in government expenditure on the national income. Let us assume that government expenditure is confined to *only purchase of goods and services* and that all other variables remain constant. The impact of change in government expenditure on the level of national income is similar to the autonomous change in investment. Recall Eq. (26.25) which gives the equilibrium level of national income in a three-sector model.

Given the national income equilibrium condition in Eq. (26.25), let the government expenditure increase by ΔG. Increase in government expenditure increases aggregate demand through a process of multiplier. The equilibrium level of national income with ΔG can be expressed as follows.

$Y + \Delta Y = 1 (a - bT + I + G + \Delta G)$

...(26.26)

$1- b$

Fig. 26.8 *Determination of National Income in Three-Sector Model*

The effect of ΔG on the level of national income can be obtained by subtracting Eq. (26.25) from Eq. (26.26). By subtraction, we get

$\Delta Y = 1 (\Delta G)$

...(26.27)

$1- b$

By rearranging the values in Eq. (26.27), we get ***government expenditure multiplier*** (*Gm*) as given below.

G

ΔY

m =

$$\frac{}{\Delta G} = \frac{1}{1-b} \qquad \ldots(26.28)$$

Note that government expenditure multiplier (*Gm*) is the same as investment multiplier given in Eq. 26.17.

26.7.2 Change in Lump-Sum Tax and the Tax Multiplier

The method of analyzing the impact of a change in lump-sum taxes on the level of income is the same as in the case of change in government expenditures.

To find out the impact of change in taxes on the level of national income, recall once again the equilibrium condition of national income (without ΔT or ΔG) given in Eq. (26.25) reproduced here as

$$Y = \frac{1}{1-b}(a - bT + I + G) \qquad \ldots(26.29a)$$

A change in tax by ΔT with constant G causes a change in Y by ΔY. By incorporating ΔT and ΔY in Eq. [26.29(a)], the equilibrium level of income may be expressed as

$$Y + \Delta Y = \frac{1}{1-b}[a - b(T + \Delta T) + I + G]$$

$$Y + \Delta Y = \frac{1}{1-b}[a - bT - b\Delta T + I + G] \qquad \ldots(26.29b)$$

Note the effect of ΔT on the level of national income can be obtained by subtracting Eq. 26.29a from Eq. 26.29b. By subtracting, we get

$$\Delta Y = \frac{1}{1-b}(-b\Delta T) = \frac{-b\Delta T}{1-b} \qquad \ldots(26.30)$$

The **tax-multiplier** (*Tm*) can now be obtained by dividing both sides of Eq. (26.30) by ΔT. Thus,

$$T_m = \frac{\Delta Y}{\Delta T} = \frac{-b}{1-b} \qquad \ldots(26.31)$$

Note that the tax multiplier (*Tm*) bears a negative sign. It is, therefore, negative. It means that tax and increase in tax have a negative impact on the national income. Note also that since $b = MPC$ and $MPC < 1$,

$$T_m = -\left[b\right] < \left[1\right] = G_m$$

|⎣

1

m

b ||

⎦⎣

1 *b* |

−

− ⎦

The comparison of T_m and G_m shows that, *T*-multiplier is smaller than *G*-multiplier. It means that if $\Delta G = \Delta T$, national income will increase. But, by how much? This question can be answered by looking at the balanced budget multiplier.

26.7.3 Balanced Budget Multiplier

We have discussed above *G*-multiplier assuming taxes (*T*) to remain constant, and *T*-multiplier assuming *G* to remain constant. In this section, we discuss the impact of a simultaneous change in *G* and *T* by an equal amount on the level of national income.

When $\Delta G = \Delta T$, the government budget is said to be in balance. The effect of **balanced budget** on the national income is analyzed by **balanced budget theorem** or **balanced budget multiplier effect.** The balanced budget theorem states that **the balanced budget multiplier is always equal to one**. That is why the balanced budget theorem is also called *unit multiplier theorem*. A proof of the theorem can be provided as follows.

Recall again the equilibrium Eq. (26.29a), i.e.,

$$Y = \frac{1}{1-b}[a - bT + I + G]$$

By incorporating ΔG and ΔT (while $\Delta G = \Delta T$) and the resulting change in income (ΔY), in Eq. (26.29a), we get equilibrium level of income as

$$Y + \Delta Y = \frac{1}{1-b}[a - b(T + \Delta T) + I + G + \Delta G] \qquad ...(26.32)$$

By subtracting Eq. (26.29a) from Eq. (26.32), we get

$$\Delta Y = \frac{1}{1-b}(-b\Delta T + \Delta G) \qquad ...(26.33)$$

Since $\Delta T = \Delta G$, by substituting ΔG for ΔT, we can rewrite Eq. (26.33) as

$$\Delta Y = \frac{1}{1-b}(-b\Delta G + \Delta G)$$

By rearranging the terms, we get

$$\Delta Y(1 - b) = -b\Delta G + \Delta G$$

$$\Delta Y(1 - b) = \Delta G(1 - b)$$

$$\Delta Y = \Delta G \qquad ...(26.34)$$

The balanced budget multiplier (B_m) can be obtained by dividing both sides of Eq. (26.34) by ΔG. That is,

$$\frac{\Delta Y}{\Delta G} = B_m = \frac{\Delta G}{\Delta G} = 1 \qquad ...(26.35)$$

An Alternative Method The balanced budget multiplier (B_m) can also be obtained by adding up the *G*-multiplier and *T*-multiplier. As shown below,

− *b*

Gm = 1 and *T*

1− *b*

m = 1− *b*

Given the *Gm* and *Tm*,

∴

Bm = *Gm* + *Tm*

1

− *b*

=

+

1− *b* 1− *b*

− *b*

= 1

= 1

...(26.36)

1− *b*

Note that *the balanced budget multiplier (Bm) is equal to unity*. It means that *if*

Δ G = Δ T, national income increases exactly by the amount of increase in government

expenditure.

26.7.4 Proportional Income Tax and Balanced Budget Multiplier

We have so far analyzed the effects of balanced budget with an autonomous lump-sum tax. In reality, however, tax system consists of both lump-sum and proportional taxes. In this section, we analyze the effect of balanced budget with a proportional income tax. The tax function used for the purpose is expressed as

T = *T* + *tY*

...(26.37)

where *T* = autonomous tax and *t* = a constant rate of income tax.

Let us begin the analysis by recalling equilibrium Eq. (26.24), given as

Y = *a* + *b*(*Y* – *T*) + *I* + *G*

...(26.38)

By substituting Eq. (26.37) for *T* in Eq. (26.38), we get

Y = *a* + *b* [*Y* − (*T* + *tY*) + *I* + *G*

...(26.39)

= *a* + *bY* − *bT* − *btY* + *I* + *G*

1

=

(*a* − *bT* + *I* + *G*)

1− *b* + *bt*

1

=

(*a* − *bT* + *I* + *G*)

1− *b*(1− *t*)

...(26.40)

Now let the government expenditure increase by Δ *G* causing an increase in national income by Δ *Y*. New equilibrium level of national income is given as

1

Y + Δ *Y* =

(*a* − *bT* + *I* + *G* + Δ *G*)

1− *b*(1− *t*)

...(26.41)

By subtracting Eq. (26.40) from Eq. (26.42), we get

Δ *Y* =

1

(Δ *G*)

...(26.42)

$1 - b(1 - t)$

By dividing both sides of Eq. (26.20) by ΔG, we get ***balanced budget multiplier*** with proportional tax as

$$\frac{\Delta Y}{\Delta G} = \frac{1}{1 - b(1 - t)} \quad ...(26.43)$$

Note that the balanced budget multiplier with a proportional tax, i.e., $1/[1 - b(1 - t)]$ is smaller than the balance budget multiplier (1) with only lump sum tax.

26.8 INCOME DETERMINATION IN FOUR-SECTOR MODEL: MODEL WITH FOREIGN TRADE

We have been concerned so far with the theory of income determination in a 'closed economy', i.e., on economy without foreign trade. In this section, we will discuss income determination in an 'open economy'. An open economy is one which carries out economic transactions with the rest of the world—mainly foreign trade. Inclusion of foreign trade makes the model a complete, four-sector model. In foreign trade, exports are injections and imports are outflows from the circular flows of incomes. In the national income analysis, however, only trade balances ($X - M$) is taken into account. That is, only $X - M$ is incorporated in the four-sector model of income determination. Injections increase the level of income and withdrawals decrease it. Therefore, if $X > M$, there is net injection and national income increases, and if $X < M$, there is a net withdrawal which causes national income to decrease.

To begin the analysis of income determination with foreign sector, let us first specify the export and import functions.

26.8.1 Export Function

Exports of a country depend on such factors as: (*i*) prices of domestic goods in relation to prices of goods in importing countries; (*ii*) tariffs and trade policies of importing countries; (*iii*) export subsidies; (*iv*) income elasticity for imports in the importing countries, and (*v*) level of imports by the domestic economy, etc.

The factors other than (*i*) and (*iii*) are beyond the control of the economy and hence are not included among the policy variables. For the sake of simplicity, therefore, it is assumed that exports (X) are determined by the factors operating outside the economy. In the income determination model, therefore, X is treated as an autonomous variable and assumed to be given as X .

26.8.2 Import Function

Imports of a country, like its exports, are determined by (*i*) import prices in relation to domestic prices; (*ii*) the level of domestic tariffs; (*iii*) domestic trade policy; (*iv*) income-elasticity of imports; (*v*) the level of incomes and (*vi*) the level of exports.

For analytical purpose, however, a simplifying assumption is made that imports (M) depend on the level of domestic income (Y) and on the *marginal propensity to import* (MPM). Under this assumption, import function is expressed as

$$M = M + gY \quad ...(26.44)$$

where M = autonomous imports, $g = \Delta M/\Delta Y = MPM$ (assumed to be constant).

Having specified the X and M functions, we may now specify the four-sector model of income determination as given below.

$$Y = C + I + G + (X - M) \quad ...(26.45)$$

where

$C = a + bYd$

$$\left.\begin{array}{l} I = I \\ G = G \\ X = X \end{array}\right\} \text{These variables are constant}$$

$Yd = Y - T$

$M = M + gY$

By incorporating these variables and functions in Eq. (26.45), the equilibrium equaiton can be written as

$$Y = a + b(Y - T) + I + G + X - (M + gY)$$

$$= \frac{1}{1 - b + g}(a - bT + I + G + X - M) \quad \text{...(26.46)}$$

The term $1/(1 - b + g)$ in Eq. (26.46) is the foreign trade multiple, on the assumption that consumption and imports are both a linear function of domestic income. This can be proved as follows. Let exports (X) increase by ΔX, while other variables remain constant. The equilibrium level of national income can then be written as follows.

$$Y + \Delta Y = \frac{1}{1 - b + g}(a - bT + I + G + X - M + \Delta X) \quad \text{...(26.47)}$$

or

$$Y + \Delta Y = \frac{1}{1 - b + g}(a - bT + I + G + X - M) + \frac{1}{1 - b + g}\Delta X$$

By subtracting Y from both sides, we get

$$\Delta Y = \frac{1}{1 - b + g}\Delta X \quad \text{...(26.48)}$$

By rearranging Eq. (26.48), we get foreign trade multiplier ($\Delta Y/\Delta X$) as

$$\frac{\Delta Y}{\Delta X} = \frac{1}{1 - b + g} \quad \text{...(26.49)}$$

Equation (26.49) can be alternatively written as

$$\frac{\Delta Y}{\Delta X} = \frac{1}{1 - (b - g)} \quad \text{...(26.50)}$$

where b is *MPC* and g is *MPM* (marginal propensity to import).

Eq. (26.50) reveals that if $b = g$, then foreign trade multiplier will be equal to unity.

26.8.3 Foreign Trade Multiplier with Tax Function

We have discussed the determination of national income and foreign trade multiplier assuming a constant tax (T). In this section, are will discuss these aspects with a tax function. It may be noted at the outset that foreign trade multiplier takes a more complex form than the one given in Eq. (26.49) or (26.50) if tax (*T*) as a function of income is included in the model. With the inclusion of tax function, the equilibrium equation will be written as

$$Y = C + I + T + G + (X - M) \qquad \text{...(26.51)}$$

where $T = T + tYd$ and rest of the variables are the same as defined in Eq. (26.45).

By substituting the variable in functional form, Eq. (26.51) written as follows:

$$Y = a + b(Y - T - tY) + I + G + X - (M - gY)$$

$$= \frac{1}{1 - b + bt + g}(a - bT + I + G + X - M)$$

$$= \frac{1}{1 - b(1 - t) + g}(a - bT + I + G + X - M) \qquad \text{...(26.52)}$$

Let us now suppose that exports (*X*) increase by ΔX, all other factors remaining the same. With increase in *X* by ΔX, the equilibrium level of national income can be expressed as

$$Y + \Delta Y = \frac{1}{1 - b + bt + g}(a - bT + I + G + X - M) + \frac{1}{1 - b + bt + g}\Delta X \qquad \text{...(26.53)}$$

By subtracting Eq. (26.52) from Eq. (26.53), we get

$$\Delta Y = \frac{1}{1 - b + bt + g}\Delta X$$

Following the procedure given above, the foreign trade multiplier ($\Delta Y/\Delta X$) can be obtained by dividing both sides by ΔX, i.e.,

$$\frac{\Delta Y}{\Delta X} = \frac{1}{1 - b(1 - t) + g} \qquad \text{...(26.54)}$$

This is a foreign trade multiplier for a four-sector model in which income tax and imports are linear functions of income.

Numerical Example. Let us now illustrate the model with a numerical example. Suppose structural functions of an economy are given as follows:

$C = 100 + 0.75\ Yd$ (where $Yd = Y - T$)

$I = 100$

$G = 50$

$T = 25 + 0.2\ Y$

$X = 25$

$M = 10 + 0.1\ Y$

(Amounts in billion `)

Substituting these parameters in Eq. (26.52), we get equilibrium level of national income (*Y*) as follows:

$$Y = \frac{1}{1 - 0.75(1 - 0.2) + 0.1}(100 - 0.2 \times 25 + 100 + 50 + 25 - 10)$$

1

= (260)

0.20

= 1300

Given the structural functions, the equilibrium of national income is determined at `1300 billion, with open economy multiplier as 1/0.2 = 5.

SUMMARY

• Theory of national income determination was formulated by J. M. Keynes. According to Keynes, the equilibrium level of national demand is determined by aggregate demand for and aggregate supply (*AS*) of goods and services.

• Aggregate demand (*AD*) refers to the total expenditure made by the people of country including households, government, and foreigners. Aggregate supply (*AS*) refers the total supply of consumer goods and services and capital goods (machinery, equipments, etc.).

• Keynesian theory of income determination is explained under three models: (*i*) two-sector model including only two sectors – households and firms; (*ii*) three-sector model including households, firms and government; and (*iii*) four-sector model including households, firms, government and foreign sector.

• Consumption expenditure (*C*) is the function of income *Y*, i.e., $C = f(Y)$. So consumption expenditure increases only when income increases. The relationship between consumption and income is given by marginal propensity to consume (*MPC*) $= \Delta C/\Delta Y$.

• Investment (*I*) is treated as autonomous expenditure, i.e., it may increase or decrease, other remaining constant. Increase in investment (*I*) leads to increase in national income by a multiple of *I*, called *investment multiplier* (*Im*). Investment multiplier depends on the *marginal propensity to consume.* Investment multiplier is measured as $Im = 1/(1 - b)$, where $b = \Delta C/\Delta Y$.

• The measures of aggregate supply and measures of aggregate demand for three different models are different. In two-sector model, $AS = C + S$ where C = supply of consumer goods and S = savings and $AD = C + I$ where C = consumer expenditure and I = investment expenditure . In two-sector model, the equilibrium level of income is determined where $AD = AS$ or $C + I = C + S$.

• In three-sector model, known also as *open economy model, AD* is defined as $AD = C + I + G$ (where G = government expenditure) and $AS = C + S + T$ (where T = tax). In three-sector model, the equilibrium level of nation income is determined where $AD = AS$ or as $C + I + G = C + S + T$. If *G* increases, other variables remaining constant, national income increases at a multiple of *G,* called *G*-m *ultiplier. G*-m *ultiplier* = 1/ (1 – b).

• In four-sector model, $AD = C + I + G + X = C + S + T + M$, where X = exports and M = imports. In four-sector model, national income equilibrium is determined where $C + I + G + (X - M) = C + S + T$.

REVIEW QUESTIONS AND EXERCISES

1. Explain the concepts of aggregate demand and aggregate supply. Using aggregate demand and supply illustrate how equilibrium of national income is determined.

2. What is the meaning of the consumption function? Suppose a consumption function is given as $C = a + bY$. How can you derive a saving function from this consumption function?

3. Suppose consumption function of an economy is given as $C = a + bY$. Derive the saving function for the economy. Illustrate both the functions graphically.

4. Suppose consumption function is given as $C = 100 + 0.75\,Y$ and investment at `100 billion. State the aggregate demand function and present it graphically.

5. Suppose consumption function of a two-sector economy is given as $C = 200 + 0.8\,Y$ and $I = 100$. Find the equilibrium level of income, consumption and savings.

6. Show graphically that the equilibrium level of income and output once determined remains stable. Show also that if some extraneous factors disturb the equilibrium, the disequilibrium itself creates conditions for the system to return to the equilibrium.

7. What is multiplier? Explain how multiplier effect of an additional investment affects equilibrium income in a two-sector economy. Draw a diagram to show that $\Delta Y > \Delta I$ when $MPC > 0$.

8. Suppose a two-sector model is given as follows:

Consumption : $C = a + 6\ Y$

Investment

: $I = I$ (constant)

Find (a) Y at equilibrium (b) $\Delta\ Y$ if I increases to $I + \Delta\ I$.

9. Suppose a two-sector model is given as follows:

Consumption : $C = 100 + 0.8\ Y$

Investment

: $I = 50$

Find (a) Equilibrium level of income (b) Consumption at equilibrium income,

(c) $\Delta\ Y$ if $\Delta\ I = 20$, all other things being given.

10. Suppose

(i) $C = 50 + 0.75\ Y$

(ii) $I = 50$ (iii) $\Delta\ I = 10$

(a) derive saving function

(b) work out the multiplier and (c) find $\Delta\ Y$.

11. What is investment multiplier? How is investment multiplier worked out?

12. Suppose consumption function in an economy is given as $C = 150 + 0.75\ Y$. Find out investment multiplier by using income determination model.

13. Explain and distinguish between the concept of static and dynamic multiplier. Assuming a consumption function given as $C = a + bY$ and investment constant at I, show the working of static and dynamic multipliers.

14. What are the leakages from the economy that prevent the application of the multiplier theory to the less developed countries? Give your answer in the light of the conditions prevailing in the Indian economy.

15. How does inclusion of the government sector in two-sector model affect the income determination model? Illustrate your answer by using the three-sector model of income determination.

16. Suppose structural equations for an economy are given as follows.

$C = a + b\ Yd$

$Yd = Y – T$

$I = 50$

$G = 50$

$T = 50$

Find the equilibrium level of national income. What will be the level of national income if G increases from 50 to 75?

17. What is balance-budget multiplier? Assuming a 3-sector model, prove that the balanced budget multiplier equals one.

18. What additional variables are added to convert a 3-sector model into a 4-sector model? What are the factors that determine exports and imports of a country? How do exports and imports affect national income equilibrium?

19. What is export multiplier? Find export multiplier for the following model.

$C = a + b\ (Y – T)$

$I = I$

$G = G$

$T = T$

$X = X$

$M = M$

Compare export multiplier with investment multiplier.

20. Suppose macro variables are given as follows.

$C = 50 + 0.8\ Yd$, where $Yd = Y – T$

$I = 100$

$G = 50$

$T = 50$

Find (a) equilibrium level of national income;

(b) government expenditure multiplier;

(c) $\Delta\ Y$ resulting from $\Delta\ G = 25$, all other things given.

21. What is balanced budget multiplier? Suppose

$C = a + bYd$ and $Yd = Y – T$

$I = I$

$\Delta G = \Delta T$

Show that balanced budget multiplier equals 1.

22. Suppose structural equations (except T) of an economy are given as in question 7 and $T = T + tYd$. Find foreign trade multiplier.

23. Suppose structural equations for an open economy are given as follows.

$C = 50 + b(Y - 50 - tY)$

$I = 100$

$G = 50$

$T = 50$

$X = 10$

$M = 5 + 0.1\ Y$

$b = 0.8$

$t = 0.20$

Find (*a*) national income at equilibrium,

(*b*) foreign trade multiplier if $\Delta X = 5$ and

(*c*) M and T at equilibrium level of income.

FURTHER READING

Dornburg, Thomas, F., *Macroeconomics: Concepts, Theories and Policies* (New Yark: McGraw-Hill Book Co. 7th Edn. 1985), Ch. 4.

Dornbusch, r., Fisher, S. and Richard Startg, *Macroeconomics* (New Delhi) Tata McGraw-Hill Inc. 9th Edn. 2004), Ch. 9.

DwiveDi, D.N., *Macroeconomics: Theory and Policy* (New Delhi: Tata McGraw-Hill Book Co., 2010), chs. 6–8.

Frozen, richarD T., *Macroeconomics Theories and Policies* (Delhi: Pearson Education, 7th Edn., 2002), ch. 5.

shapiro, eDwarD, *Macroeconomics Analysis* (New York: Har-court. Brave Jouanocich Die, 4th Edn. 1994, ch. 4.

CHAPTER

27 Monetary Sector of the Economy

CHAPTER OBJECTIVES

The objective of this chapter is to analyse the monetary sector of the economy. The basic purpose of this chapter is to describe the working of the monetary sector and to explain how equilibrium of monetary sector is determined. The following aspects of monetary sector have been discussed in this chapter.

- Definition and kinds of money
- Functions of money
- Measurement of supply of money
- Classical and Keynesian theory of demand for money
- Keynesian theory of interest and monetary sector equilibrium
- Relevance of monetary theory to business decisions

27.1 INTRODUCTION

In the preceding chapter, we have discussed the working of the product sector and how equilibrium of the product sector is determined, assuming money sector to be in a stable condition and there is no change in demand for and supply of money. In reality, however, monetary sector does not remain stable – it keeps changing. Since money market and products markets are interrelated and interdependent, changes in money sector cause changes in product market. Briefly speaking, change in money supply and demand changes the rate of interest. And, change in interest rate changes the investment which causes change in aggregate demand. Change in aggregate demand disturbs the equilibrium of the product sector. The change in product sector changes demand for money and the equilibrium of the monetary sector. This process continues until both the sectors attain their equilibrium simultaneously. We have already discussed the working of the product sector. In this chapter we discuss the framework and functioning of the money market. The interrelated functioning of both the sectors has been discussed in the next chapter.

27.2 DEFINITION AND KINDS OF MONEY

27.2.1 Definition of Money

There is no unique definition of money. "Throughout history to the present day there is no agreement on the most fundamental of questions—what *is* money?"[1] The definition of money has been rather a controversial issue. ***Conceptually, money can be defined as any commodity that is generally accepted as a medium of exchange and measure of value.*** Historically, many commodities have been used to perform these functions of money and forms of money have been changing from cattle to credit cards. Therefore, an empirical question arises as to what should be and what should not be included in actual count of money. This remains an unsettled issue. A major factor that complicates the task of defining money is the increasing number of money substitutes in the form of assets that can be converted into spendable money. The other factor that has added to the controversy is the divergence between the conceptual and empirical definitions of money. As a result, the concept of money has changed from a measurable to immeasurable things. Let us now look at the different approaches to the definition of money.

H.G. Johnson[2] has classified, the approaches to the definition of money under the following four categories:

1. The Conventional approach,
2. The Chicago approach,
3. The Central Bank approach, and
4. The Gurley-Shaw approach.

These approach to definition of money are discussed below:

The Conventional Definition. The conventional approach to the definition of money refers to the oldest and the most widely accepted approach. The conventional definition of money emphasizes the basic functions of money, that is, the medium of exchange and measure of value. Going by these functions, money is defined as, 'Money is what money does' (Stanley Withers). Conceptually, any commodity that functions as a *medium of exchange* and *measure of value* is money. If one looks back into the history of money, one finds many kinds of commodities—cattle (cow, ox, horse, pig, goat), grains, stones, cowrie shells, cigarette, metals (copper, brass, silver and gold), dried fish, coffee, leather, etc.—have been used as a medium of exchange and a measure of value[3] at different stages of human civilization and in different parts of the world. The *commodity money* had, however, some problems by today's standards. It lacked (*i*) uniformity, (*ii*) homogeneity, (*iii*) standard size and weight (*iv*) durability and storability, (*v*) portability, (*vi*) stable value, and (*vii*) divisibility. Owing to these problems, money

1. Walters, A.A., "Introduction: Money and the Economy" in his *Money and Banking* (ed.), (Penguin, 1973), p.7.

2. Johnsom, Harry G., "Monetary Theory and Policy," *Am. Eco. Rev.*, Vol. 52, No. 3, June 1962, reprint in his *Essays in Monetary Economics,* (George Allen and Unwin Ltd. London, 1969).

3. For example, till the mid-20th century, most commodities were valued by East-African tribes in terms of goat. The prices of some commodities were fixed as follows:

1 hunting knife = 10 goats; 50 bananas = 1 goat;

5 bushels of corn = 2 goats; 1 young wife = 6 goats.

See also Geoffrey Crowther, *An Outline of Money, 1958,* p. 2.

has evolved over a long time into three major forms: (*a*) metallic coins, (*b*) paper currency and (*c*) demand deposits (operated through cheque system). These forms of money perform the basic functions of money and constitute, according to the conventional approach, the total supply of money.

The first two forms of money (metallic coins and paper currency) possess two distinctive features against the third form of money (demand deposits). The metallic coins and paper currency are created and issued by the government and are *legal tenders* in the sense that they enjoy a legal status. As legal tenders, coins and paper currency are not only accepted as a medium of exchange by all the citizens of a country but are also legally enforceable in the settlement of payment obligation.[4] Demand deposits, on the other hand, are the product of the banking system and making and accepting payments by cheques is *optional.*[5] One may accept or refuse to accept payments through cheques.

The Chicago Approach. The Chicago approach was pioneered by Milton Friedman of the Chicago University and his associates,[6] called jointly as Chicago school. The Chicago school has defined money so as to include also the *time deposits with the commercial banks.* Thus, the Chicago school has broadened the definition of money to include three components: (*i*) currency, (*ii*) chequeable demand deposits, and (*iii*) *time deposits*. Although time deposits are not readily available as medium of exchange, the Chicago school gives two reasons for including it in the concept of money supply. *First,* in their opinion, *GNP* and money supply are highly correlated and money supply including time deposits has a high correlation with *GNP* than the money supply without it. Therefore, time deposit must be included in the definition of money. *Second,* the Chicago school finds that *time deposits and demand deposits are, in practice, close substitutes* because banks make time deposits available to their customers after a laps of time, say 90 days, or so. So time deposits remain unavailable for transaction only for a short period. However, it is contended that neither of the arguments make a strong case for including time deposits in the concept of money.

The Gurley-Shaw Approach. The Gurley-Shaw approach7 is attributed to John G. Gurley and Edward S. Shaw. Recall that Chicago school recognises the medium-of-exchange function of time deposits it can be a substitute for the demand deposits and includes time deposits in the supply of money. Gurley and Shaw, on the other hand, recognise the asset function of also the 'financial claims against the non-banking financial intermediaries.' They emphasize 'close substitution relationship between currency, demand deposits, commercial bank time-deposits, saving-bank deposits, saving and loan association shares, and so on, all of which are viewed by the public as alternative liquid stores of value.' According to Gurley-Shaw approach, money supply should be defined as a weighted sum of currency, demand deposits and all the deposits and claims against the financial intermediaries that can be treated as the substitutes for currency and demand deposits-weights to be determined on the basis of the degree of their substitutability. Although Gurley-Shaw approach looks theoretically

4. For example, if a person, say Rohan, causes a damage to the limb, life or property of another person, say Rohini, then the compensation fixed by the court of law is enforceable on both Rohan and Rohini. Neither can offer or demand any other mode of payment not permissible by the law of the country. For example, Rohini cannot insist on 'an eye for an eye and a tooth for a tooth' kind of settlement as it is not legally permisible.

5. In India, however, the government has made it obligatory for all official payments to be made by cheque.

6. Including David Mieselman, Philip Cagan, Anna J. Schwartz and David Fand.

7. John G. Gurley and Edward S. Shaw, *Money in a Theory of Finance,* (Motilal Banarsidas, Delhi, 1968), Ch. 5.

sound, empirically it is immensely difficult to determine the degree of substitutability of deposits and claims against the financial intermediaries and, therefore, to assign appropriate weights to measure the money supply. "Except for illustrative purposes, no attempt has been made to make the weighted sum definition operational, that is, the concept has not been

used for testing monetary theory or for carrying out monetary policy"8.

The Central Bank Approach. The central banks take still a broader view of money supply. The reason is that the central banks are entrusted with the task of controlling and regulating the credit flows in accordance with the need of the economy. To accomplish this task, they need to formulate and implement a suitable monetary policy to achieve predetermined objectives. Therefore, central banks view all available means of payment and credit flows as money. For their purpose, money supply constitutes currency plus all 'realizable assets,' that is, the assets that can be converted into money at will. The central bank approach is accredited by the Radcliffe Committee. This Committee recognizes and emphasizes 'the similarity between currency and other realizable assets or means of purchasing to the point of rejecting money in favour of some broader concept, measurable or immeasurable.'9 According to this approach *money is, in a way, the total credit flow to the borrowers.* Depending on the objective of the monetary policy and policy targets, however, central banks make and use different measures of money supply, referred to as M 1, M 2, M 3, and M 4. The various measures of money supply used by the Reserve Bank of India will be described in a subsequent section.

27.2.2 Kinds of Money

Gone are the days of commodity money. Today, all the countries—developed, less developed and backward—use modern monetary system which works through *metallic coins* and *paper money* or *paper currency.* Another important kind of money is *bank deposit.* The latest addition to the monetary system is *credit card.* The credit card works as money essentially in the way as the bank deposits without the use of cheque system. In this section, however, we will discuss only the major kinds of modern money in circulation.

1. Metallic Coins. The second most important stage in the evolution of modern monetary system was the introduction of metallic coins made of iron, copper, silver and gold—and now alloy and aluminium. The invention and introduction of metallic coins must have been necessitated by the *defects* of commodity money— *heterogeneity or non-homogeneity of money units, non-durability, perishability, non-portability, unstable value, and indivisibility.* The exact year or period of introduction of metallic coins is not known. "The first coins are believed to have been made in ancient Lydia on the Aegian Sea during the seventh century BC."10. It is believed that metallic coins were in circulation in India about 2500 years ago. The metallic coins are believed to have been first minted and introduced by the private bankers and goldsmiths, the *sahukars,* who used to certify the weight of the coin and the purity of the metal (gold and silver) and put their seal. With the passage of time, the monetary system was taken over by the government or the government authorities with a view to making coins uniform and giving them a legal status. This gave the currency a general acceptability

8. Dwayne Wrightsman, *An Introduction to Monetary Theory and Policy,* (Free Press,. NY, 1971), p.20.

9. Johnson, H.G., *op. cit.,* p. 35.

10. Thomas D. Simpson, *Money, Banking, and Economic Analysis,* (Prentice-Hall 1987), p. 16.

and also a legal status. Except silver and gold coins, however, other metallic coins were and are only *token money*— a token money has no *intrinsic value.* For example, metallic coins in circulation in India are token money.

2. Paper Money. The paper money consists of the currency notes printed, authenticated and issued by the state and the central bank of the country. In some countries, there is a dual system. For example, in India, one-rupee currency notes and coins are issued by the Government of India and currency notes of higher denominations—rupees 2, 5, 10, 20, 50, 100, 500 and 1000—are issued by the Reserve Bank of India (RBI). The currency issued by the RBI is in the form of *promissory notes.* Each currency note issued by the RBI is in the form of *promissory notes.* Each currency note issued by the RBI bears a promise by the Governor of the RBI Governor—"I PROMISE TO PAY THE BEARER A SUM OF ....RUPEES". Here, the 'sum' means one-rupee currency notes or coins issued by the Government of India and the RBI currency notes of other denominations.

Like metallic coins, the exact time of introduction of the paper currency is not precisely known. The factors that might have contributed to the advent of paper money are: (*i*) supply of gold and silver lagging far behind the demand for money due to rapid increase in the supply of goods and services, (*ii*) lack of portability of large sums of metallic money, (*iii*) loss of weight and value due to depreciation and debasement of coins by the people, and (*iv*) increasing other uses of metals like gold and silver.

Bank Deposits. The third form of common money is bank deposit. Bank deposits include three kinds of deposits: current account deposits, saving bank deposits and time deposits. The current account deposits are available on demand. That is why current account deposits are widely referred to as *demand deposits* which can be transferred and used as medium of exchange by the instrument of cheque. Demand deposits are also known as 'bank money' and 'credit money.' According to the conventional approach to the definition of money, *only demand deposit is treated as money, because it is nearly as liquid as cash in hand.* However, the Chicago approach (as described above) treats *saving* and *time deposits* as a close substitute for cash and demand deposits. Therefore, according to the Chicago approach, saving and time deposits are also included in money and money definition.

Credit Cards. The most modern form of money is credit card and debit card issued by the banks. Credit and debit cards serve as a medium of exchange. The credit card holders do not need to carry cash for shopping. Given the facility of authorization and inter-bank co-ordination, one can transact the business, all over the country and in case of credit cards issued by international banks, all over the world.

27.3 FUNCTIONS OF MONEY

Money was devised initially as a medium of exchange and measure of value. However, it acquired over time some other functions also. The following couplet brings out the major functions of money.

Money is a matter of functions four:

A medium, a measure, a standard, a store.

As this couplet reveals, money performs four major functions which are discussed below in some detail along with their importance in modern economy.

(i) Money as a Medium of Exchange. Money functions as a medium of exchange between any two goods. This is the most important and unique function of money. The importance of this function lies in that it has solved one of the biggest problems of the barter system. In the barter system, for exchange to take place, there must be 'double coincidence of wants.' The double coincidence of wants exists when, between any two persons, one is willing to accept what the other person is willing to give in exchange. Until this condition is fulfilled, exchange cannot take place. For example, a weaver cannot exchange his cloth for shoes unless the shoemaker wants cloth. In a modern, market economy, the problem of 'double coincidence of wants' is solved by money. Since money is acceptable to all, the weaver can sell his cloth to anybody (say, to a farmer or to a carter) for money and buy the shoes in exchange for money. This system works efficiently because money can buy anything, it has purchasing power and is acceptable to all.

The uniqueness of the medium-of-exchange function of modern money compared to other commodities lies in the following unique merits of money: (*a*) general acceptability, (*b*) easy portability, (*c*) divisibility, (*d*) difficult to counterfeit[11], (*e*) value guaranteed by the government, and (*f*) legal enforceability as mode of compensation.

(ii) Money as a Measure of Value. The second function of money is to work as a measure of value. All values are measured in terms of money. As a measure of value, money works as a *common denominator, as* a *unit of account* and as a *numeraire*. Today, the value of all the goods and services is expressed in terms of money. Money being a common denominator, the values of different goods can be added to find one value of all possessions of a person, of a firm and of a nation. In fact, money makes computation of national income possible. In the absence of money, measuring value would be an extremely difficult proposition in a modern economy.

In modern times, a society produces, buys and sells, and consumes goods and services in such a large number, variety and quantity that measuring and expressing values in terms of commodities would be a rather impossible task. If possible, it would mess up the entire exchange economy. Money has made the task easier. Not only each good and service has a price, but also one can find and compare the relative prices in terms of money.

(iii) Money as a Store of Value. The third basic function of money is to work as a *store of value*. The need to store value must have arisen for such reasons as (*i*) need for storing surplus produce because production and consumption or exchange of goods and services are not instantaneous in most cases, (*ii*) need for storing value for future use due to uncertainties of life, and (*iii*) accumulative nature of the people. The advent of money has provided a means to store value for future use. Even the most perishable goods can be converted into money. If prices do not increase over a long period, value can be stored for long without the loss of value. However, in case of rising prices, money stored loses its value in proportion to the general price rise.

(iv) Money as a Standard of Deferred Payments. Borrow today and repay tomorrow or buy tomorrow and pay later has been an old practice. One necessary condition of deferred payment has been that the value returned must be the same. During the barter days, it might be a difficult problem to judge whether the value returned after a lapse of time was the same.

11. Pakistanis are, however, counterfeiting Indian currencies of high values and make it flow to the Indian market via Nepal and Bangladesh.

For example, whether a quintal of wheat borrowed this year and returned one year later had the same value was a difficult question to which answer lay in customs and practices of those days. However, with the expansion of economic activities, the area and volume of borrowing and lending and sale and purchase on credit expanded enormously. Personal borrowing and lending expanded to professional activity by a class of people including moneylenders, the *sahukars,* to modern banking system and growth of credit market, involving payment of interest and principal at a later date. The deferred payment system expanded to purchase of raw materials, payment of wages, salaries and pensions, payment by wholesalers to producers and by retailers to wholesalers and consumers to retailers. In the absence of money, the economic system would have either not grown to today's level or would been extremely chaotic. The advent of money has solved the problem of deferred payment by its merits: (*i*) it is generally acceptable (*ii*) it is legally enforceable, and (*iii*) it has a relatively more stable value than other commodities.

27.4 SUPPLY OF MONEY

The supply of and demand for money are the two most important variables of the monetary sector. Money supply and demand determine two important macroeconomic variables: (i) the rate of interest, i.e., the return on money, and (ii) the general price level. Interest rate and general price level play a very significant role in determining many other macroeconomic variables, especially the level of *GDP* and employment. In this section, we discuss the composition of money supply and the factors that determine money supply. Demand for money will be discussed in the next section.

27.4.1 Measures of Money Supply

Theoretically, money supply is defined and measured under two concepts of money supply: (*i*) *ordinary money*, i.e., the stock of money or the money in circulation, and (*ii*) *high power money*, i.e., the base money . The *ordinary money (M)* is defined and measured as

$$M = C + DD \quad \ldots(27.1)$$

(where C = currency with public, and DD = demand deposits with banks) .

The *high power money* (H) is defined and measured as

$$H = C + R \quad \ldots(27.2)$$

(where C = currency with public, and R = cash reserves with central bank and other banks).

According to the theory of money supply, given the *high power money* (H), the total supply of money (M) in a country is determined by a *money multiplier* (m). It can be expressed as

$$M = m\,H \quad \ldots(27.3)$$

Given the total money supply (M), $m = M/H$.

The *money multiplier* (m) can be measured with the application of the two concepts of money supply, as given in Eqs. (27.1) and (27.2), shown below.

$$m = \frac{C + DD}{C + R} \quad \ldots(27.4)$$

27.4.2 Measuring Money Supply in India

The Reserve Bank of India (RBI) makes the measure of money supply in India. The RBI measures the money supply (M) as follows.

$$M = M1 + M2 + M3$$

where

$M1 = C + DD + OD$ (OD = other deposits)

$M2 = M1$ + Saving Deposits with Post Offices

$M3 = M1$ + Net Time Deposits with Banks.

27.5 DEMAND FOR MONEY

The economists of different generations have offered different theories of demand for money. The theories of demand for money can be classified as (*i*) classical theory of money demand, (*ii*) Keynesian theory of demand for money, and (*iii*) money-demand theories of post-Keynesian era. However, the Keynesian theory of demand for money is still treated as the fundamental theory of money demand. Although, economists of post-Keynesian era have offered other theories of money demand, these theories are basically the different versions of improvement in the Keynesian theory. Given our objective and scope of the book, we confine here to the classical and Keynesian theories of demand for money.

27.5.1 Classical Approach

C**lassical economists** did not offer any specific theory of demand for money. But, classical theory of demand for money has been derived from the ***classical quantity theories of money.*** For instance, according to the Fisher's quantity theory of money,

$$MV = PT \quad \ldots(27.4)$$

(where M = money with public; V = velocity of money; P = price of goods; and T = transaction, the rate of exchange of goods).

Given Eq. (27.4), MV gives the supply of money and PT (= $P \times T$) gives the measure of national income (Y) as it gives the money value of all economic transactions. So Eq. (27.4) can be written as

$$MV = Y$$

And

$$M = 1/V\,(Y) \quad \ldots(27.5)$$

According to classical postulate, an economy is always in equilibrium and money supply (Ms) is always equal to money demand (Md), i.e., $Ms = Md$. In Eq. (27.5), M denotes money supply (Ms). By designating $1/V$ as k, money sector equilibrium can be expressed as

$$Ms = kY \quad \ldots(27.6)$$

Since at equilibrium,

$Ms = Md$, Eq. (27.6) can be expressed as

$$Md = kY \quad \ldots(27.7)$$

Eq. (27.7) gives the classical theory of money based on *quantity theory of money.*

27.5.2 Keynesian Theory of Demand for Money

Keynes developed analytically a very sound theory of demand for money what he called ***the liquidity preference theory.*** Keynesian theory explains why people prefer liquidity

or holding cash. However, other economists interpret it as theory of demand for money.

According to Keynes, money is demanded by the people for three *motives*:

(*i*) Transaction motive,

(*ii*) Precautionary motive, and

(*iii*) Speculative motive.

The demand for money for these motives is discussed here briefly.

(i) Transaction demand for money. Transaction demand for money refers to money held by the people to carry out necessary transactions to meet the needs of life. People hold money for transaction purpose because there is often a time gap between need for expenditure and income earning. Income is earned periodically – weekly, monthly or annually. But need for spending money to buy goods and services arises regularly. Therefore, people hold money for transaction purpose. Money held for transaction purpose is termed as *transaction demand for money.*

According to Keynes, demand for money depends on two factors: (*i*) income level, and (*ii*) rate of interest12. As regards the ***relationship between income and money demand for transaction***, demand for money is positively related to income – the higher the income, the higher the transaction demand for money at constant prices. The relationship between transaction demand for money (Mt) and income (Y) is expressed in function form as

$$Mt = f\,(Y) \quad \ldots(27.1)$$

According to Keynes, there is constant relationship between income (Y) and money demand for transaction motive. The empirical transaction money demand function is then expressed as

$$Mt = kY \quad \ldots(27.2)$$

(where $k = \Delta Mt / \Delta Y$, a proportion of income demanded for transaction purpose).

The relationship between income and transaction demand for money is shown graphically in Fig. 27.1.

ΔMt

M

Δ

= k

Y

t

M = kY

t

ΔMt

Money Demand

ΔY

Interest Rate (i)

O

O

Real Income (Y)

Transaction Demand for Money

Fig. 27.1 *Income and Transaction Demand* **Fig. 27.2** *Interest and Transaction*

Demand for Money

Demand for Money

12. In the context of demand for money, interest rate is considered as the *opportunity cost* of holding idle cash balance.

As regards the relationship between ***interest rate and transaction demand for money***, according to Keynes, the transaction demand for money is *interest-inelastic*. The reason is that whatever the rate of interest, people have to pay for food, house rent, electricity bill, medical bill, kids' school fee, etc. However, if interest rate shoots up very high, people tend to economize on routine expenditure. The relationship between interest rate and transaction demand for money is shown graphically in Fig. 27.2.

(ii) Precautionary demand for money. *Precautionary demand* for money refers to the money held by the households and firms to meet contingent expenditure. The need for contingent expenditure arises due to unforeseen and unpredictable events like fire, theft, sickness, accidents, loss of job, etc. Since there are many uncertainties in human life, people hold some idle cash balance as safety measure. Keynes called money held with the purpose of meeting the contingent expenditure as the *precautionary demand for money*. According to him, precautionary demand for money also depends on the income level of the people. In other words, like *transaction demand*, the precautionary demand for money (Mp) is also a constant function of income.

$$Mp = kY \quad \ldots(27.3)$$

(iii) Speculative demand for money. According to Keynes, people hold a part of their income in the form of idle cash balance for *speculative purpose* with the objective of taking advantage of changing market conditions, especially in *money market*. The money held in the form of idle cash balance for speculative purpose is the *speculative demand for money*. Money held for speculative purpose has an opportunity cost, i.e., the loss of interest on cash held as idle cash balance. In general, speculative demand for money depends on the market rate interest. If market rate of interest is very high, then the opportunity cost is very high and hence people hold smaller idle cash balance. When interest rate declines, the opportunity cost goes down and people hold a higher idle cash balance for speculative purpose. According to Keynes, therefore, speculative demand for money (Msp) is the function of interest (i), expressed as

$$Msp = f(i) \quad \ldots(27.4)$$

The Liquidity Trap: Keynes has also pointed out a *remote possibility* of people holding a very large or unlimited amount of idle cash balance when market rate interest is critically very low. Keynes called this kind of idle cash holding as *liquidity trap*. It implies that at a critically low interest rate people fall in liquidity trap. The relationship between market rate of interest and speculative demand for money is illustrated graphically in Fig. 27.3.

Fig. 27.3 *The Speculative Demand for Money*

27.5.3 Aggregate Demand for Money: Keynesian Approach

Having described the various components of demand for money, their determinants and money demand functions, we proceed now to present the Keynesian approach to

measure the aggregate demand for money. As noted above, according to Keynes, money is demanded for three motives: (*i*) transaction motive, (*ii*) precuationary motive, and (*iii*) speculative motive. Since *transaction* and *precautionary demands* are both a function of income, these both kinds of money demand are treated as *transaction demand for money.* Thus, the aggregate demand for money can be expressed as

$Md = Mt + Msp$

...(27.5)

Since $Mt = f(Y)$ and $Msp = f(i)$, the Keynesian aggregate money demand function can be expressed as

$Md = f(Y) + f(i)$

...(27.6)

The relationship between the aggregate demand for money and the interest rate is crucial to the Keynesian theory of demand for money and the theory of interest. The relationship between the two is shown by a total-money-demand curve *(Md)* in relation to the interest rate. The derivation of *Md*-curve has been illustrated in Fig. 27.4. Panel (a) of the figure shows the transaction demand for money (*Mt*) in relation to the interest rate. Since transaction demand for money is assumed to be interest-inelastic, *Mt* is shown by a straight vertical line. As shown in the figure, whether the interest rate is i 1, i 2 or i 3 , the transaction demand for money remains constant at *Mt*.

Fig. 27.4 *The Total Demand for Money*

Panel (b) presents the speculative demand for money *(Msp)* in relation to the interest rate. The *Msp* is inversely related with the interest rate.

Panel (c) presents the total demand for money (*Md*). The total-money-demand curve, i.e., *Md* curve, is simply a horizontal summation of *Mt* and *Msp* curves. The *Mt* and *Msp* curves of panels (a) and (b) are reproduced in panel (c) and shown by the dotted lines. The *Md* curve gives the total demand for money *(Md)* in relation to the interest rate. The curve *Md* is the Keynesian demand curve for money.

27.5.4 Criticism of the Keynesian Theory of Demand for Money

The Keynesian theory of demand for money was undoubtedly a radical improvement over the classical and neoclassical theories of money. His theory has however been criticised on the following grounds.

First, Keynes' division of demand for money between transaction, precautionary and speculative motives is unrealistic. For, the people do not maintain a separate purse for each motive. They have one purse for all purposes. Besides, empirical evidence shows that, contrary to Keynes' postulate, even the transaction demand for money is interest-elastic.

Secondly, critics reject the Keynesian postulate that there exists a 'normal' rate of interest and the current rate of interest may not necessarily be the same as the normal rate: there may always be difference between the two. According to Keynes, the speculative demand for money is governed by the difference between the 'normal' and the current rates of interest. But, the critics argue that if the current rate of interest remains stable over a long period of time, people tend to take it to be the normal rate. Consequently, the difference between the current rate and the normal rate disappears. With it, disappears the basis for speculation and the speculative demand for money.

Thirdly, Keynes assumed unrealistically that the people hold their financial assets in the form of either idle cash balance or bonds. In fact, people hold their assets in a combination of both the assets.

27.6 KEYNESIAN THEORY OF INTEREST AND MONEY MARKET EQUILIBRIUM

Having discussed the Keynesian theory of demand for money, we discuss now the Keynesian theory of interest rate determination and the money market equilibrium. According to the Keynesian theory of interest, market rate of interest is determined by the aggregate demand for money and the total supply of money. And, the equilibrium rate of interest is determined at the rate at which money market is in equilibrium. Money market is in equilibrium where

$Md = Ms$

The determination of the equilibrium rate of interest is illustrated in Fig. 27.5 The derivation of the demand curve for money

(*Md*) has already been explained in the foregoing section. It is reproduced Fig. 27.5 as shown by the *Md*-curve.

As regards the supply of money, in the Keynesian model, the supply of money (*Ms*) is assumed to remain constant in the short run. It is constant because the supply of money in any country is determined by the central bank of the country in view the overall monetary

Fig. 27.5 *Determination of the Interest*

needs of the country. The supply of money

Rate: The Keynesian Theory

in India, for example, is determined by the Reserve Bank of India. Central banks do not increase or decrease the supply of money in response to the variation in the rate of interest. Therefore, the supply of money in any time period is assumed to be given, as shown by the vertical line *Ms* in Fig. 27.5. It is therefore deemed to be *interest-inelastic*.

As Fig. 27.5 shows, the money demand curve (*Md*) and money supply schedule (*Ms*) intersect at point *E*. At this point, *Ms* = *Md*. Therefore, the equilibrium rate of interest is determined at i_1. This rate of interest is supposed to be stable. For, at any other rate of interest, *Md* ≠ *Ms*. For example, if the interest rate rises to i_2 for some reason in any period of time, *Md* will decrease by *AB*. The reason is, with the increase in the interest rate, bond prices go down and the speculative demand for money decreases. This situation of disequilibrium sets the market forces in motion to restore the equilibrium. *How*? When the interest rate goes up, people prefer to hold bond because of its low price and reduce their idle cash balances. The bearish factor in the market speculates a fall in the interest rate and the consequent rise in the bond price—an expectation that will fetch capital gain. This creates demand for an idle cash balance. This implies movement from point *A* towards point *E* on the *Md*- curve forcing down the interest rate. As a result, the excess supply of money disappears and equilibrium is restored.

Similarly, when the rate or interest falls, for some reason, from i_1 to i_0, the speculative demand for money increases because at a lower rate of interest the preference for cash holding increases. As a result. the aggregate demand for money *(Md)* increases by *CD.* Consequently; demand for money exceeds supply of money by *CD.* Since there is shortage of money in the money market, people begin to expect a rise in the interest rate and, therefore, demand for money begins to decrease and continues to decrease until the equilibrium point *E* is restored.

27.7 CHANGE IN MONEY DEMAND AND SUPPLY AND INTEREST RATE

We have discussed above the determination of interest rate in the money market given the money demand function and supply of money. In reality, however, both money demand and supply are subject to change – increase and decrease – over a period of time. The change in money demand and supply causes change in the interest rate. For instance, given the supply of money, increase in demand for money causes increase in the interest rate and *vice versa.* Similarly, given the money demand, increase in money supply leads to decrease in the interest rate and *vice versa.* The effect of change in money demand and supply is illustrated graphically in Fig. 27.6.

Fig. 27.6 *Change in Money Demand and Supply and Interest Rate*

Recall that money supply is determined by the central bank of the country and it is interest-inelastic. Given this condition, suppose money supply is given by the straight vertical line Ms_1 and money demand curve is given by the curve marked Md_1 . As shown in Fig. 27.6, money demand curve (Md_1) and money supply line (Ms_1) intersect at point E_1 determining the interest rate at i_1 . At this interest rate, demand for money is equal to supply of money, i.e., *Md* = *Ms*. Therefore, money market is in equilibrium.

Let us now see how changes in money demand and money supply change the interest rate. Suppose money market is in equilibrium and demand for money increases so such that money demand curve, Md_1, shifts rightward to the position of Md_2. As a result, the interest rate increases from i_1 to i_3. Given the equilibrium rate of interest at i_3 , suppose money supply increases so much that money-supply line *Ms* 1 shifts to the position of *Ms* 2 intersecting *Md* 2 at point *E* 3. Thus, the increase in money supply brings down the interest to *i* 2. This indicates that changes in money demand and supply change in interest rate.

27.8 CRITICISM OF THE KEYNESIAN THEORY OF INTEREST

The Keynesian theory of interest is undoubtedly superior to the classical and loanable funds theory of interest. Ironically, however, the Keynesian theory of interest has been criticised on the grounds Keynes criticised the classical theory. Keynes' argument against the classical theory of interest can be summarised as follows. Since $S = f(Y)$, saving schedule cannot be known unless income (*Y*) schedule is known. Since $Y = f(I)$, income schedule cannot be known unless investment function is known. Since $I = f(i)$, investment schedule cannot be known unless interest rate (*i*) is known. And, interest rate (*i*) cannot be known unless saving and investment schedules are known. Thus, according to Keynes, the indeterminateness of the variables make the classical theory of interest indeterminate. According to Hansen, 'exactly the same criticism applies to Keynesian theory in its simpler form.13 He reiterates, 'Keynes' criticism of the classical theory applies equally to his own theory.14 His argument may be summarised as follows. 'According to the Keynesian theory the rate of interest is determined by the intersection of the supply schedule of money ... and the demand schedule for money ... .' This theory 'also is indeterminate' because, even if money supply is fixed by the monetary authority, 'the liquidity preference schedule will shift up or down with changes in the income level.' In the Keynesian system, we cannot know the liquidity preference schedule unless we know the income level. Income level cannot be known unless we know the speculative demand for money and speculative demand for money cannot be known unless interest rate is known. Thus, 'the Keynesian theory, like the classical, is indeterminate."15

Leijonhufvud remarked that the Keynesian theory of interest is 'incredibly tortuous formulation.16 According to him, the main trouble lies in his definition of 'savings' as

13. Alvin H. Hansen, *Guide to Keynes,* (McGraw-Hili, NY, 1953), p.140.

14. *ibid.* p.141.

15. Hansen, A.H., *ibid.*

16. Alex Leijonhufvud, on *the Keynesian Economics and the Economics of Keynes,* (London, Oxford University Press), 1968, p.28.

'non-consumption' taken from the 'pure' theories of interest. This definition might be appropriate in the pure theories of interest, but not in the Keynesian system. Keynes' 'ex-ante savings' is not clearly distinguished from the demand for money for speculative purpose and demand for non-monetary assets.

SUMMARY

- Monetary sector consists of all the organizations, institutions and agents that deal in monetary transactions – borrowing and lending. It includes central banks, commercial banks, financial institutions and monetary dealers.
- Money is the prime element of transaction in the monetary sector. Money has been defined by the economists of different generations in different ways. In general, money is defined as any commodity and monetary documents that is generally acceptable as a medium of exchange and store of value. In general, in modern times, money in circulation consists of metallic coins, paper currency, bank deposits and credit cards.
- Money functions as (*i*) medium of exchange, (*ii*) measure of value, (*iii*) store of value, and (iv) a standard means of deferred payment.
- Money supply is measured as under two concepts of money supply: (*i*) ordinary money, and (*ii*) high power money. Under the concept of ordinary money, money is measured as $M = C + DD$, where C = currency with public, and DD = demand deposits with banks. Under the concept of high power money (*H*), money supply is measured as $H = C + R$, where C = currency with public, and R = *c* ash reserves with central bank.
- Money demand is analysed primarily under classical and Keynesian theories. Under classical theory, money is demanded for transaction purpose and money demand (M d) = kY where Y = income and k = proportion of income used for transaction.
- According to the Keynesian theory of money demand, money is demanded for

three purposes: (*i*) transaction purpose (*M* t), (*ii*) precaution purpose (*Mp)*, and (*iii*)
speculative purpose (*M*). Since
are both function of income (
sp
Mt and *Mp*
Y), these
demand are joined as *M*
is the function of interest rate. Thus, total demand
T, and *Msp*
for money *M* =
d
MT + *Msp*.

• Keynesian theory of demand was developed with the purpose of formulation of the theory of interest. According to the Keynesian theory of interest, money supply (*Ms)* is interest-inelastic and interest rate is determined at the level at which *M*
.
d = *Ms*

REVIEW QUESTIONS

1. What are the reasons for holding money in the Keynesian system? How is the Keynesian theory of demand for money different from the classical theory?
2. Define and explain the speculative demand for money. Why is the speculative demand for money interest-elastic?
3. Why does the speculative demand for money change with the change in the interest rate? Explain in this regard the relationship between the interest rate and the bond prices.
4. When the transaction demand for money is interest-inelastic and speculative demand for money is interest-elastic, how can the total money demand curve be interest-elastic?
5. Why does an increase in the interest cause a decline in the bond prices? What is its effect on the demand for money?

6. What is the effect of simultaneous change in money demand and money supply if (*a*) both change in the same direction at equal and different rates, and (*b*) both change in opposite direction at equal and different rates?
7. What is meant by the 'Liquidity trap'. Why does a change in money supply up to a certain limit does not help the economy out of the liquidity trap?
8. Explain the Keynesian theory of the interest rate determination. How is the Keynesian theory different from the classical theory of interest?
9. Explain the derivation of the Keynesian money demand function. What factor causes the shift in money demand function?
10. Suppose money demand function is given as $MD = kY + f(i)$. Explain and illustrate graphically how change in Y and i would affect MD.
11. The Keynesian theory of interest is as indeterminate as the classical theory. Comment.
12. "Keynes's criticism of the classical interest theory applies equally to his own theory." Explain and justify this statement.

FURTHER READING

Ackley, Gardner, *Macroeconomics: Theory and Policy,* (Macmillan, NY), 1978, Chs. 5 and 9.
DwiveDi, D.N., *Macroeconomics: Theory and Policy* (New Delhi: Tata McGraw-Hill, 2010), Chs, 5-8.
lAiDler, D.W., *The Demand for Money: Theories and Evidence,* (Allied Publishers, Bombay, 1972).
ShApiro, E. *Macroeconomic Analysis,* (Galgotia, New Delhi, 1994), Ch, 11.

CHAPTER

28 IS-LM Model of the Economy:
The General Equilibrium Model

CHAPTER OBJECTIVES

The objective of this chapter is to present the analysis of the determination of the general equilibrium of the economy. An economy attains its general equilibrium when both product and money sectors reach a point of equilibrium simultaneously at the same level of national income. This chapter deals with the following aspects of the general equilibrium model of the economy.

• Interdependence of the product and money markets

• Formation of the *IS- LM* model

• Formation of *IS- function* and derivation of the *IS* curve

• Formation *LM- function* and derivation of the *LM* curve

• Determination of the general equilibrium i.e., how an economy attains its general equilibrium.

28.1 INTRODUCTION

In Chapter 26, we have discussed Keynesian theroy of national income determination, i.e., his theories related to the product sector. And, in Chapter 27, we have discussed his monetary theory and the determination of the monetary sector equilibrium. It may be recalled from these chapters that Keynes formulated his theory of national income determination in isolation of the monetary sector and developed his monetary theory in isolation of the product sector. In other words, Keynes formulated his theory of national income determination assuming that monetary sector variables are constant and have no effect on the product sector. Similarly, he developed his monetary theory assuming that product sector variables are constant and changes, if any, have little to do with the monetary sector. In reality, however, working of the product and monetary sectors is interrelated, interdependent and interactive. As such, a major change in product market affects working of the money market and a major change in money market affects the working of the product market. It was J. R. Hicks who highlighted these facts in 1937 – one year after the publication of Keynes's *The General Theory* in 1936 – and claimed that unless both product and monetary sectors reach equilibrium simultaneously, the economy will not reach a stable general equilibrium.

In order to explain how two sectors interact and can reach equilibrium simultaneously at a level of national income, Hicks developed an analytical model, known as *IS-LM* model1. In

1. The term *IS* marks the abbreviation of $I = S$. Here, both I (investment) and S (savings), are related to the product sector. And, the term *LM* is the abbreviation of $L= M$. In this case, L indicates 'liquidity', i.e., demand for money, and M indicates 'money supply', both being related to the monetary sector.

this model, the term *IS* represents the equilibrium condition ($I = S$) of the product sector and the term *LM* represents the monetary sector equilibrium ($L = M$). The purpose of the *IS-LM* model is to analyze theoretically how product and monetary sectors can reach a point of equilibrium simultaneously.

In this chapter, we present a brief description of the *IS-LM* model as developed by J. R. Hicks. Although Hicks had formulated his general equilibrium theory through *IS-LM* model in a simple economy (two-sector) framework, economists of the later generations extended his *IS-LM* model to three-sector and four-sector models. Three-sector model takes into account three sectors of the economy including (*i*) households, (*ii*) firms, and (*iii*) the government. It is also referred to a closed economy model. And, four-sector model takes into account also the foreign sector. Following the general practice, we present the *IS-LM* model of general equilibrium in three models.

1. Two-sector model – the 'simple economy model',
2. Three-sector model – the 'closed economy model', and
3. Four-sector model – the 'open economy model'.

We begin our discussion on the *IS-LM* model by having a view of basics of the formation of the *IS-LM* model, i.e., by looking at interdependence of the product and the monetary sectors.

28.2 INTERDEPENDENCE OF THE PRODUCT AND MONEY MARKETS

As noted above, the working of both product sector and monetary sector is interrelated and interdependent. In simple words, a change in any variable of one sector affects the interrelated variable of the other sector. The two important variables that create inter-linkages between the two sectors are (*i*) *investment* (I), a product-sector variable, and (*ii*) *interest rate* (i) , a monetary-sector variable. Let us now see how product sector is affected by change in the interest rate (i), the monetary sector variable, and how monetary sector is affected by the change in investment (I), the product-sector variable.

28.2.1 Dependence of Product Sector on Monetary Sector

Let us first look at the ***dependence of product sector on monetary sector***. Recall the condition of the product sector equilibrium, i.e., $AD = AS$, where AD = aggregate demand, and AS = aggregate supply which represents the level of national income (Y) at equilibrium. As such, the product market is in equilibrium where national income, $Y = C + I$. Although Keynes assumed investment to be an autonomous factor, in reality, the variable I depends on the interest rate, a monetary sector variable. Investment function is expressed as $I = f(i)$. Investment function when estimated empirically, it is expressed as

$I = \hat{I} - hi$, ($h > 0$)

…(28.1)

The investment function implies that a change in the interest rate (i) causes a change in investment (I). Suppose interest rate decreases for some extraneous reason. Given the investment function (28.1), decrease in interest rate (i) leads to increase in investment from I to $I + \Delta I$.

Since $Y = f(I)$, the increase in investment by ΔI causes an increases in the equilibrium level of national income from $Y = C + I$ to $Y + \Delta Y = C + I + \Delta I$. It implies that ΔI leads to ΔY.

Since $C = f(Y)$, change in income changes C also . That is, ΔY causes ΔC. This analysis shows that a change in a single monetary factor, interest rate (i), causes changes in the equilibrium level of national income. It shows the dependence of product sector to on the monetary sector.

28.2.2 Dependence of Monetary Sector on Product Sector

Now let us see the ***dependence of monetary sector on the product sector.*** Recall that according to Keynesian monetary theory, monetary sector is in equilibrium when money demand equals money supply, i.e., money sector is in equilibrium when $Md = Ms$. Recall also that $Md = Mt + Msp$ and $Mt = f(Y)$ and $Msp = f(i)$. Note that product sector variable (Y) is an important factor in determining money market equilibrium. Since $Mt = f(Y)$, money market equilibrium cannot be attained until equilibrium level of Y is determined. And, equilibrium level of Y cannot be attained until investment (I) is determined and equilibrium level of I cannot determined unless interest rate is determined, and so on. The foregoing brief analysis shows the interdependence of the working of the product and monetary sectors. This analysis of interdependence of the two sectors lays the foundation for the formation of *IS-LM* model.

28.3 FORMULATION OF THE IS-LM MODEL

Hicks has formulated is *IS-LM* model by integrating the money market variable interest rate (i) into product market equilibrium and the product market variable (Y) into the money market equilibrium. To begin with the product sector, recall the product market equilibrium condition in two-sector model given as

$Y = C + I$

Hicks incorporates a money market variable interest rate (i) into Keynesian model of income determination in a functional form and expresses product market equilibrium as

$Y = C(Y) + I(i)$

…(28.2)

Eq. 28.2 gives the condition for the ***product sector equilibrium*** and it is treated as *IS* function.

Similarly, Hicks incorporates a product-sector variable (Y) into the equilibrium condition of the monetary sector. By Keynesian theory, money sector equilibrium condition is given as $Md = Ms$ (where $Md = Mt + Msp$).

Given the components of Md, the monetary sector equilibrium is given as

$Mt + Msp = Ms$

…(28.3)

In Eq. 28.3, $Mt = f(Y)$ and $Msp = f(i)$. Suppose $Mt = f(Y)$ is given as $Mt = kY$ and speculative money-demand function $Msp = f(i)$ is given as $L(i)$. Given Mt and Msp functions, the ***monetary sector equilibrium*** can be expressed as

$M\text{s} = kY + L(i)$

…(28.4)

Eq. (28.4) represents the *LM* function.

Given the product-sector equilibrium in Eq. (28.2) and monetary-sector equilibrium in Eq. (28.4), the general equilibrium model, is expressed as given below.

$C(Y) + I(i) = kY + L(i)$

…(28.5)

Recall that in Eq. (28.2) represents the *IS* function and Eq. (28.4) represents the *LM*

function. The *IS-LM* model can be formulated by combining the *IS* and *LM* functions.

Thus, the *IS- LM* model can be expressed as

$IS = LM$

By substituting *IS* function and *LM* function for *IS* and *LM*, respectively, we get

IS-LM model as

$C(Y) + I(i) = kY + L(i)$

…(28.6)

Having presented the basics of the *IS-LM* model, now we proceed to present the model graphically. We begin the analysis of *IS-LM* model with two-sector model.

28.4 TWO-SECTOR IS-LM MODEL: THE SIMPLE ECONOMY MODEL

As noted above, Hicks had formulated his *IS-LM* model in the framework of the two-sector economy consisting of only two sector, viz., households and firms. It is assumed that there is no government and if the government does exist, it does not perform any economic activity.

As explained above, in two-sector model, the product-market equilibrium condition is given as

$Y = C(Y) + I(i)$

…(28.7)

Eq. (28.7) provides the *IS* function which is the basis for the *derivation of the IS curve.*

Similarly, the equilibrium condition for the monetary sector is given as

$Ms = Md = kY + L(i)$

…(28.8)

Eq. (28.8) provides the *LM* function which in the basis for the derivation of the *LM* curve.

Given the *IS* and *LM* functions which provide the basis of deriving *IS* and *LM* curve, let us now explain and illustrate the derivation of *IS* and *LM* curves.

28.4.1 Derivation of the *IS* Curve

The *IS* curve shows the relationship between the interest rate and the equilibrium level of national income, under the condition that $I = S$ at different rates of interest.

Derivation of the *IS* curve requires, therefore, the knowledge of the functional relationship between interest and investment, between saving and investment at equilibrium and between saving and income. All these functional relations, except one between the interest rate and investment, have already been discussed in the previous chapters. The nature of relationship between the investment and the interest rate is given by the investment function (28.1). In order to derive the *IS* schedule, let us recall here all the functional relations that figure in the analysis of the product market equilibrium. According to the Keynesian theory, product market equilibrium at the level of national income at which $I = S$.

Since

$I = I - hi$

and

$S = Y - C(Y)$,

the product-market equilibrium condition can be specified as

$I - hi = Y - C(Y)$

…(28.8)

In order to present these functions graphically, let us suppose that these functions are estimated factually and are given as follows.

$I = 200 - 2000\,i$

$C(Y) = 10 + 0.5\,Y$

$S = Y - (10 + 0.5\,Y)$

or

$S = -10 + 0.5\,Y$

…(28.9)

The derivation of the *IS*-schedule on the basis of these functions is illustrated graphically in Fig. 28.1. This figure is divided in four quadrants. The use of these quadrants for the derivation of the *IS* schedule are narrated below.

Fig. 28.1 *Derivation of the LM Curve*

Quadrant (a) presents the investment function: $I = 200 - 2000\,i$. The investment schedule shows an inverse relationship between the interest rate and investment. Recall that the product market is in equilibrium where $I = S$. This condition must hold for all the levels of investment at different rates of interest. For instance, as Fig. 28.1 shows investment at interest rate of 8 percent is `40 billion. So, for the product market to be in equilibrium, savings must be equal tò40 billion. And, when the interest goes down to 6 per cent, investment rises tò80 billion. So, for the product market to be in equilibrium, savings must rise tò80 billion. This relationship between the different levels of investment and savings is presented by the $I = S$ line in quadrant (b) of Fig. 28.1 with savings measured on the vertical axis and investment on the horizontal axis with the same scale as in quadrant (a). Since $S = I$ at all the levels of investment, the $I = S$ line is a straight 45° Iine.

The 45° -line in quadrant (b) gives the equilibrium levels of savings which will keep the product market in equilibrium, at different levels of investment. It implies that when

investment increases, savings must increase to the same extent. Given the saving function as $S = –10 + 0.5\,Y$, income must increase to yield savings enough to match with equilibrium. This relationship is given by the saving function in quadrant (c) of Fig. 28.1 with savings measured on the vertical axis and income on the horizontal axis. The saving function plotted in quadrant (*c*) gives the measure of the equilibrium levels of income at different levels of $I = S$. For example, at 8 percent interest, saving equals investment at `40 billion. As shown in quadrant ( *c*), a saving and investment of `40 billion produces an equilibrium level of income of `100 billion as shown by the point $J'$ . Similarly, when interest rate goes down to 6 per cent, $I = S$ rises tò80 billion. This level of $S = I$ generates income ( $Y$) of `180 billion, as shown by the point K' . Similarly, point L' can be located on the saving function.

Now what we need to do is to link the equilibrium levels of income with the corresponding interest rate and derive the *IS* schedule. The *IS* schedule has been obtained in quadrant (d) by linking the interest rates and the equilibrium levels of income.

At the equilibrium rate of interest of 8 per cent, $S = I$ at `40 billion. The saving schedule in quadrant (c) shows that at the equilibrium level of saving and investment of `40 billion, the equilibrium level of income is `100 billion. When we link up this level of income with the interest rate of 8 per cent, we get point *J* in quadrant (d). When interest fall to 6%, $S = I$ rises tò80 billion. With the rise in $S = I$ tò80 billion, the equilibrium level of Y rises tò180 billion. By linking this level of income to interest rate 6% in quadrant (d), we get a point *K.* We can similarly locate points *L*. By joining points *J, K* and *L,* we get the IS-schedule. ***The IS curve is a locus of points showing equilibrium points of the product market at various combinations of interest rate (i), investment (I), savings (S), and income (Y).***

The *IS* curve has ***two important implications. One,*** *IS* curve represents all the various combinations of interest (i) and income (Y), and investment (I) and saving (S) that keep the product market in equilibrium. The product market will not be in equilibrium at any point away from the *IS* curve. The reason is, all other points violate the equilibrium condition ($I = S$) of the product market. For example, at any point to the right of the *IS* curve, $S > I$, and at any point to the left of the *IS* curve, $S < I$. So the product market equilibrium has to be only on the *IS* curve. ***Two,*** the *IS* curve has a negative slope in relation to the income. Theoretically, it implies that national income is inversely related to the interest rate.

28.4.2 Derivation of the LM Curve

The LM curve shows the relationship between the interest rate and the equilibrium level of national income with money market in equilibrium. The *LM*- schedule can be derived straightaway from the money market equilibrium condition.

$M\text{s} = M\text{d}$

where,

$M\text{d} = M\text{t} + M\text{sp}$

$M\text{t} = kY$

$M\text{sp} = L(i)$

Thus, the money market equilibrium condition can be written as

$M\text{s} = kY + L(i)$

…(28.9)

Eq. (28.9) gives the *LM function*. It can be used to derive the *LM* curve. The derivation of the *LM* curve is illustrated in Fig. 28.2. This figure is also divided in four quadrants.

Quadrant (*a*) presents a *hypothetical* Keynesian *M* sp curve. The curvilinear *M* sp function for speculative demand for money is based on a function $M\text{sp} = L(i)$. Quadrant (*b*) shows the relationship between *speculative* demand (*M* sp) and *transaction* demand (*M* t) for money, given the *M* s. It gives the measure of *M* t at the equilibrium level of the money market, given the total supply of money (*M* s). Since at equilibrium $M\text{s} = M\text{d}$ and $M\text{d} = M\text{t} + M\text{sp}$, with *M* s constant, *M* t decreases when *M* sp increases and *vice versa*. In simple words, given the money supply, there is inverse relationship between M_t and M_{sp}, i.e., if one increases, the other decreases. This relationship is shown by a line marked $M_t = M_s - M_{sp}$. The significance of this line is that it gives M_t given the M_{sp} at different interest rates. For instance, given the supply of money (M_s), say, at `150 billion and $M_{sp}$ at `60 billion at interest rate of 6 per cent, the transaction demand for money, M_t = 150 – 60 = `90 billion.

(Amount in billion `)

Fig. 28.2 *Derivation of the LM Curve*

Quadrant (c) shows the derivation of M_t-demand curve at different levels of income (Y). It is based on the M_t-function given as $M_t = kY$ (where k = 0.5). Thus, $M_t = 0.5\,Y$. The M_t-curve gives the relationship between M_t and Y given the M_t-function. Given the M_t-function, if interest rate (i) is known, M_{sp} and M_t can be easily known, given the supply

of money. And, when M_t is known, the equilibrium level of income (Y) corresponding to M_t can be known. For example, if interest rate is 10 per cent, then M_{sp} = `30 billion and $M_t$ = `150 billion – `30 billion = `120 billion. Given the M_t function as $M_t = 0.5\,Y$, the equilibrium level of income can be obtained as

$M_t = 0.5Y$

$Y = M_t / 0.5$

Since M_t = 120 billion, Y = 120/0.5 = `240 billion

The equilibrium combination of M_t and Y at interest rate 10 per cent is shown by point *H* in quadrant (c). Similarly points *G* and *F* show the combination of M_t and Y at interest rates 6 per cent and 4 percent, respectively. By drawing a line through points *F, G* and *H,* we get M_t-function.

Quadrant (d) of Fig. 28.2 shows the derivation of the *LM* curve. The *LM* curve is derived by linking the different interest rates and the equilibrium levels of income as shown in quadrant (d). For example, at the interest rate of 10 per cent, M_{sp} = `30 billion. $M_t$, = `120 billion and the equilibrium level of income is `240 billion. By linking the equilibrium income of `240 billion with the interest rate of 10 per cent, we get point *H′* in quadrant (d). Similarly, when the interest rate decreases to 6 per cent, M_{sp} rises tò60 billion, and M_t decreases tò90 billion. By linking the equilibrium income of `180 billion with the interest rate of 6 per cent, we get a point *G′*. A number of such other points, for example, points *F′* can be located. By joining these points we get the *LM* curve, as shown in quadrant (d). ***The LM is a locus of points showing equilibrium points of the money market at different levels of interest, income and demand for money.***

28.4.3 Product and Money Market Equilibrium:

The Graphical Method

Having derived the *IS* and *LM* curves, we can now integrate them to find the general equilibrium, i.e., the simultaneous equilibrium of the product and money markets at the same interest rate and the level of income. Fig. 28.3 presents the *IS* and *LM* curves derived in figures 28.1 and 28.2, respectively. The *IS* curve shows the equilibrium levels of *Y* at different levels of internal rate (i) with the condition that $I = S$. Similarly, *LM* curve shows the equilibrium levels of *Y* at different levels of *i* with

Fig. 28.3 *The General equilibrium of Product and Money Markets*

the condition that $M_d = M_s$. As

shown in Fig. 28.3, the *IS* and *LM* curves intersect at point *E*. Point *E* gives the unique combination of *i* and *Y* that satisfies the equilibrium conditions of both the product and the money markets, that is, $I = S$ and $M_d = M_s$ at the same interest rate (*i*) and income (*Y*). Point *E* is, therefore, the point of *general equilibrium*. At point *E*, both product and money markets are in equilibrium at interest rate 6 per cent and income of `180 billion.

At all other interest rates, there is a divergence between the *IS* and *LM* curves. This divergence creates the condition of disequilibrium. For instance, suppose market rate of interest is given at 8%. At this interest rate, $I = S$ at the income level of `100 billion, as indicated by point *J*. But, for the money market to be in equilibrium at this rate of interest, an income level of `220 billion would be required, as indicated by point *K*. Thus, at interest rate of 8 per cent, product and money markets are not in equilibrium at the same level of income. There is, therefore, no general equilibrium at this interest rate. Similarly, at interest rate of 4 per cent, money market is in equilibrium at point *M*, with corresponding income of `120 billion, and product market is in equilibrium at point *T*, with corresponding income of `260 billion. The economy is, therefore, in disequilibrium. The system has, however, a tendency to converge to point *E*, the point of the general equilibrium. In fact, the disequilibrium conditions themselves create conditions for the sectoral adjustment making the economy move towards the point of general equilibrium.

28.5 THREE-SECTOR IS-LM MODEL: THE CLOSED ECONOMY MODEL

In this section, we proceed to discuss the three-sector *IS-LM* model, known also as 'closed economy model'. The three-sector model consists of three sectors of the economy, viz., households, firms, and the government. Obviously, three-sector model is created by adding the government sector to two-sector model. Inclusion of the government sector into the *IS-LM* model requires adding government related variables to the *IS-LM* model. Government related variables are (i) *fiscal variables* including government taxes (*T*) and expenditure (*G*), and (ii) *monetary variables* including changes in money supply (M_s) and in money demand (M_d) effected by the government – in fact, by the central bank of the country.

It may be added here that analysing the effects of fiscal changes on the economy is somewhat more complex that the effects of monetary changes. The changes made by the government in money supply and demand only make the *LM* curve shift to the right or to the left. Increasing money supply makes *LM* curve shifts to the right. It causes decrease in the interest rate. And, decreasing money supply makes *LM* curve shift to the left causing an increase in the interest rate. In our analysis of *IS-LM* model in three-sector framework, therefore, we will concentrate on the effects of the fiscal variables, i.e., *tax* (*T*) and *expenditure* (*G*). The effect of changes in money supply and demand will be shown in the final framework of the three-sector model. Now we proceed to explain the derivation of the *IS* curve in three-sector model.

28.5.1 Derivation of the *IS* Curve with Government Sector

An easy way to derive *IS* curve with the government sector is to recall the three-sector model of product market equilibrium. The three-sector product market model is constructed

by incorporating government sector variables, *G* and *T*, into the two-sector model. Recall that the two-sector product market model is given as

$$C + I = C + S \quad \ldots(28.10)$$

Let us now incorporate the government expenditure (*G*) and taxes (*T*) into the model given in Eq. (28.10) to construct the three-sector model. After the addition of *G* and *T*, the three-sector product market equilibrium condition is expressed as follows.

$$C + I + G = C + S + T \quad \ldots(28.11)$$

Having recalled the Keynesian model of product-sector equilibrium for a three-sector economy, we may now proceed to construct the Hicksian *IS* model. The Hicksian *IS* model is presented here under the assumption that government imposes a fixed tax, i.e., lump sum tax[2].

Let us begin by constructing the *IS* Model under Hicksian framework. The basic purpose of the *IS* Model is to present a simple framework for the derivation of the *IS* schedule. The Hicksian *IS* Model, as presented below, is based on the following assumptions.

(i) The government spending (G) is determined exogenously and remains constant;

(ii) The lump-sum tax (T) is determined by the government and remains constant; and

(iii) The government follows a balanced budget policy, i.e., G = T.

Given these assumptions, the product-market equilibrium condition given in Eq.

(28.11) can be reduced to

$I + G = S + T$

…(28.12)

Eq. (28.12) provides the basis for the for the derivation of the *IS* function for three-

sector model. For the purpose of deriving the *IS* schedule, the variables in Eq, (28.11)

need to be converted into functional form. For the purpose of specifying the *IS* model,

let us suppose that the variables in Eq, (28.11) are given as follows.

$C = a + b\ Yd$ (where $Yd = Y – T$)

…(28.13)

$I = I – hi$ ($h = \Delta I/\Delta i$ and $h > 0$)

…(28.14)

$S = – a + (1 – b)\ Yd = – a + (l – b)\ (Y – \text{T})$

…(28.15)

$G = \hat{G}$

…(28.16)

$T = \check{T}$

…(28.17)

and

$\hat{G} = \check{T}$

…(28.18)

Given these functions. the *IS* model given in Eq. (28.12) can be expressed as follows.

$I + G = S + \check{T}$

$\bar{I} – hi + \hat{G} = – a + (1 – b)\ (Y – \check{T}) + \check{T}$

…(28.19)

Eq. (28.19) gives the theoretical basis of deriving the *IS* curve. However, for the sake of

clarity and convenience in drawing the *IS* curve, let us suppose that the numerical forms of

investment function (Eq. 28.14), saving function (Eq. 28.15), and G and T are given as follows.

$I = 200 – 2000\ i$

…(28.20)

$S = – 10 + 0.5\ (Y – T)$

…(28.21)

$\hat{G} = \check{T} = 40$

…(28.22)

2. The three-sector *IS* curve can also be drawn by asing a tax function rather than lump sum tax. For details, see

D.N. Dwivedi, *Macroeconomics: Theory and Policy* (Tata McGraw-Hill, New Delhi, 2010). Ch. 17.

By substitution, the *IS* model given in Eq. (28.19) can be written as follows.

$200 – 2000\ i + 40 = – 10 + 0.5\ (Y – 40) + 40$

…(28.23)

$240 – 2000\ i = 0.5\ Y + 10$

…(28.24)

$Y = 460 – 4000\ i$

…(28.25)

Eq. (28.25) can be straightaway used to derive the *IS* curve. However, we would

illustrate here the entire process of graphical derivation of the *IS* curve for three-sector

model on the basis of Eq. (28.24).

The graphical derivation of three-sector *IS* curve on the basis of the Eq. (28.24) is

illustrated in Fig. 28.4. For the sake of brevity, LHS of Eq. (28.24) has been denoted as

Fig. 28.4 *Derivation of* IS *Curve for Three-Sector Model-I*

$I(i) + G$ and RHS of the equation as $S(Y) + T$. Quadrant (a) presents the $I(i) + G$ curve along the investment function. The $I(i) + G$ curve has been drawn on the basis of the LHS

of Eq. (28.24). This curve shows $I + G$ at different rates of interest. For example, at the

interest rate of 8 per cent, $I + G$ = 240 – 2000 (0.08) = 80 as denoted by point A and at 6 per cent, $I + G = 240 – 2000\ (0.06) = 120$ as denoted by point *B*. Thus, the $I + G$ schedule has been drawn by joining the points *A* and *B*. Once $I + G$ is determined, we need to find out $S + T$ equal to $I + G$ because for the product market to be in equilibrium, $S + T$ must be equal to $I + G$. The process of finding $S + T$ that equals $I + G$ is shown in quadrant (b) of Fig. 28. 4. The line marked $I + G = S + T$ gives the $S + T$ at different levels of $I + G$.

After estimating the value of $S + T$ at different levels of $I + G$ at different rates of interest, what we need now is to find the equilibrium level of income (Y) that can generate the equilibrium levels of $S + T$. The process of estimating the equilibrium level of Y is illustrated in quadrant (c) of Fig. 28.4. The schedule marked $S(Y) + T$ is drawn on the basis of the RHS of Eq. (28.24), i.e., $S + T = 0.5Y + 10$. Given the $S + T$ function, the equilibrium level of income can be determined by linking the different equilibrium levels of $S + T$ with Y measured at the horizontal axis of quadrant (c). For example, point *L* shows that for $S + T = 80$, an equilibrium income of 140 is required. Similarly point *M* shows that for $S + T = 120$, an income (Y) worth 220 is required.

Once equilibrium level of Y is determined for different levels of $I + G = S + T$, at different interest rates the *IS* curve can be derived by linking the equilibrium levels of Y measured in quadrant (c) with different rate of interest given in quadrant (a). The process is illustrated in quadrant (d) of Fig. 28.4. The linking process is shown by the lines linking the values in different quadrants. For example, at the interest rate of 8 per cent, $I + G = 80$. For the product market to be in equilibrium $S + T$ *must* be equal to 80. To generate $S + T = 80$, an income of 140 is required. When we link $Y = 140$ with interest rate 8 per cent, we get point *J* in quadrant (d). Similarly, we locale point *K*. By joining points *L* and *K* and extending it up and down, we derive the *IS* curve for the three-sector model.

28.5.2 Monetary Changes and Money-Market Equilibrium

In this section, we discuss changes in money supply and its effect on the money market and on the *LM* schedule. For the sake of convenience in deriving the *LM* schedule and analysing the effects of monetary changes, we assume *a linear aggregate money demand function*. We continue to assume that money supply (M_s) is determined exogenously at $M_s = 200$ (` billion) and it is interest-inelastic.

In order to derive the *LM*-function, let us recall our M_d equation given as:

$M_d = M_t + M_{sp}$

For further elaboration of the monetary sector model, let us use here our earlier M_t-function, given in section 28.3.2 as

$M_t = 0.5\ Y$

…(28.26)

and a modified M_{sp} function as $M_{sp} = 100 – 2500\ i$

…(28.27)

Given the M_t and M_{sp} functions, in Eq. (28.26) and (28.27), respectively, M_d-function can now be formulated by combining the two functions as

$M_d = M_t + M_{sp}$

$M_d = 0.5\ Y + 100 – 2500\ i$

…(28.28)

Given the money supply at M_s = `200 (billion), the money-market equilibrium condition can be expressed as

$M_s = M_d$

$200 = 0.5\ T + 100 – 2500\ i$

…(28.29)

From Eq. (28.29), the *LM-function* can be derived as given below.

$Y = 200 + 5000\ i$

…(28.30)

The *LM* function given in Eq. (28.30) is shown graphically in Fig. 28.5 by the curve LM_0. The curve LM_0 shows the money-market equilibria at various rates of interest and corresponding levels of income.

Change in Money Supply and Shift in the *LM* Curve

Let us now explain the effect of change in money supply on the *LM* curve. A change in money supply causes a shift in the *LM* curve. Suppose that the central bank increases the money supply by `100 billion so that the money supply rises tò300 billion 200bn

+ 100bn. The money-market equilibrium condition can then be written as

$M_s + \Delta M_s = M_d$

…(28.31)

Since M_s = 200 (billion) and ΔM_s = 100 (billion), by substitution, the money market equilibrium can be expressed as

$200 + 100 = 100 + 0.5\,Y - 2500\,i$

$200 = 0.5\,Y - 2500\,i$

…(28.32)

Given the money market equilibrium in Eq. (28.32), the *LM* function can be derived as

$Y = 400 + 5000\,i$

…(28.33)

The *LM* function given in Eq. (28.33) is graphically presented by the curve LM_1 in Fig. 28.5. Note that, with the increase in money supply, the *LM* curve makes a parallel shift rightward from LM_0 to LM_1. The shift in the *LM* curve equals $1/k \times \Delta M_s$.

(Billion `)

Fig. 28.5 *The shift in LM curve*

Since by assumption, $k = 0.5$ (see Eq. 28.26) and ΔM_s = 100 billion, the shift in the *LM* curve can be measured as

Shift in *LM-curve* = 100 bn (1/0.5) = 200 billion

28.5.3 Product and Money Market Equilibrium in Three-Sector *IS-LM* Model

In this section, we combine the analytical models of the product and monetary sectors and show how both product and money markets attain equilibrium simultaneously at the same level of income and interest rate. This analysis shows how general equilibrium is determined in three-sector *IS-LM* model.

To present a complete analysis of the general equilibrium in the *IS-LM* model, let us recall the product and money market models used above.

The product-sector *IS*-function is given in Eq. (28.25) as $Y = 460 - 4000\,i$. For graphical convenience, we modify this *IS* function as

$Y_{IS} = 750 - 5000\,i$

…(28.34)

The *LM* function is given in Eq. (28.30) as

$Y_{LM} = 200 + 5000\,i$

…(28.35)

The *IS* and *LM* function can now be used to show the determination of the general equilibrium.

Recall that, **general equilibrium** takes place where product and money markets are simultaneously in equilibrium. This condition is satisfied where

IS = *LM*

The *IS*-schedule is given in Eq. (28.34) and *LM*- schedule in Eq. (28.35). So the general equilibrium condition can be expressed as

$750 - 5000\,i = 200 + 5000\,i$

…(28.36)

Now the equilibrium values of *Y* and *i* can be obtained from Eq. (28.36) as shown below.

$750 - 5000\,i = 200 + 5000\,i$

$550 = 10{,}000\,i$

$i = 0.055$ or 5.5%

By substituting 0.055 for *i* in the *IS* function (or in the *LM* function), we get the equilibrium level of income (*Y*) as follows.

$Y = 750 - 5000\,(0.055) = 200 + 5000\,(0.055)$

= 475 (billion)

It may thus be concluded that the product and money markets are simultaneously in equilibrium at interest rate 5.5% and national income of `475 billion.

Graphical Solution

The *IS* and *LM* schedules given in Eq. (28.34) and Eq. (28.35), respectively, are graphed in Fig. 28.6. As Fig. 28.6 shows, the *IS* and *LM* schedules intersect at point *E* where both the product and money markets are simultaneously in equilibrium. Thus, point *E* is the point of general equilibrium where Y = `475 billion and i = 5.5%.

Fig. 28.6 *The general equilibrium in Three-sector IS-LM Model*

Given the *IS* and *LM* schedules, the economy cannot be in the state of the general equilibrium at any point other than point *E*. For, at no other point the product and the money markets are simultaneously in equilibrium. For example, at the interest rate of 8%, the product market is in equilibrium at point *B* and the money market is in equilibrium at point C. Between points *B* and *C*, the economy will be in disequilibrium. Similarly, at interest rate of 3%, the product market is in equilibrium at point *D* and the money market is in equilibrium at point *A*. Between points *A* and *D*, the two markets are not in equilibrium simultaneously. This shows that at any interest rate other than 5.5% and income level other than `475 billion, the two markets are not simultaneously in equilibrium. Therefore, there cannot be general equilibrium at any other point.

28.5.4 General Equilibrium with Simultaneous Fiscal and Monetary Changes

Let us now analyse the effect of simultaneous fiscal and monetary changes on the general equilibrium and look into their net effect. Suppose initial *IS* and *LM* curves are the same as given in Fig. 28.6. These curves are reproduced in Fig. 28.7 as IS_0 and LM_0, respectively, and the general equilibrium is determined at point *A* at income level of `475 billion. Given the general equilibrium, let us suppose that the government increases its expenditure by `100 billion and money supply by an equal amount, so that $\Delta G = 100 = \Delta M_s$. This means deficit financing of public expenditure. The increase in government spending by ΔG makes the *IS* schedule to shift upward from IS_0 to IS_1. And increase in money supply by ΔM_s makes the *LM* schedule to shift rightward from LM_0 to LM_1. The *IS*-schedule with fiscal change of $\Delta G = 100$ is given by the schedule IS_1 in Fig. 28.7, and *LM*-schedule with money supply change by $\Delta M_s = 100$ is given by the schedule LM_1 in Fig. 28.7. As the figure shows, the IS_1 and LM_1 schedules intersect at point *D* which represents the point of simultaneous equilibrium of the product and money markets. At point *D*, both product and money markets are simultaneously in equilibrium.

That is , $I + G = S + T = M_s = M_d$. At the general equilibrium, Y = `700 billion and i = 6 per cent.

Fig. 28.7 *The IS-LM Model with Fiscal and Monetary Changes*

The determination of the general equilibrium at point *D* can be proved algebraically.

With ΔG = 100 bn, the *IS*-schedule change from $Y_{IS} = 750 - 5000\,i$, given in Eq. (28.34), to

$$Y_{IS} = 1000 - 5000\,i$$

and with ΔM_s = 100 bn, the *LM* function changes from $Y_{LM} = 200 + 5000\,i$ given in Eq. (28.35) to

$$Y_{LM} = 400 + 5000\,i$$

By rearranging the terms of these equations, we get the *IS* function as

$$Y + 5000\,i = 1000 \qquad (i)$$

and *LM* function as

$$Y - 5000\,i = 400 \qquad (ii)$$

Thus, we have two simultaneous Eqs. (*i*) and (*ii*), with two unknowns, *Y* and *i*. By solving these equations, we get the equilibrium rate of *i* and the equilibrium level of *Y*, as shown below.

By subtracting Eq. (*ii*) from (*i*), we get

$$10{,}000\,i = 600$$

$$i = 0.06 \text{ or } 6 \text{ per cent.}$$

By substitution, we get equilibrium level of income from equation (*ii*) as:

$$Y = 400 + 5000\,(0.06)$$

= `700 billion.

Thus, at general equilibrium Y = `700 billion and interest rate = 6 per cent. It may thus be concluded that with $\Delta G = \Delta M_s$ = 100 billion, given the *IS* and *LM* function, there is rise in both equilibrium level of income and interest rate.

28.6 FOUR-SECTOR IS-LM MODEL: THE OPEN ECONOMY MODEL

This section presents the analysis of the four-sector model – the open economy model – of the *IS-LM* model of the general equilibrium. The four-sector *IS-LM* model is constructed by adding foreign sector to the three-sector model. Adding foreign sector to the *IS-LM* model requires incorporating *international transactions*. There are two kinds of international transactions: (*i*) *autonomous transactions*, the transactions that are made to meet the economic needs of the country, and (*ii*) *induced transactions, i.e.,* the transactions that arise due to the autonomous transactions. Specifically, *autonomous transactions* include export and import of goods, services and capital. And, *induced transactions* consist of payments received for exports (*X)* and payments made for imports (*M*). The net of exports and imports is known as trade balance which equals *X less M, i.e.,* trade balance = *X* – *M*. In our analysis of the four-sector model, however, we use a simplified model including only two variables, viz., exports (*X*) and imports (*M*).

It may be added at the outset that in our simplified four-sector *IS-LM* model, we will limit our analysis to the effects of exports and imports on the *product sector* and explain the derivation of only the *IS* curve. As regards the derivation of the *LM* curve, we make simplifying assumption that *monetary sector* remains unaffected by the international transactions and we will use the *LM* as derived in three-sector analysis of the *IS-LM* model.

28.6.1 Derivation of the *IS* Curve

To begin with, let us recall our three-sector product-market equilibrium model given as

$C + I + G = C + S + T$

To this three-sector model, we now add the foreign trade—the exports (*X*) and imports (*M*). With the addition of *X* and *M,* the four-sector product-market equilibrium condition is written as

$C + I + G + (X – M) – C + S + T$

…(28.37)

The variables *X* and *M* need some explanation and quantification. Exports (*X*) of a country depend on a variety of factors governing the foreign demand for its goods and services. The inclusion of foreign demand parameters in the domestic model of a country is neither an easy task nor a necessity for a simplified model. Therefore, *X* is assumed to be a constant factor, that is,

$X = X$

As regards imports, imports (*M*) of a country are a function of a number of factors. However, for the sake of analytical simplicity, imports are treated as the function of the country's national income (*Y*). That is, import function takes the following form.

$M = M + mY$

where, *M* is autonomous import and *m* is *marginal propensity to import*, i.e., the proportion of marginal national income spent on imports.

With *X* and *M* defined, the four-sector product-market equilibrium condition given in Eq. 28.37 can be rewritten as

$Y = C + I + G + X - M – mY = C + S + T$

…(28.38)

Fig. 28.8 *Derivation of the IS Curve in Four-sector Model*

Eq. (28.38) provides the basis for the derivation of the *IS* curve in the four-sector model. Fig. 28.8 illustrates the derivation of the *IS*-curve without and with foreign trade. The *I* + *G* schedule in quadrant (a), *S* + *T* schedule in quadrant (c), and *IS* c schedule in quadrant (d) present the product-market equilibrium conditions for an economy without foreign trade—the closed economy. Let us now look at the change in all these functions caused by the inclusion of foreign trade into the *IS- LM* model.

Note that, exports are *injections* and imports are *withdrawals*. The net of injections and withdrawals is measured as *X* – *M*. With the inclusion of foreign trade in the model, the net injection (assuming *X* > *M)* increases and, therefore, the *I* + *G* curve in quadrant (a) shifts to *I* + *G* + (*X* – *M*) curve. Note that the shift is parallel because exports (*X*) are assumed to be exogeneously determined and *X* – *M* is assumed to remain constant. The rest of the process of deriving *IS* function is the same as one used in previous section. However, the derivation of the *IS* schedule in four-sector model is explained here briefly.

With foreign trade, the product market equilibrium condition is given by $I + G + X = S + T + M$. This condition is shown in quadrant (b) of Fig. 28.8 by the 45° line. The relationship between $S + T + M$ and income (Y) is shown by the $S + T + M(Y)$ function with foreign trade. Note that the gap between the $S + T$ and $S + T + M(Y)$ schedules goes on increasing

with the increase in income. The divergence between the $S + T$ and $S + T + M(Y)$ schedules depends on the value of m (the marginal propensity to import) in the import function given as $M =$

$+ mY$. The greater the value of m, the greater the divergence and *vice versa*.

The $S + T + M(Y)$ function when linked to Y and i produces IS_0 schedule in quadrant (d) which represents the IS function with foreign trade—subscript 'o' denotes 'open economy'. Note that the IS schedule becomes steeper with the inclusion of the foreign trade.

Let us explain briefly the derivation of IS schedule with foreign trade, as exhibited in Fig. 28.8. The IS_c schedule represents the closed economy, i.e., three-sector model, drawn on the basis of $I + G$ schedule in quadrant (a). Now, given the interest rate i_2, let us add foreign sector to $I + G$. schedule. As noted above, foreign sector net is given by $X - M$. With the addition of $(X - M)$ with $X > M$, the $I + G$ schedule shifts to $I + G + (X - M)$, under the condition that $X > M$. Note that $X - M = MN$ in quadrant (a) . As such, for the economy to be in equilibrium, imports (M) must rise by MN. This is shown by QR in quadrant (b). Note that $MN = JK = QR$. The rise in imports (M) by QR, given the income at Y_1, makes point A shift to point B in quadrant (c). When Y_1 and i_2 are linked, as shown in quadrant (d), point E on IS_c is arrived at. Note that point E falls also on the IS_c scheudle. It implies that there is no change in product-market equilibrium.

Now let the interest rate fall to i_1. As a result, investment (I) increases by NT. Increase in investment causes an increase in income (Y). With no increase in foreign trade, income would have increased to Y_3 as shown in quadrant (c). But in open economy case, increase in income causes increase in M because $M = f(Y)$. Imports, being a leakage, reduce the income expect ed from $\Delta I = NT$. As a result, income increases but not to Y_3 – income increases to Y_2 only.

When we link the new $S + T + M$ with a relatively lower level of income (Y_2), we get point C in quadrant (c). By linking Y_2 with i_1, we arrive at point D in quadrant (d). By joining points E and D by a line and extending it further, we get open economy IS-schedule as shown by IS_F. Note that the IS_F curve is steeper than the IS_c curve.

28.6.2 Four-sector General Equilibrium

Having explained the derivation of the four-sector IS function, we proceed now to explain the determination of the general equilibrium in the open economy model. We will also explain the effect of shift in the IS schedule on the equilibrium level of the output and the interest rate. Let us assume once again that money market is not affected by the introduction of foreign trade and that the LM function remains unaffected. Figure 28.9 illustrates the general equilibrium based on these assumptions. Suppose that the initial IS schedule is given by IS_2—it represents IS_F schedule in quadrant (d) of Fig. 28.8. The IS_2 and LM schedules intersect at point E_2. Point E_2 is therefore the point of general equilibrium in the four-sector model. This point of equilibrium will remain stable until there is a shift in IS and LM schedules.

Let us now look at the effect of the shift in the IS schedule on the general equilibrium, assuming no shift in the LM schedule. A shift in the IS curve is caused by a change in the internal factors including, I, G, S and T and external factors including X and M. Confining our analysis to the foreign sector, however, let us assume that the internal factors— I, G, S and T—

remain constant and the shift in the IS curve is caused by the change in the external factors (X and M).

The shift in the initial *IS* 2 curve
(leftward or rightward) results either from
a change in *X* or a change in *M* or in
both, other things remaining the same.
When *X* increases or
in the import
function *(M =*
+ mY) decreases, the
IS curve shifts rightward as shown by
its shift from *IS* 2 to *IS* 3. And, when *X*
decreases or
increases, the *IS* curve
shifts leftward as shown by its shift from
IS 2 to *IS* 1 in Fig. 28.9. In case *X* and *M*
change simultaneously, and if Δ *X* > Δ *M*,
IS schedule shifts upward and if Δ *X* <
Δ *M*, the *IS* curve shifts downward. Let us
now look at the consequences of upward

Fig. 28.9 *Shift in the IS Curve and the*

and down shifts in the *IS* schedule, *LM*

General Equilibrium

schedule remaining the same.

It can be seen in Fig. 28.9 that an upward shift in the *IS* schedule, *LM* curve remaining the same, causes a rise in the equilibrium levels of both the interest rate and the income. For example, with an upward shift in the *IS* schedule from *IS* 2 to *IS* 3, the equilibrium point shifts from *E* 2 to *E* 3. With this shift in the equilibrium point, the equilibrium rate of interest increases from *Oi* 2 to *Oi* 3 and the equilibrium level of income increases from *OY* 2 to *OY* 3. Similarly, a downward shift in the *IS* schedule from *IS* 2 to *IS* 1, *LM curve remaining the same*, the equilibrium point shifts from *E* 2 to E1. The downward shift in the *IS* schedule causes a fall in the equilibrium interest rate from *Oi* 2 to *Oi* 1 and the equilibrium level of income decreases from *OY* 2 to *OY* 1.

This analysis of the four-sector *IS- LM* model takes us to the end of the discussion on the Hicksian analysis of determination of the general equilibrium in an open economy.

SUMMARY

- The *IS-LM* model was constructed by J.R. Hicks in 1937. It was an improvement in the revolutionary macroeconomic theories of John Maynard Keynes including his theory of income determination and monetary theory.
- Keynes had constructed his theories of income determination and money market in isolation of each other while working of both the product and monetary sectors are interdependent and interrelated.
- Hicks highlighted the interdependence and interrelations of the product and monetary sectors and argued strongly that unless both the sectors attain their equilibrium simultaneously at the same level of national income, none of these sectors reach equilibrium, nor will the economy attain the stable equilibrium.
- Hicks developed his *IS-LM* model by combining Keynesian theory of income

determination and monetary theory. According to Keynes, the equilibrium of national income is determined by the aggregate demand (*AD*) and aggregate supply (*AS*). In a simple economy, $AD = C + I$, where C = consumer demand, and I = investment demand, and $AS = C + S$, where S = savings . The determined C being common to

both *AD* and *AS,* national income equilibrium is determined where $I = S$. Hicks abbreviated this terms as *IS.*

- According to Keynes, monetary sector reaches equilibrium when demand for money is equal to supply of money. Keynes designated money demand as liquidity preference (L) and demand for money as M. Thus, the equilibrium condition for money market is $L+M$. Hicks abbreviated it *LM*.
- By showing the interlinkage of product and monetary sectors and combining equilibrium conditions of the two sectors, Hicks developed his *IS-LM* model.
- Hicks showed the dependence of monetary sector to the product sector by linking transaction demand for money to income, i.e., $Mt = f(Y)$. He showed the dependence of product sector on monetary sector by linking investment demand to interest rate, i.e., $I = f(i)$.
- Having showed the interrelationship between the product and monetary sectors, Hicks derived the product sector equilibrium condition as $Y = C(Y) + I(i)$, with $I = S$ at different levels of Y and i. The equation $Y = C(Y) + I(i)$ provides the basis of the derivation of the *IS curve.* Similarly, he derived the monetary sector equilibrium condition as $Ms = kY + L(i)$. This equation provides the basis for the derivation of the *LM* curve, with $L = M$ at different levels of Y and i.
- Having propounded the *IS* and *LM* formula , Hicks offered the general equilibrium model as $IS = LM$ or $C(Y) + I\ (i) = kY + L(i)$. This model is related to two-sector framework. Other economists extended the two-sector model to three-sector economy (closed economy) and four-sector economy (open economy) models.

REVIEW QUESTIONS AND EXERCISES

1. Distinguish between the product and money markets. Explain the interdependence of the two markets. How does it matter in determining the general equilibrium of the economy?
2. What is the basis of derivation of the *IS* curve? Derive the *IS* graphically in two-sector model and explain the relationship between interest rate and national income.
3. Suppose consumption function and investment functions are given as follows.

$C = 20 + 0.5\ Y$

and

$I = 200 – 2000\ i$

Find the equilibrium level of income at interest rate of 8 per cent.

4. Explain the basis of the derivation of the *LM* curve. Derive the *LM curve* graphically and explain the determination of the equilibrium at different rate interest and the level of income.
5. Suppose money market functions are given as follows.

$M\text{t} = 0.5\ Y$

$M\text{s} = 100 – 1500\ i$

$M\text{s} = 150$

Derive the *LM* function.

6. Explain the concept and conditions for the general equilibrium. Show graphically the determination of the general equilibrium by drawing the *IS* and *LM* curves. Show also how general equilibrium is disturbed by change in the interest rate.

7. Suppose in two-sector economy, product-sector functions are given as follows.

$C = 10 + 0.5\ Y$

$I = 200 – 2000\ i$

And monetary sector functions are given as

$M\text{s} = 150$

$M\text{t} = 0.5\ Y$ and

$M\text{s} = 150 – 1500\ i$

Find (a) *IS* and *LM* functions, and

(b) Find income and interest rate the general equilibrium.

8. How does addition of the government sector to the two-sector model change the *IS-LM* model of

the general equilibrium analysis? What new variables are added to the model and how do they ffect the *IS* and *LM* curves?

9. Suppose a product market model is given as follows.

$C = a + b(Y - T)$

$I = \hat{I} - hi$

$G = G$

$T = T + tY$

Derive the product market equilibrium condition and the *IS* curve .

10. Suppose product sector model is given as $C = 100 + 0.75\,Y$ and $I = 120 - 0.5\,i$ and money market model is given as $M_s = 150$ and $M_d = 0.2Y - 4\,i$. Find equilibrium level of *Y* and interest rate.

11. What is an open economy? How does addition of foreign sector to closed economy model affect the *IS-LM* model?

12. What are the determinants of export and imports of a country? Why are exports considered as an injection into and imports as leakages from the economy?

13. Explain and illustrate graphically how exports and imports affect the *IS* function of the three-sector model and the *IS* curve. Explain also how *IS* curve is affected if exports exceed imports and imports exceed the exports.

14. Given the closed economy equilibrium condition as $C + I = S + T$, suppose $X = X$ and import function is given as $M = M - m\,Y$. Derive graphically the *IS* curve for both the closed and open economies.

15. Suppose economic functions of an economy are given as follows.

$C = 100 + 0.8\,Y$

$I = 150 - 6\,i$

$G = 100$

$T = 0.25\,Y$

$M_d = 0.2\,Y - 2\,i$

$M_s = 300$

$X = 100$

$M = 20 + 0.1\,Y$

Find (*i*) The equilibrium level of income and interest rate in closed economy case,

(*ii*) The equilibrium level of income and interest rate in open economy case.

FURTHER READING

Ackley, G., *Macroeconomics: Theory and Policy* (London: Macmillan, 1978).

BrAnson, William H., *Macroeconomic Theory and Policy,* 3rd Ed. (New Delhi: East-West press, 2005).

cArlin, Wendy and Soskice, David, *Macroeconomics: Imperfections, Institutions & Policies* (New Delhi: Oxford University Press, 2006).

DernBurg, T. F., *Macroeconomics: Concepts, Theories and Policies, 7th* Ed. (New York: McGraw-Hill, 1985).

DornBusch, R. and Fischer, S. and Startz, R., *Macroeconomics,* 9th Ed., (New Delhi: Tata McGraw-Hill, 2004).

DwiveDi, D. N., *Macroeconomics: Theory and Policy,* 3rd Ed. (New Delhi: McGraw-Hill, 2010).

eDgmAnD, Michel R., *Macroeconomics: Theory and Policy,* 2nd Ed. (New Delhi: Prentice Hall of India, 1983).

Froyen, Richard T., *Macroeconomics: Theories and Policies,* 7th Ed (Delhi: Pearson Education, Inc., 2002).

mAnkiw, N. Gregory, *Macroeconomics,* 5th Ed. (New York: Worth Publishers, 2003).

shApiro, E., *Macroeconomic Analysis,* 5th ed. (New Delhi: Galgotia Publications (P) Ltd, 1994).

CHAPTER

29 Modern Theories of Economic Growth

CHAPTER OBJECTIVES

The objective of this chapter is to present a brief discussion on the modern theories of economic growth. The theories of economic growth give a view of what factors play important role in economic growth of the country and how they contribute to economic growth. A basic understanding of the growth theories guides the business managers how to assess the business prospects in the country. In this chapter, we discuss the following aspects.

• Meaning of economic growth,

• Determinants of economic growth of a country,

• The Harrod-Domar theory of economic growth,

• The Neo-Classical theory of economic growth, and

• How growth theories reveal the business prospects.

29.1 INTRODUCTION

The overall economic trend in a country is the most important element of its business environment. A growing economy with a considerably high and sustained growth rate in its *GDP* over a long period of time provides a promising business prospect and builds business confidence. For instance, the Indian economy had registered an annual growth rate of 9.6 per cent in 2006–07 and 9.4 per cent in 2007–08. A growth rate above 9 per cent was one of the highest in the world. This growth rate, along with economic reforms and liberalization, had built a very optimistic business environment. So much so that multinationals have entered the Indian market in a large way. Even though growth declined to about 5 per cent in 2013–14, the Indian economy is being called 'the future economic power of the world' and 'Indian economy is emerging as a world economic power' and so on, despite occasional recessions and hiccups. The annual inflow of Foreign Direct Investment (*FDI*) has increased from $ 155 million in 1991 to $ 4.7 billion in 2005–061 though *FDI* inflow to India is much lower than *FDI* in China.2 However, a downtrend in the economy or recession like situation beginning

1. *Economic Survey*—2006–2007, Ministry of Finance, GOI, p. 128.

2. According to the World Investment Report (*UN*), India received an estimated *FDI* of US $ 3264 million in 1997 whereas China received *FDI* of US $ 45300 million in this year, reported in *Economic Survey*—1998–1999, Ministry of Finance, GOI, p. 86, Table 6.7. The trend has not changed significantly since then.

in 2011–12 has reduced the business activities and confidence. The economic downtrend reduced overall consumer demand and, therefore, business prospects. Capital has begun to flow out of the country, accelerating the process of downslide. It is, therefore, important for the business decision-makers to understand the logistics of economic growth and to have a general awareness about growth parameters. This is all the more important where business decisions have long-run repercussions.

In this chapter, we discuss some major aspects of economic growth and growth theories. Our purpose here is not to provide a full-length discussion on economics of growth. We will concentrate on some basic aspects of economic growth. We will first discuss the meaning of economic growth. This will be followed by a brief discussion on the determinants of economic growth. Finally, we will discuss two modern theories of economic growth, viz., (*i*) the Harrod-Domar theory of growth and (*ii*) neo-classical theory of economic growth.

29.2 MEANING OF ECONOMIC GROWTH

In general, economic growth means continuous increase in *GDP* or *GNP* over a period of time at a reasonably high rate. In real sense of the term, however, economic growth means a sustained increase in *per capita* national output or net national product over a long period of time. It implies that the rate of increase in total output must be greater than the rate of population growth. For instance, India's *GNP* (at 1999–2000 prices) had increased at 9.1 per cent in 2005–06, and population had increased at about 1.7 per cent in this year. Consequently, *per capita*, *GNP* had grown at only 7.4 per cent. In real sense of the term, therefore, *GNP* has grown only at 7.4 per cent. It may be asked here : 'Is there no growth in a country where nation's output and population increase at the same rate so that per capita output remains constant ?' And, is there growth in a country where both output and population decrease—the former decreasing at a lower rate than the latter—so that per capita output increases? The answer to these questions is certainly in the negative. For, if output and population grow at the same rate, there would be no increase in per capita income and there would be no improvement in the general standard of living despite increase in the output. Such a growth is considered to be as good as stagnation in the economy. On the other hand, increase in per capita income as a result of decrease in population at a rate higher than the rate of decrease in output amounts to general decay in the economy: there is no growth despite increase in per capita income. Thus, economic growth implies a considerable and sustained increase in per capita income with or without increase in population.

Another qualification of economic growth is that the national output should be composed of such goods and services which satisfy the maximum wants of the maximum number of people. Besides, for economic growth to be genuine, the increase in output must be sustained over a long period of time. A short-run increase followed by a similar decrease in the output does not mean economic growth. Also seasonal, occasional and cyclical increases in output do not satisfy the conditions of sustained economic growth.

29.3 Determinants of Economic Growth

There are five most important determinants of economic growth of a country, viz.,

1. Human resources and its quality,

2. Natural resources of the country,
3. The rate of capital formation,
4. Technological development, and
5. Political and Social environment.

The first four factors are considered as the "four wheels" of economic growth.

The social and political factors make the fifth factor. Let us now see how these factors contribute to economic growth of a country.

29.3.1 Human Resource and its Quality

Human resource of a country is the most crucial factor in its economic growth. Human resource is comprised of the available labour force and its quality. Quality of labour force depends on the level of its education, training, skills, and its inventive and innovative abilities. Quantity and quality of manpower are both equally important. However, an excees supply of unskilled labour force, as is the case in most *LDCs* including India, is of little consequence. On the other hand, scarcity of skilled labour in the US is proving a serious constraint to its economic growth. The labour force along with its skill is the source of all goods and services.

Apart from quantity and quality, an appropriate combination of labour with different skills is also very important in making optimum use of human resources. An excess of labour force of any kind works as a barrier rather than an impetus in economic growth. An important aspect of human resources is that both excess and scarcity of labour force are big constraints in the process of economic growth. The excess of labour force in India has proved a burden on the economy and a barrier to rapid economic growth, particularly the uneducated, untrained and unskilled manpower. According to NSSO's 6th round survey, 9.1 per cent of rural and 8.8 per cent of urban labour force was unemployed in 2004 in India (reported in *ES* 2005–2006, p. 208). Unemployed people consume without producing—it reduces the rate of savings and investment. On the other hand, scarcity of labour in oil rich middle-east countries constrained their real growth severely during the 1970s and 1980s. They had to depend on imported labour for all kinds of their manpower needs and they still do to a great extent.

As regards the quality of labour, it includes not only skill and productivity but also discipline, honest and sincere work efforts, commitment to productivity and professionalism. A highly trained and qualified labour will be much less productive than its potential if it lacks these qualities. That is what makes a difference between Indian and Japanese labour force, whatever might be the reason, historical, social, natural or otherwise.

29.3.2 Natural Resources

Natural resources of a country include the area of usable land, and resources on the land surface and underground. Land surface resources include sources of natural water (rivers and lakes), forests, landscape, etc. Underground resources include oil and natural gas and minerals. Favourable climatic and environmental conditions add to the natural resources endowments of a country. The countries with rich natural resource endowments have a much larger growth potential than those lacking natural resources.

However, natural resources are passive factors of growth. The exploitation and use of natural resources depends on the quality of manpower, availability of capital and technology.

The countries endowed with rich natural resources and a highly skilled and motivated manpower can do miracles in economic growth. These are the factors which may be said to have contributed to rapid growth of the United States, France, Germany, U.K., Canada and Australia. However, there are countries which are richly endowed with only one natural resource, i.e., oil. For example, Saudi Arabia, UAE and Kuwait, etc., are endowed with only one natural gift (oil) but they have the highest per capita income in the world. In contrast, there are tiny countries like Hong Kong, Singapore and Taiwan which have a small resource endowment, but a very high rate of economic growth. Japan is one of the most prominent examples of countries having achieved the highest growth rate during the post-World War

II period with its small area of land and few natural resources. The quality of manpower has played a vital role in the economic growth of these countries. On the other hand , large countries like India and China are still counted among the developing countries. These examples apart, countries rich in natural resources, and skilled manpower with high level of motivation and drive provide a stronger foundation for a high growth rate.

29.3.3 Capital Formation

Capital is defined as man-made means of production. In practual sense of the term, capital includes machinery, plant and building, means of transport and communication, electricity plants, and social overheads like roads, railways, schools, colleges, hospitals, etc. Creating or acquiring man-made means of production is known as *capital formation* or *capital accumulation*. Capital formation enhances the availability of capital per worker. A high capital/labour ratio enhances the productivity of labour. In other words. With a high rate of capital formation, a larger quantity of goods and services are produced per unit of time. This means a high growth rate.

Capital formation requires saving men and material resources from their use in consumer goods and transforming them into producer goods. In economic jargon, capital formation means sacrificing current consumption and saving incomes to be invested in capital goods (machinery, plant, building and equipment etc.). In general, the countries with a high rate of saving and investment have a higher rate of economic growth. Also, as the rate of saving and investment increases, the rate of economic growth increases too. For instance, during 1990s, the rate of gross capital formation in India was around 24 per cent and growth rate was 5 per cent. After 2004–05 gross capital formation crossed 30 per cent and growth rate shot up to 9 per cent.

29.3.4 Technological Development

Technology used in production is the fourth vital determinant of economic growth. Technology refers to scientific methods and techniques of production. In effect, technology means the nature and kind of machinery and technical equipments used with a given amount of labour. Capital-labour ratio is a broad measure of technology. Technological development means improving the technique of production through research and innovations. It results in a larger output from a given number of labour, materials and time. Invention of steam engines and railways, telephones and wireless, electricity, airplanes and computers are a few examples of technological developments over the past two centuries. Historical evidence shows that countries which achieved technological development at a rapid pace have made big strides in

the field of economic growth. Economically backward countries which are unable to make technological development on their own have to import technology from other countries. Countries using inefficient technology have evidently grown at a slower pace.

An important aspect of technological development or technological innovation is the choice of appropriate technology. Appropriate technology means a technology which is in accordance with the available resources and need of the country. The choice and use of inappropriate technology often leads to a high social cost of production. Choice of labour-saving technology by a labour-surplus economy leads to growth of unemployment. Under such conditions, not only are the advantages of technological development lost but also there are social and political tensions in the country due to growing unemployment and unequal distribution of incomes.

29.3.5 Social and Political Factors: The Fifth Factor

Social and political systems, organizations, institutions, social values, etc., also play an important role in the development process of an economy. Social factors like customs, traditions, beliefs, institutions, social (communal) harmony, and attitude towards the material life and well-being, determine, to a considerable extent, the pace of economic growth. A society of illiterate and ignorant people living with superstitions and unscientific beliefs resists modern ways of life and rational organization of society. Such a society finds it very difficult to achieve a high growth rate.

The form of government and its economic roles and policies matter to a great extent in determining the level and the rate of economic growth of a country. A government that plays a promotional role, provides adequate and efficient industrial infrastructure, builds an efficient system of public utilities (health and education system), invests in industries in which private investment is inadequate and removes weaknesses of the market system—helps economic growth. On the other hand, a government that throttles business and, thereby, activities through its restrictive economic policies, controls and regulations—as did the Indian government prior to the 1991 reforms through its licence, permit, quota *raj*—encourages inefficiency and malallocation of resources and restrains economic growth. Furthermore, political stability has always proved conducive to economic growth by encouraging industrial endeavours. An honest, sincere and efficient government builds public confidence, optimism and the right kind of attitude towards the society and the country, and commitment towards the nation and public welfare. In contrast, if the government is dishonest and inefficient, manned by corrupt and dishonest ministers, bureaucrats and government administrative infrastructure, as is the case at present in India—one of the most corrupt countries of the world—it promotes inefficiency even in the private business, increases cost of production, encourages inefficient allocation of resources, profiteering, black marketing and encourages malpractices in the private sector. All these hamper growth. More dangerously, it builds a corrupt system, a corrupt society and degrades social and human values. In this social environment, individuals care only for their private gains, not for the nation.

29.4 THEORIES OF ECONOMIC GROWTH

We have discussed above the factors that contribute to economic growth of a country. In this section, we discuss the theories of economic growth. Theories of economic growth bring out how economic factors contribute to the economic growth of a country.

The emergence of economic growth theories can be traced back to Adam Smith's *Wealth of Nations*.[3] With the passage of time, many other theories of economic growth were formuted, but not of great significance. It was during the 1930s and 1940s, that R.F. Harrod and E. Domar developed a path-breaking theory of economic growth—the capital accumulation growth theory—popularly known as the Harrod–Domar growth model. Since the mid-fifties, a number of other significant contributions[4] have been to the theory of economic growth. A collective name given to these contributions is *neoclassical theories of economic growth*. Of all the various theories of economic growth, we will discuss here only the two most important growth theories, viz., Harrod-Domar theory and the neo-classical theory of economic growth. The reason for omitting other theories[5] is that they do not have much relevance to modern economic growth. On the other hand, Harrod-Domar and neo-classical theories have a high analytical power and these theories explain more closely the modern growth behaviour.

29.4.1 Harrod-Domar Theory of Growth

Harrod-Domar growth model is an extension of Keynesian short-term analysis of full employment and income theory. It provides "a more comprehensive long period theory of output". The attention of economists was drawn towards the problem of a steady and sustained economic growth by the Great Depression of 1930s, and subsequently, by the economic devastation caused by the Second World War.[6] The central issue of the inquiry was to explore the requirements and conditions for steady growth in output and employment. Harrod and Domar had in their separate writings[7] examined and explained the conditions and requirements of steady economic growth. Although their models differ in details, their conclusions are substantially the same. Their models are, therefore, known as the Harrod-Domar model of growth.

Both Harrod and Domar consider *capital accumulation* as the key factor in the process of economic growth. They emphasize that capital accumulation (i.e . , net investment) plays a double role in economic growth. It generates income, on the one hand, and increases production capacity of the economy, on the other. For example, establishment of a new factory generates income for those who supply labour, bricks, steel, cement, machinery and equipment, etc., and at the same time, it increases the total stock of capital and thereby the production capacity of the economy. The newly generated income creates demand for goods and services. A necessary condition of economic growth is that the new demand (or spending) must be adequate enough

3. Smith, Adam, *An Enquiry into the Nature and Causes of Wealth of Nations* (1776), edited by Edwin Cannan (Random House, New York, 1937).

4. The most significant ones are : R.M. Solow's, 'A Contribution to the Theory of Economic Growth', *Quarterly Journal of Economics*, Feb. 1956; J.E. Meade's ' *A Neo-Classical Theory of Economic Growth*' (Oxford University Press, 1961); E.S. Phelps'. 'The New View of Investment: A Neo-Classical Analysis', *Quarterly Journal of Economics*, Nov., 1962; and H.G. Johnson's 'The Neo-Classical One-Sector Model: A Geometrical Exposition and Extension to a Monetary Economy', *Economica*, Aug., 1966.

5. A brief exposition of classical and other growth theories is given by Hahn, H.F. and R.C.O. Mathews, “The Theory of Economic Growth: A Survey”, *Economic Journal,* December 1964. For a detailed discussion, see Benjamin Higgins, *Economic Development* (Central Book Depot, Allahabad, 1961).

6. Kindleberger. C.P. and B., Herrick, *Economic Development* (McGraw-Hill Kogakusha, Ltd., 1977), International Students Edition, Ch. 3, p. 45.

7. Roy, Harrod, “An Essay in Dynamic Theory” in *Economic Journal*, March 1939 and Evsey D. Domar, “Expansion and Employment”, *Am. Eco. Rev.* , March 1947.

to absorb the output generated by the new investment. Otherwise, there will be excess or idle production capacity. This condition should be fulfilled year after year in order to maintain full employment and to achieve steady economic growth in the long-run. This is the central theme of this model. The model itself is formally described below.

29.4.2 Assumptions of the Harrod-Domar Growth Model

The Harrod-Domar growth model is based on the following two basic assumptions.

(1) ***Capital-Output Ratio Remains Constant.*** Harrod-Domar model assumes a constant capital-output ratio. It implies a production function with constant capital-output co-efficient. In simple words, the Harrod-Domar model assumes that national income is proportional to the stock of capital, i.e.,

$$Y = kK, \ (k > 0) \quad \text{...(29.1)}$$

where Y = national output; K = total stock of capital; and k = capital output ratio.

Since capital/output ratio is assumed to remain constant, any increase in national output (ΔY) must be equal to k times ΔK, i.e.,

$$\Delta Y = k\Delta K \quad \text{... (29.2)}$$

It follows from Eq. (29.2) that the growth in national output per time-unit depends on and is limited by the growth of capital stock per time-unit. If an economy is in equilibrium and the existing stock of capital is fully employed, then capital/output ratio (k) can be easily worked out. Once k is known, then additional capital required to produce a given additional output can also be easily worked out by using Eq. (29.2).

Since increase in capital stock (ΔK) in any period equals the net investment (I) of the period, Eq. (29.2) may be rewritten as

$$\Delta Y = kI \quad \text{...(29.3)}$$

(2) ***Saving-Income Ratio Remains Constant.*** Another important assumption of the Harrod-Domar model is that society saves a constant proportion of the national income, i.e., total saving (S) is a function of income Y, and the saving function can be written as

$$S = sY, \ (s > 0) \quad \text{...(29.4)}$$

where S = saving per unit of time, and s = constant propensity to save.

At equilibrium level of output, the desired savings must equal the desired investment, i.e., at equilibrium,

$$S = I = sY \quad \text{...(29.5)}$$

29.4.3 Harrod-Domar Growth Model

Given the assumptions, the growth rate, defined as $\Delta Y/Y$, may be obtained as follows[8]. It may be inferred from Eq. (29.3) that in period ‘ t’,

$$\Delta Y_t = kI_t \quad \text{...(29.6)}$$

Here, ΔY_t equals income in period t *minus* income in period, t–1. That is,

$$\Delta Y_t = Y_t - Y_{t-1}$$

8. $\Delta Y/Y = (Y_t - Y_{t-1})/Y_{t-1}$ where Y_{t-1} = national output in period t–1; Y_t = national output in period t (i.e., one time-period later than period t – 1).

By substitution, Eq. (29.6) may be written as

$$Y_t - Y_{t-1} = kI_t \quad \text{...(29.7)}$$

Eq. (29.5) says that the Harrod–Domar model assumes that at equilibrium in period t, $I_t = S_t = sY_t$. Since $I_t = sY_t$, by substituting sY_t for I_t in Eq. (29.7), we get

$$Y_t - Y_{t-1} = k \times sY_t \quad ...(29.8)$$

By dividing Eq. (29.8) by Y_t, we get growth rate ($\Delta Y/Y$) or what Harrod calls, 'the warranted growth rate' (G_w) as

$$G_w = \frac{Y_t - Y_{t-1}}{Y_t} = k \times s$$

or

$$G_w = \frac{\Delta Y}{Y_t} = k \times s \quad ...(29.9)$$

where G_w is the 'warranted growth rate.'

As Eq. (29.9) shows, the rate of growth equals the capital output ratio (k) times the constant propensity to save.9 Since growth rate, $\Delta Y/Y$, pertains to condition $I = S$, this may also be called *equilibrium growth rate*, which implies capacity utilization of capital stock. This growth rate fulfils the expectations of the entrepreneurs. Harrod calls this growth rate as *'warranted growth' rate,* (G_w). Harrod defines G_w as "that rate of growth which, if it occurs, will leave all parties satisfied that they have produced neither more nor less than the right amount".

According to the Harrod-Domar model, a *target growth rate* can be achieved either by increasing marginal propensity to save and simultaneously increasing the investment or by increasing the capital/output ratio. When marginal propensity to save increases, overall savings increase. Savings transmuted into investment increase income and production capacity of the nation. Increase in income leads to increase in demand for goods so that additional output generated through additional investment is absorbed. On the other hand, increase in production capacity in a period creates more income in the following periods. Higher incomes lead to higher savings and investment and still higher incomes in the subsequent periods. In the process, the investment increases at an accelerated rate, based on the *principle of acceleration*. 10

29.4.4 Shortcomings of the Model

The economists have pointed out the following shortcomings in Harrod-Domar growth model.

(1) ***Harrod-Domar Model is based on Unrealistic Assumptions.*** The Harrod-Domar proposition that savings will always increase to match with the investment need is based on the assumption that *warranted growth rate* (G_w) is always equal to the *actual growth rate* (G_r), i.e., expected growth rate is always realized. This is possible only under the following simplifying assumptions of the model:

(*i*) Marginal propensity to consume (*mpc*) remains constant;

(*ii*) Output/capital ratio is constant;

(*iii*) The technology of production is given;

9. Assumption 2 implies that marginal propensities to save is equal to average propensity to same.

10. The Principle of Acceleration is discussed ahead in Ch. 30.

(*iv*) The economy is initially in equilibrium;

(*v*) There is no government expenditure and no foreign trade, and

(*vi*) There are no lags in adjustments between demand and supply, and between saving and investment.

These assumptions make the model-economy unrealistic because the warranted (or expected) growth rate may not be always equal to the actual (realized) growth rate. And, if warranted and actual growth rates are not equal, it will lead to economic instability, not to stable growth as claimed by the model.

(2) ***Harrod-Domar Model is a Razor-Edge Model.*** Another major defect of the

Harrod-Domar model is that the parameters used in it, *viz.* , capital/output ratio, marginal propensity to save, growth rate of labour force, progress of labour-saving technology, are all determined independently out of the model. The model, therefore, does not ensure the equilibrium growth rate in the long-run. Even the slightest change in the parameters will make the economy deviate from the path of equilibrium. That is why this model is sometimes called a 'razor-edge' model.

In spite of these limitations, the Harrod-Domar model is widely used in growth planning of the countries. This model has been used in India's development planning. Besides, it has great analytical power.

29.4.5 The Neo-Classical Theory of Growth

As mentioned earlier, the contributions made to the growth theory by economists[11] including Tobin, Solow, Swan, Meade, Phelps and Johnson have been given a collective name as "The Neo-classical growth theory." The approach adopted by these growth theorists in their models is based on the assumptions usually made by the neo-classical economists, viz., Marshall, Wicksell and Pigou. The assumptions are following:

(*a*) There is perfect competition in commodity and factor markets;

(*b*) Factor payments are always equal to their marginal revenue productivity;

(*c*) Capital/output ratio is variable, and

(*d*) There is full employment.

Since growth theories of Tobin, Solow, *et al*, are based on neo-classical assumptions, these theories are called *neo-classical growth theories*.

There are, however, some important differences between the assumptions of the Harrod-Domar model and neo-classical growth model.

First, while the production function implicit in the Harrod-Domar growth model contains only one factor, i.e., capital, the neo-classical growth model assumes a multi-factor production function including capital (K), labour (L) and technology (T).

11. Tobin, James, 'A Dynamic Aggregate Model', *Journal of Political Economy*', 63, April 1955; Robert M. Solow, 'A Contribution to the Theory of Economic Growth,' *Quarterly Journal of Economics*, 7 February 1956; T.W. Swan, 'Economic Growth and Capital Accumulation,' *Economic Record*, Nov. 1956; J.E., Meade, *A Neo-Classical Theory of Economic Growth*, Oxford University Press, 1961; E.S. Phelps, 'The New View of Investment : A Neo-Classical Analysis', *Quarterly Journal of Economics*, Nov. 1962; H.G. Johnson, 'The Neo-Classical One-sector Model : A Geometrical Exposition and Extension to a Monetary Economy, *Economica*, Aug. 1966.

Secondly, in the Harrod-Domar model, labour and capital are deemed to be perfect *complements* of one another whereas in the neo-classical model, capital and labour are assumed to be *substitutes* for one another.

Thirdly, while Harrod-Domar model assumes a constant capital/output ratio, the neo-classical model assumes a variable capital/output coefficient (ratio). Incidentally, both the models assume that capital and labour are subject to the law of diminishing marginal returns to scale.

Finally, in addition, the neo-classical model assumes perfect competition in input market where factor prices equal their marginal revenue productivity.

The Neo-Classical Growth Model

According to the neo-classical model, rate of economic growth depends on the growth rate of (*i*) capital stock, K; (*ii*) labour supply, L; and (*iii*) technological progress over time, T. The relationship between the national output and these variables may be expressed in the form of a production function, i.e.,

$$Y = F(K, L, T) \quad ...(29.10)$$

where Y = national output (at constant price), K = stock of capital, L = labour supply, and T = the scale of technological progress.

If technology is assumed to remain constant, then the growth rate depends on K and L. The production function then takes the following form.

$$Y = F(K, L) \quad ...(29.11)$$

Given the assumption of constant returns to scale, the increase in national output (ΔY) would be equal to marginal productivity (MP) times ΔK and ΔL. Thus,

$$\Delta Y = \Delta K \,.\, MPk + \Delta L \,.\, MPl$$

...(29.12)

where *MPk* and *MPl* denote marginal physical product of capital (*K*) and labour (*L*) respectively.

By dividing both sides of Eq. (29.12) by *Y*, we get the growth rate ($\Delta Y/Y$) as follows.

$$\frac{\Delta Y}{Y} = \Delta K\left(\frac{MP_k}{Y}\right) + \Delta L\left(\frac{MP_l}{Y}\right) \qquad ...(29.13)$$

If we multiply the first term on the *RHS* of Eq. (29.13) by K/K and the second term by L/L, it will yield a useful ratio without altering the equation. By multiplying, we get

$$\frac{\Delta Y}{Y} = \Delta K\left(\frac{MP_k}{Y}\right)\frac{K}{K} + \Delta L\left(\frac{MP_L}{Y}\right)\frac{L}{L}$$

By rearrainging the terms, we get

$$\frac{\Delta Y}{Y} = \frac{\Delta K}{K}\left(\frac{K \cdot MP_k}{Y}\right) + \frac{\Delta L}{L}\left(\frac{L \cdot MP_l}{Y}\right)$$

⎝

⎠

...(29.14)

The neo-classical model assumes perfect competition in which factor prices are equal to their marginal physical product times the product price (i.e . , marginal revenue productivity). Under this condition, *MPk* × *K* and *MP l* × *L*, denote the total share of capital

(*K*) and labour (*L*) in the national output (*Y*), and *MPk* × *K*/ *Y* and *MPl* × *L*/ *Y* denote the relative share of *K* and *L* in *Y*. Thus,

k

$MP\,.\,K$

l

$MP\,.\,L$

$+$

$= 1$

...(2.15)

Y

Y

l

$MP\,.\,L$

If

k

$MP\,.\,K$ in Eq. (29.15) is assumed to be equal to *b*, then

$= 1 - b$.

Y

Y

By substitution, Eq. (29.14) can now be rewritten as

ΔY

$\left(\Delta K\right.$

$= |$

$| + (1-)\ \Delta L$

b

b

...(29.16)

Y

$\left.K\right)$

L

Equation (29.16) reveals the basic proposition of neo-classical growth theory. In this equation, *b* and (1 – *b*) show the responsiveness of output to the changes in *K* and *L*, respectively. That is, *b* denotes the elasticity12 of output with respect to change in capital stock when labour remains constant and (1 – *b*) denotes the elasticity of output with respect to change in labour, stock of capital remaining constant. Thus, the neo-classical model suggests that the rate of economic growth equals the elasticity of output with respect to increase in capital stock, plus elasticity of output with respect to increase in labour force, given the level of technology.

Neo-Classical Model with Technology Change Let us now reintroduce technological progress in the neo-classical model and designate the growth rate of output associated with technological progress as Δ *T*/ *T*. Now Eq. (29.16) may be written as

ΔY

$\left(\Delta K\right.$

ΔL

Δ

$=$

T

$b|$

$| + (1 - b)$

$+$

...(29.17)

Y

$\left(K \right)$

L

T

After the introduction of technological progress in the model, the growth rate presented in Eq. (29.17) increases by the proportional increase in the output owing to technological progress. Thus, the overall growth rate equals the elasticity of output (*Y*) with respect to capital expansion plus elasticity of output (*Y*) with respect to labour plus growth rate of output as a result of technological progress.

To conclude, the neo-classical model, like the Harrod-Domar model, is based on certain simplifying assumptions. The assumptions, no doubt, add analytical simplicity to the model, but they also lead to a very high degree of abstraction from the real problems of growth, particularly in respect of small and poor countries.[13]

Shortcomings of the Neo-Classical Growth Model

Prof. A.K. Sen[14] has pointed out some major problems of growth theories which are not resolved by the neo-classical model.

12. The elasticity of output with respect to an input is the proportional change in output associated with 1 per cent change in the (chosen) input, other inputs remaining constant.

13. Kindleberger, C.P. and B. Herrick, *op. cit.*, p. 54.

14. A.K. Sen, *Growth Economics*, (Penguin Books), pp. 21–28.

First, "for steady state equilibrium, it is necessary to assume that technological progress is entirely augmenting", i.e., neutrality of technological progress. There is, however, little empirial justification for this assumption.

Secondly, the adjustment mechanism envisaged by the neo-classical model rests on flexibility of factor prices. But the adjustability of factor prices, e.g., interest rates, may be prevented by 'liquidity trap', which does not allow the interest rate to go below a certain level, and "this may prevent the capital output ratio from being as high as may be necessary for growth equilibrium."

Thirdly, neo-classical model does not include the investment function and hence fails to explain the expectations of the entrepreneurial class and its role in capital accumulation. Therefore, if an independent investment function is introduced to the neo-classical model, instability problem will appear in the model.

Finally, neo-classical model assumes homogeneity of capital assets, which is unrealistic and misleading.

It may thus be said at the end that economic growth problems are not as simple as presented in the growth models.

SUMMARY

- Economic growth means increase in national income at a reasonably high rate persistently over a period of time. Economic growth provides a good business environment.
- Economic growth of a country depends on the availability and use of national resources. National resources include (*i*) human resources, the manpower, (*ii*) natural resources, and (*iii*) capability of a country to use the resources.
- Practically, economic growth of a country is determined by (*i*) supply and quality of manpower, (*ii*) availability and use of natural resources, (*iii*) capital formation, (*iv*) technology, and (*v*) political and social environment.
- Economists have constructed various theories of economic growth explaining the relationship between economic growth and determinants of economic growth. However, two most important growth theories relevant to modern economic conditions are (*i*) Harrod-Domar theory of growth, and (*ii*) Neo-classical theory of growth.
- According to Harrod-Domar growth theory, given the capital-output ratio and rate of savings, growth rate of a country depends on (*i*) the level of investment, and (*ii*) capital-output ratio *(k)*. Given the rate of savings (*s*), growth rate is measured at $k \times s$.
- According to the Neo-Classical theory of growth, economic growth of a country depends on three factors: (*i*) capital stock *(K)*, (*ii*) labour supply (*L*), and (*iii*) technology (*T*). Given the technology (*T*), growth rate depends on (a) increase in *K* and marginal productivity of capital (*MP* k), and (b) increase in *L* and marginal productivity of

labour *(MP* l *)*. Growth rate can be measured as

$\Delta Y/Y = \Delta K \cdot MP_k + \Delta L \cdot MP_t$

REVIEW QUESTIONS

1. What is meant by economic growth? How is economic growth measured?
2. What are the determinants of economic growth? How do the social and political factors affect economic growth rate?
3. What are the assumptions of the Harrod-Domar growth model? Briefly discuss the model and point out its shortcomings.
4. Discuss elaborately the Harrod-Domar growth model. What is meant by warranted growth rate?
5. Why is Harrod-Domar model of growth called 'razor-edge model'?
6. Discuss neo-classical theory of growth. How does this growth model incorporate the effect of technological change?
7. How is the neo-classical growth theory different from the Harrod-Domar growth theory?
8. What do theories of economic growth reveal? What can be the possible use of growth theories in business decisions?
9. Explain the neo-classical theory of economic growth and point out its limitations.
10. Explain the following concepts:

(*a*) Capital-output ratio,

(*b*) Marginal physical productivity,

(*c*) Marginal revenue productivity and

(*d*) Elasticity of output.

11. What are the remarks of Prof. A.K. Sen on the neo-classical growth theory?
12. Write a note on the relevance of the growth models to the real economic world.

FURTHER READING

Baumol, W.J., *Economic Dynamics: An Introduction* (The Macmillan Company, New York, 2nd Edn., 1959).

BenJamin, Higgins, *Economic Development* (Central Book Depot, Allahabad, 1961).

Domar, e., "Capital Expansion, Rate of Growth and Employment." *Econometrica*, Vol. 14, 1946.

HarroD, r.F., "An Essay in Dynamic Theory", *Economic Journal*, Vol. 49, 1939.

HarroD, r.F., *Towards a Dynamic Economics* (Macmillan, 1948).

KinDleBerger, C.P. and B. HerriCK, *Economic Development* (McGraw Hill Kogakusha, International Edn., 1977).

meaDe, J.e., *Neo-Classical Theory of Economic Growth* (Oxford University Press, 1961).

sen, a.K., (ed.), *Growth Economics* (Penguin Books, England, 1970) 'Introduction' by the Editor.

soloW, r.m., "A Contribution to the Theory of Economic Growth," *Quarterly Journal of Economics,* 7 Feb., 1955.

sWan, T.W., "Economic Growth and Capital Accumulation," *Economic Record*, November 1956.

ToBin James, "A Dynamic Aggregate Model," *Journal of Political Economy*, 63, April, 1955.

CHAPTER 30 Business Cycles and Control Measures

CHAPTER OBJECTIVES

Business cycles, i.e., sequential economic growth and depression, change business environment and affect business prospects – growth affects it favourably and depression adversely. By going through this chapter, you will know the following aspects of business cycles.

- Meaning and phases of business cycles
- Theories of business cycles applicable to modern conditions
- Policy measures to control business cycles and
- Efficacy of policy measures

30.1 INTRODUCTION

In the preceding chapter, we have discussed modern theories of economic growth which attempt to explain the rela tionship between the national output and growth factors, and specify the conditions under which an economy would steadily grow on the equilibrium path. The historical evidence, however, shows that economies have not grown persistently over a long period of time. Most economies of the world have experienced business cycles at different stages of their economic growth. The economic history of various economies is, in fact, a history of ups and downs, booms and slumps, prosperity and depression. Briefly speaking, business cycles have characterized the free enterprise industrial world over the past one and a half century. Although, the Great Depression of 1930s has not repeated itself even after a period of 65 years, the global recession of 2007–10 is considered to be the second worst economic recession of the global economy.

The global inflation and world economic recession of the 1980s and the second worst economic recession of 2007–10 give a clear warning against any complacency towards the dangers of economic fluctuations. The corrective measures adopted by the government to control economic fluctua tions may prove inadequate and inefficient when destabilizing forces are deep-rooted. To quote E. Burns, "... men who wish to serve democracy faithfully must recognize that the roots of business cycles go deep in the economic organization, that the ability of government to control depressions adequately is not yet assured, that

our power of forecasting is limited and true foresight requires policies for coping with numerous contingencies."[1]

Business cycles, i.e., repeated phases of economic booms and slumps are essentially a perpetual feature of the economic environment of a country. Business cycles influence business prospects tremendously and set the trend for future business. The period of prosperity opens up new and larger oppor tunities for investment, employment and production, and thereby promotes business. On the contrary, the period of depression reduces business opportunities. A profit maxi mizing entrepreneur must, therefore, analyze and take in view the economic environ ment of the period prior to making decisions, particularly those pertaining to forward planning.

This chapter is devoted to a brief discussion[2] on the phases and features of business cycles, important business cycle theories, causes of business cycles, and economic policies that aim at stabilization of the economy.

30.2 PHASES OF BUSINESS CYCLES

Business cycles refer to the periodic booms and slumps in eco nomic activities, generally compared to 'ebb and flow'. The ups and downs in the economy are reflected by the fluctuations in aggregate economic magnitudes, including total production, investment, employment, prices, wages, bank credits, etc. The upward and downward movements in these magnitudes show different phases of business cycles. Basically, there are only two phases in a cycle, *viz.* , *prosperity* and *depression*. However, considering the inter-mediate stages between prosperity and depression, the various phases of trade cycle are enumerated as follows:

1. Expansion of economic activities,
2. Peak of boom or prosperity,
3. Recession, the downtrend,
4. Trough, the bottom of depression, and
5. Recovery and expansion.

The five phases of the business cycle are presented in Fig. 30.1. The *steady growth line* shows the growth of the economy when there are no business cycles. The various phases of business cycles are shown by the *line of cycle*

Fig. 30.1 *Phases of Business Cycle*

which moves up and down

1. Quoted from Gordon, R.A., *Business Fluctuations* (Harper and Brothers, Publishers, N.Y., 1952), p. 4.

2. For detailed discussion, see Maurice W. Lee, *Economic Fluctua tions* (Richard D. Irwin Inc. Illinois, 1955), Ch. 3; D. Hamberg, *Business Cycles* (The Macmillan Companies, New York, 1951), Ch. 1; R.A. Gordon, *Business Fluctuations* (Harper and Brothers, Publishers, New York, 1952), Ch. 8.

the *steady growth line*. The line of cycle moving above the steady growth line marks the beginning of the period of 'expansion' or prosperity' in the economy. The *expansion phase* is characterized by increase in output, employment, investment, aggregate demand, sales, profits, bank credits, wholesale and retail prices, per capita output and a rise in standard of living. However, growth rate eventually slows down and reaches its peak. The phase of recession begins when the downward slide in the growth rate becomes rapid and steady. Output, em ployment, prices, etc., register a rapid decline, though the realized growth rate may still remain above the steady growth line. So long as growth rate exceeds or equals the expected steady growth rate, the economy enjoys the period of *prosperity*—high or low. But, when the growth rate falls below the steady growth rate, it marks the beginning of depression in the economy. When depression continues and hits the bottom, it marks a period of *trough*. After the period of trough, the economy begins to recuperate and treads on the path of recovery. The process is continuous.

Let us now describe in some detail the important features of the various phases of business cycle, and also the causes of turning points.

Prosperity: Expansion and Peak The prosperity phase is characterized by a rise in the nation al output, rise in consumer and capital expenditure, rise in the prices of raw materials and finished goods, and rise in the level of employment. In the later stages of prosperity, however, inputs start falling short of their demand. Additional workers are hard to find. Hence additional workers can be obtained by bidding a wage rate higher than the prevailing rates. Labour market becomes a seller's market. A similar situation also appears in other input markets. Consequently, input prices increase rapidly leading to increase in output and employment. Cost of living increases at a rate relatively higher than the increase in household incomes. Hence consumers, particularly wage earners and households of fixed income class, review their consumption expenditure. Consu mers' resistance gets momentum. Actual demand stagnates or even decreases. The first and most serious impact of decline falls on the demand for new houses, flats and apartments. Incidently, this is what happened in the US in 2007–08 causing economic depression. Following this, demand for cement, iron and steel, and construction labour tends to halt. This trend subsequently appears in other durable goods industries like automobiles, refrigerators, furniture, etc. This marks reaching the *peak*.

Turning-Point and Recession As already mentioned, once the economy reaches the peak, increase in demand is halted. The demand even starts decreasing in some sectors, for the reason stated above. Producers, on the other hand, unaware of this fact continue to maintain their existing levels of production and investment. As a result, a discrepancy arises between output supply and demand: supply exceeds demand. The widening of discrepancy between supply and demand is so slow that it goes unnoticed for some time. But, the persistence of this problem makes the producers believe that they have indulged in 'over-investment' and over-production. Consequently, future investment plans are given up; orders placed for new equipments, raw materials and other inputs are cancelled. Replacement of worn-out capital is postponed. Demand for labour tends to decrease; and temporary and casual workers are laid off in a bid to bring demand and supply in balance. The cancellation of orders for inputs by the producers of consumer goods creates a chain-reaction in the input market. Producers of capital goods and raw materials cancel their orders for their input. This is the *turning point* and *the begin ning of recession*.

Since demand for inputs has decreased, input prices, e.g., wages, interest, etc., show a gradual decline leading to a simul taneous decrease in the incomes of wage and interest earners. This ultimately causes *demand recession*. On the other hand, producers lower their price in order to get rid of their inventories and also to meet their financial obligations. Consumers, is their turn, expect a further decrease in price, and hence, postpone their purchases. As a result, the discrepancy between demand and supply continues to grow. When this process gathers speed, it takes the form of irreversible *recession*. Investments start declining. The decline in investment leads to decline in employment, income and consumption. When investments continue to decline, production and employment continue to decline causing in further decline in demand for both consumer and capital goods. Borrowings for investment decreases; bank credit shrinks; stock prices de crease; unemployment increases even though there is a fall in wage rates. At this stage, the process of recession is complete. When growth rate falls below the steady growth line, the economy enters the phase of *depression*.

Depression and Trough During the phase of *depression*, economic activities slide down their normal level. The growth rate becomes negative. The level of national income and expenditure declines rapidly. Prices of consumer and capital goods decline steadily. Workers lose their jobs. Debtors find it difficult to pay off their debts. Demand for bank credit reaches a low ebb and banks experience mounting of their cash balances. Investment

in stock becomes less profitable and less attractive. At the depth of *depression*, all economic activities touch the bottom and the phase of *trough* is reached. Even the expenditure on maintenance is deferred in view of excess production capacity. Weaker firms are eliminated from the industries. At this point, the process of *depression* is complete.

How is the Process Reversed? The basic factor that reverses the recessionary trend is the limit to which an economy can shrink. When the econo my hits the bottom and stays there for some time, it marks the end of pessimism and beginning of optimism. This reverses the process. The process of reversal generally begins in the labour market. The widespread unemployment forces workers to work at wages less than the prevailing rates. The producers anticipating better future try to maintain their capital stock and offer jobs to some workers here and there. They do so also because they begin to take an optimistic view of the situation due to the halt in decrease in price in the trough phase. Consumers on their part expecting no further decline in price begin to resume their postponed consumption and hence demand picks up, though gradual ly. Bankers having accumulated excess liquidity (idle cash re serve) try to salvage their financial position by lowering the lending rate and by investing their funds in securities and bonds, even if rate of return is very low. Consequently, investment picks up and employment gradually increases. Following this recovery in production and wage income, demand for both consumer and capital goods starts picking up. Since banks have accumulated excess cash reser ves, bank credit becomes easily available and at a lower rate. For all these reasons economic activities get accelerated. Due to increase in income and consumption, the process of multiplier gives further impetus to the economic activities, and the phase of *recovery* gets underway, depending on the speed of recovery.

The Phase of Recovery As the recovery gathers momentum, some firms plan addi tional investment; some undertake renovation programmes; and some undertake both. These activities generate construction activities in both consumer and capital goods sectors. Individuals who had postponed their plans to construct houses undertake this task now, lest

cost of construction should mount. As a result, more and more em ployment is generated in the construction sector. As employment increases despite wage rates moving upward, the total wage incomes increase at a rate higher than the employment rate. Wage income rises and so does the consumption expenditure. Businessmen realizing a quick return with high profitability, speed up the production machinery.

Over a period, as factors of production become more fully employed, wages and other input prices move upward rapidly, though not uniformly. Investors, therefore, select the best of the alternative investment opportunities. As prices, wages and other factor prices increase, a number of related developments begin to take place. Businessmen start increasing their inventories, consumers start buying more and more of durable goods and variety items. With this process catching up, the economy enters the phase of expansion and prosperity. The cycle is thus complete.

30.3 THEORIES OF BUSINESS CYCLE

A number of theories have been offered to explain business cycles. A systematic study of business cycles, however, is a matter of rather recent development. Most of the important contribu tions to the theory of business cycle were made in the first half of the twentieth century though business cycles had taken place throughout the nineteenth century. Between 1890 and World War I, a number of important contributions were made to the trade cycle theory[3]. Although many important contributions were made to the theory of business cycle[4] prior to the Great Depression, the study of business cycle still remained outside the general economic theory. It was Keynes,[5] who provided a general theoretical frame work in which the theory of business cycle could be interwoven. In his *General Theory* he provided standard tools for analyzing the economic fluctuations though he himself had said little about the causes of cycli cal fluctuations. Hicks[6] has remarked that Keynesian economics has done all for understanding of business fluctuations but has left out the analysis of

business cycle itself. In the post-Keynesian era, the main contributors7 to the cycle theory
include Metzler, Harrod, Kalecki, Samuleson, Kaldor, Hicks, Goodwin and Duesenberry.

The trade cycle theories are generally classified under the following categories:

1. The Pure Monetary Theory,
2. The Monetary Over-investment Theory,
3. The Non-monetary Over-investment Theory,
4. Innovation Theory,
5. Acceleration Principle of Trade Cycle,
6. Psychological Theory,

3. The only important contribution prior to this period was made by a non-economist, Clement Juglar in 1860.

4. The important contributions were made by M. Tugan Baranowsky of Russia, Aurther Spiethoff and J.A. Schumpeter of Germany, Knut Wicksell of Sweden, D.H. Robertson and R.G. Hawtrey of England, Albert Aftalion and Jean Lescure of France, Thorstein Veblen and W.C. Mitchell of the United States. (For details, see Gordon, R.A., *op. cit.* , p. 306).

5. *The General Theory of Employment, Interest and Money*, 1936.

6. *A Contribution to the Theory of Trade Cycle* (Oxford University Press, London, 1950), p. 1.

7. See also Clark, J.J., and M. Cohen, (eds.), *Business Fluctua tions, Growth and Economic Stabilization* (Random House, 1963), Bibliography.

7. Under-consumption Theories,
8. Exogenous Forces Theories,
9. Mitchell’s Theory of Cycle,
10. Theories of Keyensian System, and
11. Modern Theories of Trade Cycle based on interaction of the multiplier and the accelerator.

A detailed discussion on all these theories of trade cycle falls outside the purview of this book. Therefore, we will discuss here only some selected theories, for the simple reason that they have a relatively greater relevance to modern business conditions. The theories that we will discuss include the following.

1. Pure Monetary Theory,
2. Monetary Over-investment Thoery,
3. Schumpeter’s Innovation Theory,
4. Multiplier-Accelerator Interaction Theory, and
5. Hicks’ Theory of Trade Cycle.

30.3.1 Pure Monetary Theory

The early business cycle theorists lay major emphasis on the monetary and credit system in their analysis of business cycles. Their theories of business cycle are, therefore, jointly known as *monetary theory of business cycle*. According to this theory, the main cause of business cycle is the fluctuation in monetary and credit markets. The fluctuation in the supply of money and bank credit is the basic causal factor at work in the cyclical pro cess. Hawtrey, the main proponent of this theory, maintained that business cycles are nothing but successive phases of infla tion and deflation. According to him, all changes in the levels of economic activities are only reflections of changes in money flows. When money supply expands, prices rise, profits increase, and the total output increases, and, finally, growth takes place.

According to Hawtrey, the principal factor affecting the money supply is the credit mechanism. In modern economies, the principal source of money supply is the volume of credit created by the banking system. The upswing of the cycle begins with an expansion of bank credit and continues as long as credit expan sion continues. Banks expand credit facility because conditions are such that banks find it profitable to offer credit on easier terms (i.e., at a relatively lower interest rate). This is the most effective inducement for entrepreneurs to undertake productive activities. Consequently, bank credits flow into different types of capital formation activities including both widening and deepening of capital. Thus, credit expansion accelerates the process of economic expansion, on the one hand, and helps rise in prices, on the other.

The process of prosperity thus brought about by the banking credit mechanism is reversed when banks begin to restrain their credit expansion. The banks may find it difficult to expand the credit further at the prevailing rate because their cash reserves stand depleted due to (*i*) increase in loans and advanc es; (*ii*) reduced inflow of deposits, and (*iii*) withdrawal of deposits for quicker returns and more profitable uses. As credit expansion comes to an end, businessmen can no longer obtain bank credit for furthering their business activities. Therefore, the process of expansion is slowed down. Due to lack

of credit, businessmen find it difficult to meet their payment obligations and to maintain inventories at the existing level. So they allow their inventories to de plete. Besides, they cancel their orders for the purchases of inputs. This marks the beginning of downswing.

Evaluation of Pure Monetary Theory The pure monetary theory remained in vogue for quite some time and it still retains its relevance in explaining modern business fluctuations. It has, however, been criticized on the following grounds.

First, economists argue that business cycles are not purely a monetary phenomenon. Economic activities have also fluctuated because of change in non-monetary factors like aggregate demand, demand for new investments, cost structure, expectations of businessmen, etc.

Secondly, although monetory factors play an important role in the cumulative process of expansion and contra ction, they do not fully explain the turning points. At turning points, non-monetary factors have been found to have played a major role.

Thirdly, monetary theorists' conviction that businessmen are highly sensitive to the changes in interest rates is highly doubtful. A more important factor in business decisions are future business prospects and the marginal efficiency of capital.

In spite of these shortcomings, the pure monetary theory of business cycles has been regraded as a sound reasoning and logi cal explanation of economic fluctuations.

30.3.2 Monetary Over-investment Theory

The monetary over-investment theory emphasizes the role of *imbalance between the desired* and *actual investments* in economic fluctuation, i.e., actual investment exceeding the desired investment. F.A. Hayek, the pioneer of this theory, stresses that to keep the economy in equilibrium, investment pattern should correspond to the pattern of consumption and *for the economy to remain in stable equilibrium, it is necessary that voluntary savings are equal to the actual investment.* The total investment is so distribut ed between the various industries that each industry produces only as much as is demanded by the consumers, i.e., in case of each industry, supply is equal to demand. Given these equilibrium conditions, there will be no tendency to expand consumption and the entire economy would remain in the state of stable equilibrium.

The equilibrium and stability of the economy would be upset by changes in the money supply and saving-investment relations. The saving-investment relations may change due to increase in investment without corresponding increase in voluntary savings. Saving remaining constant, investment may increase due to such reasons as increase in mar ginal efficiency of capital , fall in the rate of interest, over-optimism about the future prospects, etc. If new investments are financed through increased bank credit, this leads to *over-invest ment*, mainly in the capital goods industries. Thus, there may be expansion of investment without contraction in consumption, and increase in savings. New investments generate additional income and, thereby, additional consumer demands. Since there is a time-lag between demand and supply, excess demand causes inflation.

The rise in consumer goods prices overtakes the rise in capital goods pric es. Therefore, profitability in consumer goods industries becomes higher than that of capital goods industries. This causes a shift in investment from the consumer goods to capital goods industries. As a result, demand for bank credit increases in the consumer goods industries.

But due to bankers' unwillingness and inability to meet the credit de mand, more so when there is a competitive demand for funds from capital goods industries, a financial crisis develops. This leads to a sharp decline in the capital goods production because of (*i*) fall in investment under the pressure of rising cost, and (*ii*) fall in marginal efficiency of capital. This causes unemployment in capital goods industries. This unemployment created in the capital goods sector is much too rapid to be absorbed in the consumer goods sector. Consequently, large-scale unemployment results; income stream is broken; and a downswing begins in general business activities leading to a depression.

Critical Evaluation The main shortcomings of the over-investment theory are following.

First, the theory presumes that when market rate of interest is lower than the natural rate—the rate at which demand for funds is equal to the supply of funds—the new bank credit flows to the capital goods industries. This would apply only under the condition of full employment. But, business cycles have taken place even when resources were not fully employed.

Secondly, the monetary over-investment theory stresses upon the change in the inter est rate as the main determinant of investment. It ignores many other important factors such as businessmen's own expectations, cost of capital equipments, etc.

Finally, the monetary over-investment theory lays undue emphasis on the imbalance between the investment in capital goods and consumer goods industries. In a modern economy, such imbal ances are self-correcting and do not create serious depressions.

30.3.3 Schumpeter's Theory of Innovation

According to Schumpeter, "business cycles are almost exclu sively the result of innovations in the industrial and commercial organizations." By innovations he means "... such changes of the combination of the factors of production as cannot be effected by infinitesimal steps or variations on the margin. [Innovation] consists primarily in changes in methods of production and trans portation, or changes in industrial organization, or in the production of a new article, or opening of a new market or of new sources of material...". Innovations do not mean inventions. Innovations are simply the commercial application of new techniques, new materials, new means of transportation and new sources of energy.

According to Schumpeterian theory, innovations are the cause of cyclical fluctuations.

In his formal approach to the business cycles theory, Schum peter has developed a model in two stages which he calls the *first approximation* and the *second approximation*. The first approximation deals with the initial impact of the innovatory ideas, and the second approximation deals with the subsequent waves that are created by the application of innovations.

The ***first approximation*** of Schumpeter's model starts with the economic system in equilibrium in which there is no involunt ary unemployment; each firm has its $MC = MR$ and P (price) = AC. Under the conditions of complete equilibrium in the economy, if an innovation in the form of a new technique or a new process of production is introduced, it will have to be financed through bank credit. For, the economy being in equilibrium, there is no surplus fund to finance the new adventure. With additional funds available from the banking system, the innovating firms go on bidding higher prices for other

inputs with a view to withdraw them from other uses. Due to increased spending in the economy, prices begin to rise. This process is further accelerated when other firms imitate the innovation and acquire additional funds from the banks. With the widespread adaptation of the innovation, output begins to flow into the market. The phase of expansion gets underway. But beyond a certain level, increased output causes a decrease in price and profitability. Since further innovations do not come by quickly, there would not be additional demand for funds. Rather, the firms which had earlier borrowed from the banks start paying back. This leads to contraction in money supply. Hence prices fall further. The process of recession begins and continues until equilibrium is once again restored.

The ***second approximation*** of Schumpeter's model analyzes the secondary waves that are created by the first approxi mation. The main element in the secondary wave is *speculation*. When the primary wave of expansion begins, investors in capital goods industries, expect the upswing to be permanent. With this expectation, existing firms borrow heavily. Even consumers anticipating higher price in future go into debt to acquire durable consumer goods. This heavy indebtedness causes a problem when prices begin to fall. Debtors, both investors and consumers, find it extremely difficult to meet their obligations. This situation leads to a panic and then to depression.

Critical Evaluation

Schumpe ter's theory has been criticsed on the ground that it does not offer a sound theory of trade cycle. As M.W. Lee8 puts it, "An objective evaluation of Schumpe ter's theory of the cycle is not only difficult" but also una vailing because most of his arguments are based on 'sociological rather than economic factors'. Hence, this theory can hardly be put to test. Besides, Schumpeter's theory is not basically differ ent from over-investment theory: it differs only in respect of the cause of variation in the investment when the economy is in state of equilibrium. Not only that, this theory too, like many other theories, leaves out many other important factors causing fluctua tions. Innovation is only one of the factors and not the sole factor.

30.3.4 Multiplier-Accelerator Interaction Theory

While business cycle theorists of Keynesian tradition empha size the *multiplier* process in economic fluctuation, J.M. Clark stresses the role of *acceleration* in business fluctuations. The post-Keynesian business cycle theorists9, however, contend that neither the theory of multiplier nor the principle of accelera tion alone is an adequate tool for analyzing the business cycles. In their opinion, an integration of multiplier and acceleration principles offers a much more satisfactory explanation to business cycles. They have, therefore, developed their own models and have shown the role of interaction between multiplier and accelerator in business fluc tuations. We will discuss here two prominent models of this category, viz., Samuelson's and Hicks' models.

8. Lee, W.W., *Economic Fluctuations* (Richard D. Irwin, Inc., Illinois), pp. 317–18.

9. Including Harrod, R.F., *The Trade Cycle Theory* (Oxford Clarendon Press, 1936); P.A. Samuelson, "Interac tion between the Multiplier and the Principle of Acceleration, "*Review of Economic Statistics*," May 1939; A.H. Hansen, *Business Cycle and National Income* (W.W. Norton and Company Inc., 1951); and J.R. Hicks, *A Contribution to the Theory of Trade Cycles* (Oxford University Press, 1950).

Samuelson's Model Samuelson's model is regarded as the first step in the direction of integrating theory of multiplier with the principle of acceleration. His model shows how *multiplier* and *accelerator* interact with each other to generate income, to increase consump tion and investment demands more than expected and how this process causes economic fluctuation.

To understand Samuelson's model, let us first distinguish between autonomous investment and derived investment. **Autonomous investment** is the investment undertaken due to exogenous factors such as new inventions in techniques of production, production process, and of new market, etc. *Derived investment* is the in vestment in capital equipment which is undertaken due to in crease in consumer demand necessitating new investment.

Let us de scribe the interaction process briefly. When autonomous invest ment takes place in a society, income of the people rises and the process of *multiplier* begins. Increase in income leads to increase in demand for consumer goods depending on the marginal propensity to consume. If there is no excess production capacity, the existing stock of capital would prove inadequate to produce consumer goods to meet the rising demand. Therefore, firms try to meet the growing demand and under take new investment. Thus, increase in consumption creates demand for investment. This is ***derived investment***. This marks the begin ning of *acceleration process*. When derived investment takes place, incomes rise further, in the same manner as it happens when autonomous investment takes places. With increase in income, demand for consumer goods rises. This is how the multiplier and the accelerator interact with each other and make the incomes grow at a rate much faster than expected. In brief, exogenous factors lead to autonomous investment. This results in multipli er affect. The multiplier effect creates derived investment. This is acceleration of investment. Derived investment creates multiplier effect leading to acceleration. This is called *multi plier-acceleration interaction*.

In his analysis of interaction process, Samuelson makes the following assumptions.

(*i*) No excess production capacity;

(*ii*) One-year lag in consumption;

(*iii*) One-year lag in increase in consumption and in vestment demand, and

(*iv*) No government activity and no foreign trade.

Samuelson's model of economic fluctua tions is briefly presented below. Given the assumption (*iv*), the economy will be in equilibrium when

$$Y_t = C_t + I_t \quad ...(30.1)$$

where Y_t = national income, C_t = total consumption expenditure, and I_t = investment expenditure, all in period t.

Given the assumption (*ii*), the consumption function may be expressed as

$$C_t = a\, Y_{t-1} \quad ...(30.2)$$

where Y_{t-1} is income in period t – 1, and a is *mpc* or $\Delta C/\Delta Y$. (Recall that $\Delta C/\Delta Y$ determines the multiplier).

Investment is a function of consumption with a one-year lag, i.e.,

$I_t = b(C_t - C_{t-1})$

...(30.3)

where b represents capital/output ratio. It is important to note here that the parameter 'b' determines the accelerator.

By substituting Eq. (30.2) for C_t and Eq. (30.3) for I_t, the equilibrium Eq. (30.1) can be rewritten as

$Y_t = aY_{t-1} + b(C_t - C_{t-1})$

...(30.4)

Recall that $C_t = aY_{t-1}$ and $C_{t-1} = aY_{t-2}$. By substitution, Eq. (30.4) can be written as

$Y_t = aY_{t-1} + b(aY_{t-1} - aY_{t-2})$

...(30.5)

By simplifying Eq. (30.5), we get

$Y_t = a(1 + b)Y_{t-1} - abY_{t-2}$

...(30.6)

Eq. (30.6) is the final form of equilibrium equation. This equation reveals the two necessary clues for analyzing the business cycles:

(*i*) If values for a and b, and in comes of two preceding years are known, then income for any past or future year can be determined, and

(*ii*) The rate of variation in income would depend on the values of parameters a and b.

Samuelson has shown in his pioneering work the various kinds of cycles that would be generated by different combinations of a and b. Through a diagram reproduced here (Fig. 30.2), he has shown the various types of cycles caused by the different combinations of a and b. The various combinations of parameters a and b have been shown by areas marked by A, B, C and D, each having a different pattern of trade cycles. The different patterns of trade cycles resulting from different combinations of a and b are shown in areas A, B, C and D marked in Fig. 30.3. The various combinations of a and b and the corre sponding natures of cycles may be briefly decribed as follows.

Fig. 30.2 *Combination of Parameters a and b and Trade Cycles*

Area A. All the combinations of a and b falling in area A make the incomes move upward or downward at decreasing rates asymptotically reaching a new equilibrium. In this area, rise or fall is one-way. It creates *damped nonoscillation*, as shown in part A of Fig. 30.3.

Area B. The combinations of a and b in area B produce cycles of amplitude growing smaller and smaller until the cycles disap pear and the economy is ultimately stabilized. This, according to Samuelson, is the case of *damped* cycles or *damped oscillation*.

Area C. Combinations of a and b in area C produce a series of trade cycles with larger and larger amplitude. Combinations of a and b in this area create *explosive* cycles, as shown in area C of Fig. 30.3.

Area D. Combinations of a and b falling in area D make the income increase (or decrease) at an exponential rate until the ceiling (or bottom) is hit. This is the case of one-way explosion, that creates *explosive oscillation* as shown in area D.

Point E in Fig. 30.2 shows a special case in which cycles are of equal amplitude an continue forever.

Fig. 30.3 *Trade Cycle Patterns*

Critical Evaluation

Samuelson's model, highly acclaimed as a sound attempt to integrate the Keynesian multiplier theory and Clark's accelera tion principle, has been on account of the following shortcomings.

First, Samuelson's model is regarded by its critics as far too simple a model to fully explain what actually happens during the period of economic fluctuations. In the opinion of his critics, this model has been developed on highly simplifying assumptions.

Secondly, Samuelson's model emphasizes the role of multiplier and accelerator, and interaction between them in economic fluctua tions. Like many earlier theories, Samuelson's model too leaves out or gives little consideration to many other important factors which might play an equally important role in business cycles, e.g., the role of producer's expectations, changing psychology of businessmen, changing consumers' preferences and exogenous factors.

Thirdly, one of the major shortcomings of the model is that it assumes constancy of capital/output ratio whereas there is a great likelihood of changes in this ratio during the periods of upswings and downswings. If this assumption is dropped, cycles may have different shapes and amplitudes from those suggested by the model.

Finally, as Shapiro10 has pointed out, many cyclic patterns suggested by the model do not conform to the real world experience.

30.3.5 Hicksian Theory of Trade Cycle

Hicks combines Samuelson's multipli er-accelerator interaction model and Harrod-Domar growth model to expound his theory of trade cycle.11 In his opinion, business cycles have historically taken place against the background of economic growth, and hence, the trade cycle theory should be linked with growth theory. In his theory of trade cycle, Hicks uses (*i*) Keynesian concept of saving-investment relation and the multiplier, (*ii*) Clark's acceleration principle, (*iii*) Samuel son's multiplier-accelerator interaction and (*iv*) Harrod-Domar growth model. These are the main ingredients of Hicks' theory of trade cycle. Let us now look at his theory of trade cycle in detail.

Assumptions

1. There is an ***equilibrium rate of growth*** in the model economy in which realized growth rate (*Gr*) equals the natural growth rate (*Gn*). The autonomous investment increases at a con stant rate which always equals the rate of increases in voluntary savings.

2. Hicks assumes a ***Samuelson-type of consumption function***, i.e., $C_t = aY_{t-1}$ (with one-year lag in consumption). The reasons he gives for the lagged consumption function are: (*i*) lag of expenditure behind income, and (*ii*) lag in non-wage income behind the change in *GNP*. The saving function naturally becomes the function of past year's income. Given these assumptions, Hicks' multiplier becomes a 'mathematical truism'.

3. The ***autonomous investment is a function of current output*** and is undertaken to replace the wornout capital. Induced investment, in his model, is a function of change in output. The change in output generates *induced investment* which brings the accelerator principle in action. It may be noted here that this acceleration interacts with multiplier effect upon income and consumption.

4. Finally, unlike Samuelson's ever-widening, explosive cycles (case *C*), Hicks prescribes 'ceiling and bottom' for the upswing and downswing. The ceiling on upward expansion is imposed by the 'scarcity of employable resources'. As regards the limit of the downswing, he says 'there is no such direct limit on contraction'.12 But an indirect limit is imposed by the mechanism of accelerator on the downswing.

10. Shapiro, Edward, *Macroeconomic Analysis, op cit.* , pp. 446–47.

11. *A Contribution to the Theory of the Trade Cycle, op. cit.*

12. Hicks, J.R., *op. cit.* , p. 95.

Graphical Presentation of Hicks' Theory of Trade Cycle

Hicks has outlined his theory of trade cycle in a diagram (see Fig. 30.4). The vertical axis measures the logarithms of output and employment and the horizontal axis measures time by its semi-logarithmic scale. In Fig. 30.4, line *AA* shows the course of *autonomous investment*, increasing at a constant rate. The line *EE* is the *equilibrium path* of output, which is a constant multiple of autonomous investment. The line *FF* is the *full em ployment* ceiling. The line *LL* shows the equilibrium path during the period of slump, assuming the output level will never go below this level. The line *LL* marks the botto line of the economy.

F

P

P

1

Q

E

1

F

P 2

P

L

0

E

Q_2

Output

q

L

A

A

O

X

Time

Fig. 30.4 *Hicks' Model of Trade Cycles*

Now, Hicksian theory of trade cycle can be described as follows. Suppose that the economy has been progressing on the dynamic equilibrium path, and reaches point P_0. Suppose also that at this juncture autonomous investment takes place due to some invention. Consequently, output increases and the economy leaves the equilibrium path *EE* and moves upward. After a certain time lag begins the multipli er process caused by the autonomous investment. As a result, output and employment increase. Increase in output leads to induced investment which in turn brings the accelerator in action. This interaction between the multiplier and accelerator causes expansion in the economy and the economy moves along the expansion or oscillation path P_0P_1 until point P_1 is reached. The expansion beyond P_1 will not be possible because of full employment con straint. The most it can do is to creep along the ceiling *FF*. But it cannot do so for long. For, the initial burst of autonomous investment was supposed to be shortlived; 'thus on the upper part of the path P_0P_1 no more than the normal amount of auto nomous investment is taking place.' It imp lies that the expansion along P_0P_1 has been mainly on account of the induced investment during the preceding periods.

However, once the ceiling is hit, the expansion sustained by the induced invest ment along the line *FF* is bound to end and a downswing becomes inevitable with a time lag.

The reason is that increase in output along *FF* is not high enough to induce investment and hence the induced investment ceases to take place. The downswing may be delayed if output-investment (induced) relation has a three or four year lag. But the downfall in output is inevitable. Once the downfall has started, say, at point *P'*, it must continue till it hits line *EE*. Since there is nothing in the process to stop it at *EE*, the downfall will con tinue further. It marks the beginning of the slum process.

The course of slump is shown by curve Q_1Q_2. That is, output decreases along the curve Q_1Q_2 towards the slump equilibrium line *LL*. Another course of possible slump is shown by Q_1q, when the output plunges downward indefi nitely. However, this is a rare possibility. The normal course of slump is Q_1Q_2.

Turning to recovery, when the downswing hits the bottom, it starts moving along the lower equilibrium line *LL*. This line is linked to the autonomous investment line *AA*, and rises with it. Thus, at this stage output will again start to rise. This increase in output should bring the accelerator back into action. This marks the beginning of recovery. Once the autonomous invest ment starts coming in, the process of multiplier and, later, its interaction with accelerator makes the economy grow on the path of expansion towards equilibrium path *EE*. This completes the cycle.

Critical Evaluation

Hicksian theory of trade cycle is regarded as an up-to-date, most modern and a highly streamlined theory of trade cycle. As Lee13 remarks, "It incorporates all the best features of earlier models and scrapes most of those which have not proved out in the past." Besides, Hicks has pointed out that there is a difference between the roles of accelerator at the upswing and at the down swing. On the downswing, the accelerator is inoperative because of excess capacity. However, Hicksian theory of trade cycle has been criticized on the following grounds.

First, the Hicksian theory of trade cycle is like other theories of this tradition, it too does not provide sufficient reasons for the linear consumption func tion and a constant multiplier. It is quite likely that during the phases of expansion and contraction, incomes are redistributed affecting the marginal propensity to consume and hence the multi plier.

Second, the assumption regarding the constancy of multiplier under dynamic conditions is looked upon with skepticism. Without a sound empirical evidence, the whole discus sion on acceleration principle assumes an abstract character and, in that case, it retains only academic interest. The empirical studies have also not provided evidence to the assumption of constant accelerator14.

Third, Hisksian theory, like some other theo ries, is regarded as a highly abstract formulation which seems incapable of explaining the phenomenon of fluctuations in real life.

Despite these limitations, the Hicksian theory of trade cycle is considered to be the most sound theory of trade cycle.

13. Lee, W.W., *Economic Fluctuations, op. cit.* , pp. 414–15.

14. See, for example, Tinbergen, *Statistical Testing of Business Cycle Theories* (League of Nations, Geneva, 1939), and his "Critical Remarks on Business Cycle Theories,' *Econometrica*, April 1942; Simon Kuznets, 'Relation between Capital Goods and Finished Products in Business Cycle' *Economic Essays in Honour of Wesley Clair Mitchel* (Columbia University Press, New York, 1935).

30.4 MEASURES TO CONTROL TRADE CYCLE

30.4.1 Need for Controlling Trade Cycles

Business cycles, especially the violent fluctuations in economic activities, cause not only harm to business community but also misery to the countrymen in general by creating unemployment and poverty during the depression period. Economists and the government have, of late, felt concerned with the consequences of business cycles and suggested various ways and means to control economic fluctua tions. The need for stabilizing economy and, thereby, preventing severe economic ups and downs was realized after the Great Depression of 1930s.

The experience of the Great Depression and Keynesian revolution in mid-1930s assigned a big role to the government in maintaining economic growth and preventing unem ployment and business fluctuations. Therefore, the government interventions in the economies all over the world increased in a big way. The government in free enterprise economies not only entered the production of commodities and services but also adopted a number of fiscal and monetary measures to control and regulate the private economy and prevent violent economic fluc tuations. The governments in many developing countries like India assumed the role of a key player in economic growth, employment and stabilization.

In this section, we discuss some major stabilization policies which are relevant to the current problems of the world.

30.4.2 Concept and Objectives of Stabilization

Stabilization broadly means preventing extreme ups and downs, particularly depression, in the economy without preventing or obstructing the process of economic growth. It also implies preventing over and under-employment. Stabilization does not mean bringing about rigidities in the economy—it should permit a reaso nable degree of flexibility for 'self-adjusting forces of the economy'.

The major objectives of stabilization policies are:

(*i*) preventing excessive and prolonged economic fluctuations, and making allowance for fluctuations necessary for a long-term sustainable economic growth;

(*ii*) creating conditions for efficient utilization of labour and other productive resources as far as possible; and

(*iii*) avoiding, as far as possible, the conflict between the internal and external interests of the economy.

The two most important and widely used economic policies to achieve economic stability are (*i*) *fiscal policy* and (*ii*) *mone tary policy*. In case of dire economic necessity, particularly when fiscal and monetary measures prove to be inefficient or inadequate, the government adopts *direct controls* to supplement fiscal and monetary measures. We have discussed below the fiscal and monetary policies which are generally adopted by the govern ment to control the business cycles and to stabilize the economy.

30.4.3 Fiscal Policy: Meaning and Measures

The 'fiscal policy' refers to the government policy of changing its taxation and public expenditure programmes intended to achieve certain predetermined objectives. Taxation is a meas ure of transferring funds from the private purses to the public coffers; it amounts to withdrawal of funds from the private use. Public expenditure, on the other hand, increases the flow of funds in the economy. Taxation reduces private disposable income and, thereby, the private expenditure. The public expenditure, on the other hand, increases private incomes and, thereby, the pri vate expenditure. Since tax-revenue and public expenditure form the two sides of the government budget, the taxation and public expenditure policies are also jointly called *budgetary poli cy* or *fiscal policy.*

Fiscal or budgetary policy is regarded as a powerful instru ment of economic stabilization. The importance of fiscal policy as an instrument of economic stabilization rests on the fact that government activities in modern economies are greatly enlarged, and government tax-revenue and expenditure account for a consid erable proportion of *GNP*, ranging from 10 to 25 per cent. Therefore, the government may affect the private economic activities to the same extent through variations in taxation and public expendi ture. Besides, fiscal policy is considered by some economists to be more effective than monetary policy because the former directly affects the private decisions while the latter does so indirectly. If fiscal policy of the government is so formulated that it generates additional purchasing power during depression and it contracts purchasing power during the period of expansion, it is known as '*counter-cyclical fiscal policy*'.

Counter-Cyclical Fiscal Policy

The counter-cyclical fiscal policy is based on the relation of public expenditure and taxes to the national income, the *GNP*. The relationship between public expenditure and *GNP* and between tax and *GNP* may be expressed in the form of the following propo sitions.

Public Expenditure and GNP *An increase in public expenditure raises the level of GNP. The size of increase in the GNP as a result of additional public expenditure is determined by the multiplier.* Public expenditure in the form of purchase of goods and services increases business incomes and household incomes—wage, interest, rent and business profit—which in turn increases government's tax revenue.[15] Marginal propensity to consume being greater than zero, households spend a part of additional income on consumption, and so do the people who earn additional income due to additional consumption expenditure by the households at the first instance. The process continues and *GNP* increases at the rate of multipli er[16].

Taxation and GNP Direct taxes without an equivalent public expenditure have adverse effect on *GNP*. Direct taxes have, therefore, a deflationary impact on the economy. *Increase in taxation either due to increase in the rates of existing taxes or due to imposition of new taxes, reduces GNP. The size of decrease in GNP as a result of increase in taxation depends on the tax-multiplier.* The multiplier in case of taxation works in the reverse direction. For, taxation reduces disposable income and hence consumption expenditure cumulatively. It should be noted here that the negative multiplier will not be as high as in case of public expenditure because the payment of taxes at the first instance does not reduce *GNP* as it is only a transfer of income. *Reverse multiplier or tax-multiplier will be one less than public-expenditure multiplier,* even if *mpc* is same in both the cases. The implication of the

15. Ekstein, Otto, *Public Finance* (Prentice-Hall, Inc., New Jer sey, 1967), 2nd Edn, pp. 102–3.

16. For detailed description of multiplier process, see Chapter 24.

expenditure multiplier being higher by one than tax-multiplier is that expenditure effect of a certain amount would more than neutralize the effect of taxation of an equal amount.

Counter-cyclical Fiscal Policy: Automatic and Discretionary Changes

It may be inferred from the relationship between public expenditure and *GNP* and between taxation and *GNP* that a countercyclical fiscal policy would require increase in public expenditure and reduction in taxation to fight depression, and reduction in public expenditure and increase in taxation to control inflation. In other words, fighting depression would require a *deficit budgeting* and controlling inflation requires *surplus budgeting.*

Some of the budgetary changes are ***automatic*** and some are ***discretionary***. The automatic budgetary adjustment takes place only when fiscal policy has built-in-flexibility. The automatic budget ary changes should follow the change in *GNP*. Built-in-flexibility in the fiscal policy implies that as *GNP* falls, both income and consumption decline. Consequently, the revenue from both direct and indirect taxes declines. Government's planned and commit ted expenditure remaining the same, public expenditure exceeds its revenue, and the budget automatically runs into deficit. This effect is more quick and powerful in the countries which provide unemployment allowances and other relief benefits. When *GNP* increases, tax base expands and tax-revenue increases. Expenditure level remaining the same, the budget automatically shows surplus. The deficit and surplus resulting from fluctuation in *GNP* work as automatic stabilizers of the economy.

It is, however, generally believed that automatic stabilizers prove to be adequate and serve useful purpose only for short-term fluctuations in the economy. Automatic stabilizers prove generally inadequate to control the economic fluctuations of larger amplitude. Under such condi tions, discretionary changes in budget become necessary.17

The ***discretionary changes*** in the budget refer to the changes in the tax-structure, and in the level and pattern of public expen diture by the government on its own discretion. Discretionary changes include change in tax-rate structure, abolition of existing taxes, imposition of new taxes, increasing and decreasing the public expenditure, changing the pattern of public expenditure, etc. Discretionary changes are so designed as to arrest the inflationary and defla tionary trends in the economy and to mitigate the destabilizing forces, such as increase or decrease in aggregate demand.

Problems in Formulating Counter-cyclical Fiscal Policy

Formulating a counter-cyclical fiscal policy is not an easy task. It involves certain complications, which should be borne in mind while devising the tax and expenditure policy to stablize the economy. Some complications have been pointed out by Eckstein18 as follows.

1. All expenditures do not have the same multiplier effect. For example, transfer payments by the government do not create a demand for goods and services. Some kinds of public expenditures (e.g., those on free education and hospital facili ties) replace private expenditure.

17. See Musgrave, Richard A., *The Optimal Mix of Stabilisation Policies in* W.D. Gramp, and E.T. Weiler, (eds.), *Economic Policy: Reading in Political Economy* (Richard D. Irwin, Inc., Illinois).

18. *Op. cit.,* pp. 109–111.

2. Not all tax-changes have the same multiplier effect. For example, taxes paid by the upper income groups have lower multiplier effect than those paid by lower income groups, because of differences in their *mpc*. The multiplier effects of indirect taxes are not clearly known.

3. Deficit financing through public borrowing may reduce private investment through crowding-out effect. This kind of deficit financing reduces the multiplier effect.

4. There are practical difficulties with regard to the assessment of time-lags and accuracy of forecasts. There is uncertainty with regard to effectiveness of fiscal policy.

30.4.4 Monetary Policy

Monetary policy refers to the central bank's programme of changing monetary variables, viz., total demand for and supply of money, inter est rates and credit rationing, to achieve certain predetermined objectives. One of the primary objectives of monetary policy is to prevent inflation and to achieve economic stability. The following are the traditional monetary instruments through which a central bank carries out the monetary policies.

(*i*) Open market operations,

(*ii*) Changes in bank rate (or discount rate),

(*iii*) Changes in the statutory reserve ratios, and

(*iv*) Selective credit controls and Moral suasion.

All these instruments when operated by the central bank reduce (or enhance) directly and indirectly the credit creation capacity of the commercial banks and thereby reduce (or increase) the flow of funds from the banks to the public. The pattern of use and working of these instruments is described here briefly.

1. ***Open Market Operations.*** *Open market operation* by the central bank refers to the sale and purchase of government bonds, treasury bills etc., to and from the public. The open market operation is carried out through scheduled commercial banks. During the period of expansion, the central bank sells the government bonds and securities to the public. The sale of securities reduces their price on the one hand, and results in withdrawal of money from the public, on the other. To the extent the government securities are purchased through the trans fer of bank deposits to the central bank account, it reduces the credit creation capacity of the commercial banks. Open market operation works successfully only if (*a*) govern ment securities are popular, (*b*) people have a good deal of banking habit and (*c*) banking system is fairly developed. Under these conditions, the sale of public bonds results in monetary contraction. During the period of depression, the central bank buys the government secu rities. Its impact on money supply with the public is just opposite to the impact of the sale of securities.

2. *The Bank Rate or Rediscount Rate.* Bank rate is the rate at which the central bank discounts banks 'first class' bills of exchange or grants short-term loans to banks. When the objective is to control inflation, the central bank raises the bank rate. A rise in the bank rate increases the cost of borrowing from the central bank. Following the increase in bank rate, commercial banks raise their own discount rates for the public. The increase in cost of borrowing discourages public borrowings from the commercial banks. Thus, the flow of money towards the private economy is re strained. But this method is effective only when commercial banks do not possess excess reserves. On the contrary, during depres sion,

the bank rate is lowered with a view to facilitating and encouraging private borrowing, which leads to monetary expansion and works against the forces of depression. Monetary expansion pushes the aggregate demand up and, thereby, helps economic recovery. Bank rate is the most widely used method. The RBI used the interest rates directly to control inflation in India during April-May 2007. Now bank rate has been replaced by 'Repo Rate' (Repurchase operation rate). Repo rate is defined as the rate at which central banks (RBI in India) lend money to commercial bank as a short-term fund. Against repo rate, there is 'reverse repo rate,' the rate at which commercial banks deposit their surplus finds with the central bank. The RBI now uses the repo rate very frequently to control.

3. *The Statutory Reserve Ratio (SRR).* When the central bank wants to reduce the credit creation capacity of the commercial banks, it increases the ratio of their demand and time deposits to be held as reserve with the central bank, and *vice versa*. Therefore, an anti-inflationary monetary policy requires increas ing the liquidity ratios, and anti-deflationary policy requires lowering the liquidity ratios. When the central bank changes the *SRR*, the deposits which form the basis of credit creation are affected and it affects the capacity of banks to create credit.

Of the three instruments of monetary control, the open market operation is considered to be the most effective weapon available to the central bank, specially in the less developed countries having under-developed money markets. The open market operation is flexible and easily adjustable to the changing conditions. The other two instruments are effective only when (*i*) commercial banks do not possess excess cash reserves and (*ii*) in case of bank rate, borrowers are not highly optimistic about future business prospects.

4. *Selective Credit Controls and Moral Suasion.* In addition to the instruments discussed above, central banks often use various *selective credit control measures* and *moral suasion*. The selective credit controls are intended to control the credit flows to particular sectors without affecting the total credit, and also to change the composition of credit from an undesirable to a desirable pattern. *Moral suasion* is a persuasive method to convince the commercial banks to carry out their business in accordance with the demand of the time and in the interest of the nation.

CONCLUSION

The fiscal and monetary policies may be alterna tively or simultaneously used to control business cycles in the economy. The choice of the policy should be made in view of their suitability, applicability and their limitations. The policy needs of a country vary from problem to problem and, therefore, the choice of policy also varies. For instance, monetary policy is considered to be more effective to control inflation than to control depres sion. It is, however, always desirable to adopt a proper mix of fiscal and monetary policies to check the business cycles. It is essential also because both the policies have their own limita tions. Therefore, an appropriate mix of fiscal and monetary policies would always prove more effective than a single policy. Besides, a proper mix of the two policies is also essential because it would avoid the possible conflict between them. It is, therefore, always desirable to formulate a counter-cyclical policy with a proper coordination of fiscal and monetary po licies.

SUMMARY

- Business cycle or trade cycle refers to subsequent phases of economic growth and depression in the economy repeated over a period of time.
- In general, business cycles have five sequential phases: (*i*) economic expansion, (*ii*) peak of prosperity – a high growth rate of production and employment, (*iii*) recession in economic activities – beginning of decline in production and employment, (*iv*) trough – depression hitting the bottom line, and (*v*) phase of recovery – beginning of economy recovery.
- Economists of different era have offered different theories of business cycles. According to classical theories, known as monetary theory, business cycles take place due to fluctuation in money and credit supply.
- According to monetary overinvestment theory, trade cycles arise due to increase in monetary investment in excess of desired investment. Desired investment is determined on the basis of availability of productive resources.
- Another business cycle theory was offered by Joseph Schumpeter, known as 'theory of innovation'. This theory states that trade cycles are caused by innovations of consumer and capital goods. Innovations lead to high investment causing high rise in price. High price generates speculative tendency and collapse of the market. The process begins once again.
- Samuelson propounded a theory of business cycle by combining the theories of multiplier and acceleration. Multiplier begins with autonomous investment. The multiplier effects autonomous investment generates derived investment leading to acceleration of investment. Thus multiplier and acceleration principles interact to create different kinds of interaction. According to their positive and negative effects of interaction of multiplier and acceleration create the conditions for growth or depression.
- Hicks propounded a more profound theory of trade cycle. Hicks combined Samuelson's multiplier-accelerator interact theory and Harrod-Domar growth model to develop his theory of trade cycle. According to his theory, multiplier and acceleration combined with growth parameters make the economy to grow. But since resources are limited, growth hits the upper limit and growth rate tends to decline causing decrease in investment. This reverses the process of multiplier-acceleration interaction cause growth rate to be negative leading to economic depression. The process repeats to cause business cycle.
- In general, two kinds of economic policies are used to control the trade cycle, especially to control depression in the economy. Policy measures to control depression are: (*i*) fiscal policy, and (*ii*) monetary policy. Fiscal policy measures include increasing public expenditure and cutting down taxation to prevent decrease in demand. Monetary policy measures including cutting down the interest rate and creating conditions for credit expansion.

REVIEW QUESTIONS

1. What is meant by business cycle? What are the different phases of a business cycle? How is private business affected during the different phases?
2. Describe 'the turning points' and the factors responsi ble for them in the business cycles.
3. Compare and contrast the pure monetary and the monetary over-investment theories of business cycles.

4. How does Schumpeter's innovation theory differ from the monetary over-investment theory of trade cycle? Which of the two theories explains business fluctuations better?
5. Explain Samuelson's multiplier-accelerator interaction theory of trade cycle. Also point out its shortcomings.
6. Why is controlling trade cycles necessary? Describe the major stabilization policies and point out their limitations.
7. Describe briefly the main functions of monetary and fiscal policies. Which of the two policies is more effective in controlling trade cycles in a developing economy?
8. What is meant by built-in-flexibility in budgetary policy? Comment on its relevance to a developing economy like India.
9. Describe the various phases of a trade cycle. Discuss the steps a businessman may take to safeguard himself against the evil effects of a trade cycle.

10. Describe briefly the Hicksian theory of trade cycle and point out its merits and limitations.
11. What is meant by fiscal policy? What kind of fiscal policy is adopted to prevent economic depression and to revive the economy?
12. What is meant by monetary policy? What are the monetary measures? What kind of monetary policy is adopted to control depressionary, factors?

FURTHER READING

A.E.A., *Readings in Business Cycle Theories,* American Economic Association.
Clark, J.J. and M. Chen, (Eds.), *Business Fluctuations, Growth and Economic Stabilization* (Random House, N.Y., 1963).
dauten, C.a., *Business Cycles and Forecasting* (South-Western Publishing Co. Cincinatti, 1961), 2nd Edn., Chs. 3-7.
dwivedi, d.n. *Macroeconomics: Theory and Policy* (New Delhi: Tata McGraw-Hill, 2010), Ch. 22.
Gordon, r.a., *Business Fluctuations* (Harper and Brothers Pub lishers, New York, 1952), Chs. 8–12.
haberler, G., *Prosperity and Depression* (United Nations, New York, 1952), 3rd Edn., Parts I and II.
haMberG, d., *Business Cycles* (The Macmillan Company, New York, 1951), Chs. 3–8.
hansen, a., *Business Cycle and National Income* (W.W. Norton and Company Inc., New York, 1951).
hiCks, J.r., *A Contribution to the Theory of the Trade Cycle* (Oxford University Press, London, 1950).
lee, w., *Economic Fluctuations* (Richard D. Irwin, Inc., Illinois, 1955).
saMuelson, P.a., "Interaction between the Multiplier Analysis and the Principle of Acceleration," *Review of Economic Statistics,* May 1939.

CHAPTER

31 Theories of Inflation and Control Measures

CHAPTER OBJECTIVES

Business cycles have been a rare phenomenon in the world history but inflation has been a frequent phenomenon, especially in developing countries. Inflation has both favourable and unfavourable effects on business. The objective of this chapter is to discuss the following aspects of inflation.

- Meaning and measurement of inflation
- Kinds of inflation experienced by countries
- Classical and modern theories of inflation
- Effects of inflation on the economy and
- Policy measures to control inflation

31.1 INTRODUCTION

In the preceding chapter, we have discussed causes and consequences of business cycles and the main theories that seek to explain the factors and forces causing business cycles. Business cycles of high magnitudes were considered till the end of 20th century as the thing of past. But the global recession of 2009–13 revived the possibility of business cycles. However, inflation has been a perpetual problem. What has gone nearly uncontrolled over time, especially during the post-World War II period, is the *problem of inflation.*
The problem of inflation got accentuated since the early 1970s. It emerged as the most intractable economic problem for both theoreticians and policy -makers all over the world. A continuous rise in the general price level over a long period of time has been the most common feature of both the developed and the developing economies. (See Appendix to this chapter). Persistent inflation is perhaps the second most serious macroeconomic problem confronting the world economy—second only to hunger and poverty in the 'third world'. Some authors consider inflation as the 'dominant economic problem' in modern times. In this chapter, we will discuss the various aspects of inflation as noted above[1].

1. Readers interested in a detailed study of inflation are advised to read, Chapters 21, 22 and 23 of D.N. Dwivedi, *Principles of Economics* (Vikas Publishing House, New Delhi).

31.2 DEFINITION OF INFLATION

31.2.1 What is Inflation?

Inflation means generally an unduely high and persistent rise in the general level of prices over a period of time. However, there is no universally acceptable definition of inflation. The definition of inflation has been changing over time depending on the perception of the economists. For example, according to Pigou[2], a neo-classical economist, "Inflation exists when money income is expanding more than in proportion to increase in earning activity." To Coulborn, inflation is a situation of "too much money chasing too few goods". Modern economists have tried to define inflation more meaningfully. According to Ackley, "Inflation is a persistent and appreciable rise in the general level or average of prices."3 Harry G. Johnson defines inflation as "a sustained rise in prices."4 According to Samuelson, "Inflation denotes a rise in the general level of prices."5 Bronfenbrenner and Holzman6 have suggested a number of alternative definitions of inflation which are mostly modified versions of earlier definitions. Their alternative definitions make things more fuzzy rather than adding clarity to inflation.

However, economists seem to agree that *inflation means a 'persistent' and 'appreciable' increase in the general level of prices*. The terms like 'persistent', 'appreciable', 'sustained', 'considerable', 'continuing' and 'prolonged' are not precisely defined. In practice, however, the term '*persistent*' implies that the price rise exhibits a secular trend or continues to rise over a period of one to two years, and does not respond to anti-inflationary policies. The term '*appreciable*' is more ambiguous because it does not specify as to what rate of increase in the price level is to be considered as 'appreciable' or 'considerable'. It varies from country to country and from time to time. For example, in India, a 7–8% annual inflation in 1970s and 1980s was not so appreciable but a 6% inflation April–May 2007 and 7–8% in 2013 had become highly problematic—a matter of great concern for both the government and the RBI.

31.3 METHODS OF MEASURING INFLATION

There are two common methods of measuring inflation: (*i*) by computing change in *Price Index Numbers* (*PIN*), and (*ii*) by comparing the change in *GNP Deflator.* The two methods are explained here briefly.

31.3.1 Wholesale PINs Method

The formula used for measuring the rate of inflation through the changes in the PINs is given below.

$$\text{Rate of inflation} = \frac{PIN_t - PIN_{t-1}}{PIN_{t-1}} \times 100$$

2. Pigou, A.C., "Types of War Inflation", *E.J.*, December 1947, p. 409 and in his *The Veil of Money,* 34.

3. Ackley, Gardner, *Macroeconomic Theory, op. cit.,* 421.

4. Johnson, Harry G., "A Survey of Theory of Inflation", *Ind. Eco. Rev.,* Vol. VI, No. 4, August 1963, reprinted in his *Essays in Monetary Economics* (George Allen & Unwin Ltd., London, 1966), p. 104.

5. Samuelson, P.A. and W.A. Nordhaus, *Economics,* 15th International Edn., 1955, p. 574.

6. Bronfenbrenner Martin and Franklin D. Holzman, "A Survey of Inflation Theory", in *Surveys of Monetary Theory*, Vol. I, 1965.

where PIN_t is the wholesale price index number for the year selected for measuring inflation and PIN_{t-1} is the *PIN* in the preceding year.

For example, consider some recent price index numbers in India. The *WPI* (2004–05 = 100) for 'all commodities' increased from 143.3 in fiscal year 2010–2011 to 156.1 in 2011–12. The rate of inflation between 1999–2000 and 2000–01 can be obtained by using the above formula as follows:

$$\text{Rate of inflation} = \frac{156.1 - 143.3}{143.8} \times 100 = 8.93$$

31.3.2 GNP Deflator Method

The *GNP* deflator method uses the ratio of nominal *GNP* in a year to the real *GNP* of that year. The *GNP* deflator is defined as follows:

$$GNP \text{ deflator} = \frac{\text{Nominal } GNP}{\text{Real } GNP}$$

where *Nominal GNP* is *GNP at current prices* and *Real GNP* is *GNP at constant prices.*

Thus, *percentage change* in *GNP* deflator between any two continuous years gives the rate of inflation. For example, suppose we want to measure inflation rate in year 2008–09 by using *GNP* deflator method. Inflation rate for 2008–09 can be measured by finding the change in percentage of *GNP* deflator in years 2008–09 and 2007–08.

India’s nominal *GNP* (i.e., *GNP* at current prices) in 2008–09 was `5270640 crores and her real *GNP* (i.e., *GNP* at constant prices of 2004–05 = 100) in this year was `4133292 crores. Now, India’s *GNP* deflator for 2008–09 can be obtained as follows:

GNP deflator (2008–09) = 5270640 = 12.75

4133292

In terms of percentage, India’s *GNP* deflator in 2008–09 equals 1.28 × 100 = 128.

Similarly, *GNP* deflator for 2007–08 is worked out at 1.18 and at percentage rate of 118.

The percentage change in *GNP* deflator between the two years gives a measure of inflation. That is, the rate of inflation in India between 2007–08 and 2008–09 can be obtained as follows:

Rate of inflation = 128 −118 × 100 = 8.47%

118

It is important to note here that *GNP* takes into account all the goods and services and all the prices in the economy. Therefore, *GNP* deflator method takes into account all the final prices—both wholesale and retail. In contrast, *WPI* takes into account prices only at the wholesale level. Therefore, economists consider *GNP* deflator as a better measure of inflation than WPI.7

7. See also Dornbusch, R. and S. Fischer, *Macroeconomics* (McGraw-Hill, NY, 1994), p. 36; Baumol, W.J. and A.S. Blinder, *Economics*: *Principles and Policy* (Harcourt Brace Jovanovich, London, 1988), p. 114; Froyen, T., Richard *Macroeconomics: Theories and Policies* (Macmillan, 1990), p. 35.

Inflation in India

The decade-wise annual average rate of inflation based on wholesale price index number (*WPI*) is given in Table 31.1. As the table shows, during the 1950s, the rate of price rise was very low—only 1.5 per cent per annum. Although price rise picked up during the 1960s, inflation, in the real sense of the term, started in the 1970s. During 1970s, the rate of inflation was 9.9 per cent per annum and during the period from 1980–81 to 1993–94, it was 8.1 per cent. The annual average rate of inflation during the period firm 1950–51 to 1993–94 was 6.5 per cent.

Table 31.1 *Annual Average Rate of Inflation in India (WPI: 1980– 81 = 100)*

Period (decennial)	*Rate of Inflation (%)*
1950-51 to 1960–61	1.5
1960-61 to 1970–71	6.1
1970-71 to 1980–81	9.9
1980-81 to 1993–94	8.1
1950-51 to 1993–94	6.5

Source: Centre for Monitoring Indian Economy (*CMIE*), *Basic Statistics Relating to the Indian Economy,* August 1994, Table 22.2.

Post–2000-01 Annual Inflation (y to y basis)

(1993–94 = 100)

2000–01	5.5
2001–02	1.6
2002–03	6.5
2003–04	4.6
2004–05	5.1
2005–06	4.1

2006–07	6.1

Annual Inflation: 2005–06 to 2011–12

(2004–05 = 100)

2005–06	4.5
2006–07	6.6
2007–08	4.7
2008–09	8.1
2009–10	3.8
2010–11	9.5
2011–12	8.9
2012–13	7.4
2013–14	6.0

Source: *Economic Surveys* of different years.

India had virtually no inflation during the 1950s. It had a moderate rate of inflation (6.1% p.a.) during the 1960s. In the 1970s, however, the rate of inflation had almost reached a double-digit rate (9.9% p.a.) — what was called galloping inflation. During the 1980s, the rate of inflation declined marginally to 8.1 per cent per annum. The annual average rate of inflation during the period from 1990–91 to 1995–96 was very high at 10.6 per cent (as published in *Economic Survey*, MOF, GOI, 2005–06, p. 80). What is noteworthy, India has had intermittent bouts of galloping inflation—about 14.0% in 1966–67; 20.1% in 1973–74, 25.2% in 1974–75; 17.1% in 1979–80, and 18.2% in 1980–81. However, as can be seen in Table 31.1, India experienced a moderate rate of inflation during 2001–06, annual average inflation rate being 4.7 per cent. However, in April 2007, it shot up to 6.5 per cent.

31.4 KINDS OF INFLATION

Inflation is generally classified on the basis of its rate and causes. The types of inflation based on its cause will be discussed under the causes of inflation. Here, we take a look at the kinds of inflation based on the *rate of inflation*. Inflation on the basis of rate is classified as (*i*) moderate inflation, (*ii*) galloping inflation, and (*iii*) hyper inflation. Another kind of inflation referred to in contemporary inflation studies is 'suppressed inflation'.

(*i*) Moderate Inflation. 'A single digit' rate of annual inflation is called 'moderate inflation' or 'creeping inflation'. During the period of moderate inflation, prices increase but at a moderate rate from 1% to 9%. The 'moderate rate' may vary from country to country. However, an important feature of moderate inflation is that it is 'predictable' and people hold money as a store of value. By this definition, India has had a moderate rate of inflation during the post-independence period, except in few years.

(*ii*) Galloping Inflation. A very high rate of inflation is called "galloping inflation'. How high should be the rate of inflation to be called galloping inflation is not defined precisely. According to Baumol and Blinder,8 "Galloping inflation refers to an inflation that proceeds at an exceptionally high rate." They do not specify what rate of inflation is 'exceptionally high'. Samuelson and Nordhaus9 define 'galloping inflation' more precisely. According to them, "Inflation in the double or triple-digit range of 20, 100 or 200 per cent a year is labeled galloping inflation". This definition is equally imprecise because the double and triple-digit inflation varies from 10% to 999% per annum. There is a very wide difference between these two lower and upper rates. A country with 900 per cent inflation will have devastating effects whereas a country with 20–30 per cent inflation can manage without pressing the alarm bell though in modern times, it does alarm the bell. The post-World

War I inflation in Germany is an example of galloping inflation. The wholesale prices in Germany increased by 140 per cent in 1921 and a colossal 4100 per cent in 1922, which is, of course, the case of hyper inflation. Some recent examples of galloping inflation, i.e., the annual average rate of inflation[10], during 1980–91 are as follows: Argentina – 416.9%; Brazil – 327.6%; Mexico – 66.5 %; Peru – 287.3% and former Yugoslavia – 123.0%. These cases are often quoted as examples of hyper inflation also.

8. Baumol, W.J. and A.S. Blinder, *Economics: Principles and Policy, op. cit.,* p. 109.

9. *Economics*, 15th Edn., *op. cit.*, p. 579.

10. CMIE., *World Economy and India's Place in it,* October 1993, Table 11.6.

(*iii*) Hyper Inflation. According to the early views, hyper inflation takes place when prices shoot up at more than three-digit rate per annum. During the period of hyper inflation, paper currency becomes worthless. Germany had hyper inflation in 1922 and 1923 when wholesale price index shot up by '100 million per cent between December 1922 and November 1923.'[11] November 1923 was the worst period of hyper inflation in Germany—"from January 1922 to November 1923, the price index rose from 1 to 10,000,000,000."[12] Hungarian inflation of 1945–46 is another example of hyper inflation, infact the worst case of hyper inflation ever recorded. In Hungary, "rate of inflation averaged about 20,000 per cent per month for a year and in the last month prices skyrocketed 42 quadrillion[13] per cent."[14]

In recent times, Argentina, Brazil, and Peru had hyper inflation in 1989 and 1991, as shown below:

Country	*1989*	*1990*
Argentina	3079.8%	2314.0%
Brazil	1287.0%	2937.8%
Peru	3398.6%	7481.7%

Source: CMIE, *World Economy & India's Place In It,* October 1993, Table 11.6.

The Suppressed Inflation

Another category of inflation that often comes across in contemporary writings on inflation is *suppressed inflation.* In contrast to open inflation (i.e., price rise without any control and regulation by the government), some countries experience ***suppressed inflation***. When prices continue to rise in spite of price control but at less than potential rate, it is called *suppressed inflation.* Price control through various direct or indirect price control measures has become a common feature of economic policy of most developed and developing economies. Price controls take the form of statutory fixation of the price or fixation of a price ceiling; rationing the consumption of scarce goods, controlled distribution of goods through public distribution system and subsidization of commodities with high inflation potentials. In spite of these control measures, prices do rise and inflation does take place but at a rate lower than the potential rate in the open system. This kind of inflation is called suppressed inflation.

31.5 THE CLASSICAL THEORY OF INFLATION

31.5.1 The Classical Keynesian View

Inflation has a long history. So is the case with the theory of inflation. Theories of inflation seek to explain what causes inflation.

11. Baumal, W.J. and A.S. Blinder, *Economics: Principles and Policy, op. cit.,* p. 109.

12. Samuelson, P.A. and W.D. Nordhaus, *op. cit.,* p. 579.

13. In USA and France, 1 quadrillion = 1,000,000,000,000,000 and in UK and Germany,
1 quadrillion = 1,000,000,000,000,000,000,000,000.
or 1 quadrillion = 1,000,000,000 zillion
or 1 quadrillion = 1,000,000 septillion.

14. Baumol, W.J. and A.S. Blinder, *op. cit.,* p. 109.

Looking chronologically at the development of inflation theory, the classical economists sought, for the first time, the cause of inflation through the *quantity theory of money*. In their opinion, the general level of prices rises in proportion to the increase in money supply, real output remaining the same. Thus, the classical theory emphasizes the role of money and ignores the real or the non-monetary factors causing inflation. This theory is therefore considered one-sided and incomplete. Keynes attributed inflation to excess aggregate demand at full employment level or the level of potential output. The excess aggregate demand is called *inflationary gap*. Keynes emphasized the role of a non-monetary factor, i.e., the aggregate demand in real terms and ignored the influence of monetary expansion on the price level. His theory too does not fully explain the phenomenon of inflation. The modern monetarists tried to revive the classical monetarism in a modified form, again emphasizing the role of money vis-a-vis inflation. Modern theories of inflation, on the other hand, recognize the role of both *demand-side* and *supply-side* factors on the price level. It explains the causation of demand-side and supply-side factors in the general equilibrium framework. Modern theories of inflation have a great relevance in the formulation of anti-inflationary policy.

In this section, we discuss only the classical view, known also as monetarists' view. The modern theories of inflation and the structuralists' view on inflation will be discussed in the next section.

31.5.2 Monetarist View

Modern monetarists, led by Milton Friedman, revived and modified the classical monetary theory of inflation. The modern monetarists[15] hold that the general level of price rises only due to increase in money supply but not proportionately. According to Milton Friedman, "Inflation is always and everywhere a monetary phenomenon… and can be produced only by a more rapid increase in the quantity of money than in output."[16]

More importantly, while classical economists, especially Irving Fisher, considered increase in the stock of money as the sole cause of inflation and price rise being proportional to increase in money supply, modern monetarists do not agree with proportionality of increase in price level. Recall here Fisher's quantity theory equation,

$$MV = PT$$

and

$$P = MV/T$$

where MV = money supply = currency × velocity of money; P = general price level; and T = total number of transactions (sale and purchase).

Thus, quantity theory reveals that, the volume of transactions (T) remaining constant, P increases in proportion to increase in MV, the total supply of money. However, this proposition is not acceptable to modern monetarists. In Friedman's own words, "In its most rigid and unqualified form the quantity theory asserts strict proportionality between

15. "Leading monetarists include the late Karl Brunner of the University of Rochester, Allan Meltzer of Cannegie-Mellon University, William Poole of Brown University, Anna Schwartz of the National Bureau of Economic Research and Hunter College, and Robert Barro of Harvard University". Quoted from Rudiger Dornbusch and Stanley Fischer, *Macroeconomics,* 6th edition (McGraw-Hill, Inc., New York, 1994), p. 209.

16. Milton, Friedman, *The Counter-Revolution in Monetary Theory,* Occasional Paper No. 33, Institute of Economic Affairs, London, 1970, p. 24.

the quantity of what is regarded as money and the level of prices. Hardly any one has held the theory in that form."[17] That is, modern monetarists do not agree that *there is proportional relationship between the supply of money and the price level.*

31.6 MODERN THEORIES OF INFLATION

The modern theories of inflation follows the theory of price determination. That is, the general price is determined by *aggregate demand* for and *aggregate supply* of goods and services and the variation in the aggregate price level is caused by the level of shift in the aggregate demand and aggregate supply curves. The modern theory of inflation is, in fact, a synthesis of classical and Keynesian theories of inflation. The modern analysis of inflation shows that inflation is caused by one or both of *demand* and *supply-side* factors. The factors which operate on the demand side are called *demand-pull factors,* and those operating on the supply-side are called *cost-push factors*. Accordingly, there are two kinds of inflation:

(*i*) **demand-pull inflation** and (*ii*) **cost push inflation**. The two kinds of inflation are discussed here briefly. However, a section of economists argues that demand-pull and cost-push factors interact to cause inflation. The theory of inflation based on interaction of demand-pull and cost-push factors has been discussed subsequantly.

31.6.1 Demand-Pull Inflation

The demand-pull inflation occurs when the aggregate demand increases at a much higher rate than the aggregate supply. In other words, demand-pull inflation occurs when, given the aggregate supply, aggregate demand increases substantially. Increase in aggregate demand may be caused by (*a*) monetary factors, i.e., increase in money supply and/or (*b*) real factors, i.e., increase in demand for real output. Let us now see how monetary and real factors cause inflation.

(a) Increase in Money Supply and Demand-Pull Inflation. One important reason for demand-pull inflation is increase in money supply in excess of increase in potential output. Whether increase in money supply in excess of increase in output is the only cause of inflation is a controversial issue. But the fact is that monetary expansion in excess of increase in the level of output is one of the most important factors causing demand-pull inflation.

Let us look at the process of demand-pull inflation caused by monetary expansion. When monetary and real sectors are in equilibrium at the same level of output and prices, the economy is said to be in general equilibrium. The general price level corresponding to the general equilibrium is called equilibrium price level. Now let money supply increase, other things remaining the same. The increase in money supply causes a decline in the interest rate. The decrease in the interest rate increase in transaction demand for money, especially for investment. Increase in investment causes increase in the level of income. Increase in income causes a rise in consumption expenditure. The rise in investment and consumer expenditures increases aggregate demand, aggregate supply remaining the same. This rise in aggregate demand is exactly proportional to the rise in the money stock. The rise in aggregate demand,

17. "Friedman, Milton, Money: The Quantity Theory" in *The International Encyclopedia of Social Sciences,* Vol. 10 (London, Corwell Collier and Macmillan, Inc.), 1968, pp. 432–47. Quoted in Rudiger Dornbusch and Stanley Fisher, *op. cit.,* p. 209.

given the aggregate supply, causes increase in the general price level. Thus, increase in money supply causes demand-pull inflation.

The German inflation of 1922–23 is often cited as an example of demand-pull inflation caused by the increase in money supply. During 1922–23, the German government had fallen under heavy post-war debts and reparations of payment obligations. The government, left with no option, printed and circulated billions and billions of paper currency. As a result, the general price level rose a billion fold. In recent times, the excess supply of money caused demand-pull inflation in Russia in 1990s 'when the Russian government financed its budget deficit by printing roubles'. Due to rapid increase in money supply, the general level of prices had risen in Russia during the early 1990s at an average rate of '25 per cent per month [or 100 × (1.2512 – 1) = 1355 per cent per year].'[18]

(b) Real-Factor Demand-Pull Inflation. Real-Factor demand-pull inflation can be caused by any or many of the following real factors.

(*i*) Increase in the government expenditure without change in tax revenue;

(*ii*) Cut in tax rates without change in the government expenditure;

(*iii*) Increase in investment;

(*iv*) Increase in consumer demand;

(*v*) Increase in exports given imports; and

(*vi*) Decrease in imports given the exports.

The first four factors straightaway increase the level of disposable income. Increase in aggregate income increases aggregate demand causing demand-pull inflation. For example, suppose that the government increases its spending financed through borrowings abroad. The rise in the government spending generates additional demand and, therefore, aggregate demand increases. Since there is, by assumption, full employment, additional resources can be acquired only by bidding a higher price. This pushes the prices up without increase in the output. Therefore, the transaction demand for money increases. In order to meet the additional transaction demand for money, people sell their financial assets—bonds and securities. Consequently, bond and security prices go down and the rate of interest goes up. In the product market, prices increase to such an extent that the additional government spending is absorbed by the price rise. This is how other real factors also cause inflation.

31.6.2 Cost-Push Inflation

There are instances of inflation which could not be fully explained by the demand-side factors. The 1958-recession in the western countries is a famous instance. During the period of recession, the aggregate demand had declined. Yet there was no decrease in the general price level. Instead, it tended to rise. In recent times, it is a common experience that prices generally do not decrease during the period of recession. Furthermore, even when there is stagnation in the economy and there is no inflationary pressure, the general price level generally continues to increase. It implies a situation of **stagflation.** The investigation into this kind of phenomenon, particularly for the 1958-puzzle, had led to the emergence of *supply-side theories of inflation*, popularly known as **cost-push theory** and **supply-shock** theory of inflation.

18. Samuelson, P.A. and W.D. Nordhau, *Economics*, *op. cit.*, p. 584–85.

Cost-push inflation is generally caused by monopolistic groups of the society, like labour unions and firms in monopolistic and oligopolistic market setting. Strong labour unions often succeed in forcing money wages to go up causing prices to go up. This kind of rise in price level is called **wage-push inflation.** Also, firms enjoying monopoly power have also been found using their monopoly power to raise prices which in turn leads to rise in the general price level. The monopolistic and oligopoly firms push their profit margin up causing a rise in the general price level. This kind of inflation is called **profit-push inflation.** Another kind of cost-push inflation is said to be caused by supply shocks, i.e., decrease in the aggregate supply. This is called **supply shock inflation.** *Minimum-wage legislation* and *administered prices* are other supply side factors which not only keep price level up but also create conditions for increase in the prices. In this section, we will discuss briefly these kinds of cost-push inflation.

(*i*) Wage-Push Inflation. Wage-push inflation is attributed to the exercise of monopolistic power by the labour unions to get their money wages enhanced more than the competitive labour market conditions would permit. Organized and powerful labour unions exercise their monopoly power and force their employers to increase their money wages above the competitive level without matching increase in labour productivity. Increase in money wages causes an equal increase in the cost of production. The increase in cost of production forces the firms to increase the price. This trend causes the aggregate supply curve to shift backward. A backward shift in the aggregate supply causes an upward movement in the price level. However, every rise in the money wages is not always inflationary. The rise in money wages due to the following factors is not said to be inflationary.

(*i*) Increase in wage rate due to increase in productivity,

(*ii*) Rise in wage rates due to inflation,

(*iii*) Wage rise where unionized wage bill is very small, and

(*iv*) Wage rise due to shortage of labour supply.

(*ii*) Profit-Push Inflation. In contrast to wage-push inflation, profit-push inflation is caused by the use of monopoly power by the monopolistic and oligopolistic firms to enhance their profit margin which results in the rise in price and inflation. Today, monopolistic competition and oligopolies characterize the real market situation all over the world. The monopolies, monopolistically competitive firms and oligopolies tend to increase their price to increase their profit margin, given the market conditions, of course. Therefore, a profit-push type of inflation is certainly a theoretical possibility.

Profit-wage Spiral It may be added here that ***wage-push and profit-push inflation go hand in hand***, whichever may be the leading cause. Labour unions may be the first to force wage rate to go up but firms raise the price level often more than proportionately. Or else, monopolistic firms may be the first to push the product price up forcing labour unions to demand a higher wage rate. It then has its repercussions on the money wages. Higher prices and profits induce demand for higher wages. The powerful labour unions force their employers to raise their wages. Following the wage hike, firms raise the product prices. When this process gets going, it takes the form of **'profit-wage spiral'.**

(*iii*) Supply-Shock Inflation. Supply shock is generally caused by unexpected decline in the supply of major consumer goods or key industrial inputs. For example, vegetable and foodgrain prices shot up in India by more than 100 per cent in the last quarter of

2013, making a big election issue in 2014. Prices of some key industrial inputs like, coal, steel, cement, oil, basic chemicals, etc., go up because of short supply caused by labour strikes, natural calamities, etc. Also, rise in the price may be caused by supply bottlenecks in the domestic economy or international events (generally wars) causing bottlenecks in the movement of internationally traded goods and causing thereby shortage of supply and rise in imported industrial inputs.

The sudden rise in the OPEC oil prices of 1970s due to Arab-Israel war is the famous example of the supply shock inflation all over the world. The OPEC had more than quadrupled the oil prices between 1972 and 1974. The oil price (Arabian Lights/Dubai) had increased from $ 1.90 per barrel in 1972 to $ 10.41 per barrel in 1974.[19] Due mainly to rise in the oil prices, the rate of inflation in India was 20.1% in 1973–74, 25.2% in 1974–75; 17.1% in 1979–80, and 18.2% in 1980–81 compared to the annual average of 6.1 per cent inflation during the preceding and about 8 per cent inflation during the succeeding decades. The other factors which had contributed to the high price-rise were failure of crops in 1972–73, aftermath of 1971 war and influx of Bangladesh refugees. For these reasons, prices had risen by 32 per cent in September 1974. This kind of inflation falls in the category of supply-shock inflation.[20]

31.6.3 Interaction Between Demand-Pull and Cost-Push Factors

One may conclude from the preceding discussion that, inflation is caused either by demand-pull or by cost-push factors. In fact, as mentioned above, demand-pull and cost-push factors work as cause-and-effect to cause a sustained increase in prices. Many economists hold the view that any one of these factors alone cannot cause inflation. To quote Machlup, “There is a group of economists contending that there cannot be a thing as cost-push inflation because, without an increase in purchasing power and demand, cost increases will lead to unemployment and depression, not to inflation.”[21] There is another group of economists who contend that ‘demand-pull is no cause of inflation, it takes a cost-push to produce it.’[22] The conclusion of these arguments is that neither cost-push nor demand-pull alone can cause and sustain inflation. In reality, cost-push and demand-pull interact to sustain the inflation over a period of time, whichever may be the cause of initial inflation. And, the dichotomy between the demand-pull and cost-push inflation is only a convenient way of classifying inflation. Besides, in the process of their interaction, cost-push and demand-pull factors get so intermixed that it is difficult to identify whether it is demand-pull or cost-push type of inflation. Some economists even contend that “the distinction between demand-pull and cost-push inflation is unworkable, irrelevant or even meaningless.”[23] It may be argued that demand-pull and cost-push inflation may be distinguished on the basis of which of the two factors caused rise in the price for the first time. But the question arises: *first* since what

19. Centre for Monitoring Indian Economy, Economic Intelligence Service, *World Economy and India’s Place In It,* October 1993, Table 11.11.

20. For more examples of supply-shock inflation, see Dernburg, Thomas F. *Macroeconomics: Concepts, Theories and Policies* (McGraw-Hill, New York, 1985), 7th Edn., Section “12.2. Food and Energy Shocks”, pp. 278–82.

21. Machlup, F., “Another View of Cost-Push and Demand-Pull Inflation”, *In Rev. of Eco., and Stat.,* Vol. 42, 1960, reproduced in R. Ball and Peter Doyle (eds), *Inflation: Selected Readings,* (Penguin Books, 1969 references from Ball and Doyle).

22. Machlup, F., in Ball and Doyle, p. 153.

23. Machlup, Fritz, *op. cit.*

time? If prices and wages have risen in turn, especially in a cause-and-effect manner, in successive years, ‘the choice of a base period is quite arbitrary and a conclusion assigning the leading role to one factor or the other would be equally arbitrary’. Nevertheless, demand-pull factor is regarded as the leading factor.

31.7 INFLATION IN LESS DEVELOPED COUNTRIES: THE STRUCTURALIST VIEW

Theories of inflation discussed above are all based on the characteristics and experience of the western developed countries. A question that may be asked here is: Can inflation problem faced by the less developed countries (LDCs) be explained by inflation theories developed with reference to the western industrialized nations? It is often argued that inflation theories based on the economic characteristics of the developed countries can hardly be applied to explain the nature of inflation in the developing countries and are, therefore, of little consequence in the formulations of anti-inflationary policy in LDCs.

The reason is that the characteristics and the institutional set up of the developed countries do not exist in the LDCs. Besides, in the framework of the orthodox theories, inflation takes place only when the economy is in the state of full employment with 'natural rate of unemployment', if any. The rise in the general price level prior to the state of full employment in the economy is not considered to be inflationary because that price level was lower than what was necessary to bring about full employment in the economy. *In contrast, in the less developed economies, inflation and large scale unemployment go hand in hand*. This has been the experience of most developing economies trying to achieve a high growth rate through the public sector investments.

As regards the institutional factors, the less developed economies are characterized by highly fragmented markets, market imperfections, immobility of factors, wage rigidities, disguised unemployment and underemployment, 'low equilibrium trap' and sectoral imbalances with surplus in some sectors and scarcity in others. Furthermore, inflation in the LDCs has generally been an inevitable consequence of their growth efforts. For these reasons, inflation theories built on the experience and in the background of the developed countries have little relevance to LDCs.

However, it is equally incorrect to say that inflation theories built in the background of developed countries cannot be applied to explain inflation in LDCs. The economic literature offers two plausible explanations or, in other words, two approaches to the phenomenon of inflation in the developing countries,24 viz., (*i*) monetarist approach, and (*ii*) structuralist approach. Monetarist approach to inflation has already been discussed above. In this section, we briefly discuss the structuralist approach to inflation in the developing countries.

Structuralists' Approach to Inflation

As mentioned above, inflation theories worked out in the background of the developed countries do not offer a reasonable explanation to inflation in the LDCs. Economists like Myrdal and Streeten argue strongly against straightaway application of the so called modern theories of inflation to LDCs. Their effort to find an appropriate explanation to inflation in

24. Ayre, P.C.I., "Money, Inflation and Growth" in Subrat Ghatak (ed.), *Monetary Economics in Developing Countries* (Macmillan, 1981), p. 68.

LDCs has led to the emergence of a new school of economists called '**structuralists**' and a new class of inflation theories known as **structuralist theories of inflation**. Some significant contributors of this school of thought are Myrdal,25 Streeten,26 and several Latin American economists.27 The structuralist view on inflation is briefly explained here.

According to the structuralist view, inflation in LDCs is an unavoidable result of their ambitious development programmes and is caused mainly by the structural imbalances in such economies. The structural imbalances in LDCs are:

(*i*) *food scarcity*: the imbalance between demand for and supply of food,

(*ii*) *input imbalance:* shortage of capital and surplus labour, shortage of fuel and oil,

(*iii*) *foreign exchange bottleneck:* imbalance between exports and imports and balance of payment deficits,

(*iv*) *infrastructural bottlenecks*: inadequate supply of electricity, transport and communication, and telecommunication, and

(*v*) social and political constraints.

In general, inflation in LDCs is caused by an admixture of factors including increase in money supply, effects of international factors, continued deficit financing, droughts and floods causing fall in agricultural products, high indirect taxes and persistent imbalance in demand for and supply of primary goods. However, economists have not offered a systematic theory of interaction of the structural imbalances in causing a persistent inflation in LDCs.

31.8 EFFECTS OF INFLATION

The economic effects of inflation are all pervasive. Inflation affects favourably or unfavourably all those who depend on the market for their supplies. Its effects may be high or low, beneficial or harmful depending on the rate of inflation and the kind of economic behaviour. A high rate of inflation is called 'economic evil' because it affects economy as a whole in many adverse ways. It affects different sections of the society in different ways and to different extents. The effects of inflation may be favourable or unfavourable depending the rate of inflation. But galloping and hyper-inflation have not only social and political implications but also prove disastrous for the economy. In this section, however, we will discuss only economic effects of inflation on the economy as a whole and on different sections of the people. We will discuss here the effects of inflation on (*i*) production and economic growth, (*ii*) employment of labour, (*iii*) distribution of income and (*iv*) distribution of wealth.

31.8.1 Effects on Production and Growth

Theoretically, the rate of economic growth depends primarily on the rate of capital formation and the rate of capital formation depends on the rate of saving and investment. Therefore, how inflation affects economic growth depends on how it affects savings and investment. Many economists hold the view that *inflation is conducive to economic growth* and that *there is positive relationship between inflation and economic growth*. For example,

25. Myrdal, G., *Asian Drama: An Inquiry into the Poverty of Nations* (Hormondsworth, Penguin, 1968).

26. Streeten P., *The Frontiers of Development Studies* (Macmillan, London, 1972).

27. For details, see Kirkpatrick, C.H. and F.I., Nixon, "The Origins of Inflation in Less Developed Countries: A Selected Review," in M. Parkin, and G. Zis, (eds.), *Inflation in Open Economies* (Manchester University Press, 1976).

Harry G. Johnson says, "…some degree of inflation—but a moderate degree only—is the logical concomitant of efficient economic mobilization."[28] He argues that apart from helping growth through redistribution of income, a moderate rate of inflation breaks the characteristic "rigidities and immobilities" of the underdeveloped economies and can "draw labour and resources out of traditional or subsistence sectors into the developing sectors of the economy" and helps efficient allocation of resources.[29]

Johnson puts forward two arguments in support of his view. ***First*,** during the period of inflation, output prices rise first and input prices follow. It implies that there is a **time-lag**[30] between the rise in output prices and rise in input prices, especially the wage rate. In fact, output prices increase first and wages increase after a time-lag. This time-lag between the output price rise and the wage rise is called **wage-lag**. If wage-lag persists over time, it enhances the profit margin which provides incentive for investment and investible funds to the firms. This results in increase in investment, production capacity and a higher level of output.[31]

***Second*,** inflation redistributes incomes in favour of higher income-groups who have higher propensity to save. Inflation-induced redistribution of incomes increases total savings because upper-income classes have a *higher propensity to save*. As a result, the level of savings increases which lowers the rate of interest. Lower interest rate induces new investment. Increase in investment enhances production capacity of the economy leading to increase in the total output. It means economic growth.

Does Inflation Help Economic Growth Really?

Empirically, there does not seem to be a clear evidence of positive relationship between inflation and economic growth, at least in the long run. The economists have different opinion on this issue. "Looking back to the record of the eighteenth and the nineteenth centuries, some economists find a positive relationship between inflation and economic growth in various countries."[32] Samuelson and Nordhaus refute this argument on the basis of the US experience: "Until the 1970s, high inflation usually went hand in hand with high employment and output. Rising inflation occurred when investment was brisk and jobs were plentiful…But a more careful examination of the historical record has revealed an interesting fact: The positive association between output and inflation appears to be only a temporary relationship. Over the longer run, there seems to be no sustained relation between a country's inflation rate and its level or growth of output or employment."[33] Different kinds of relations between inflation and growth have been observed during the post-World War II period:

(*i*) low rate of inflation and high rate of growth (West Germany);

28. Johnson, Harry G., "Is Inflation the Inevitable Price of Rapid Development or a Retarding Factor in Economic Growth?", *Malaysian Economic Review*, Vol. II, No. 1, April 1966, reproduced in Gerald M. Meier (ed.). *Leading Issues in Economic Development* (Oxford University Press, Delhi, 1995), pp. 179–82.

29. Johnson, Harry G., *op. cit.,* p. 179.

30. For the evidence of wage-lag, see Alchian, A.A. and R.A. Kessel, "The Meaning and the Validity of Inflation Induced Lag of Wages Behind Prices", *Am. Eco. Rev.,* March 1960, pp. 43–66; T.F. Cargil, "An Empirical Investigation of Wage-lag Hypothesis", *Am. Eco. Rev.,* December 1969, pp. 806–16.

31. For the effect of wage-lag on economic growth, see D. Felix, "Profit Inflation and Industrial Growth", *Qly. Jl. of Eco.*, August 1956, pp. 441–63, and E.J. Hamilton, "Prices as Factor in Business Growth", *Jl. of Eco. Hist.*,

December 1952.

32. Shapiro, Edward, *Macroeconomic Analysis*, 5th Edn. (Galgotia Publication (P) Ltd., New Delhi), p. 489.

33. Samuelson P.A., and W.D. Nordhaus, *Economics*, 15th Edn., pp. 582–83.

(*ii*) high rate of inflation and high rate of growth (Japan);

(*iii*) high rate of inflation and low rate of growth (United Kingdom)34 and

(*iv*) low rate of inflation and low rate of growth (India).35

However, a review of literature on the issue reveals that economists generally agree that a *moderate rate of inflation is conducive to economic growth*. In the words of Samuelson and Nordhaus themselves, *"While economists may disagree on the exact target for inflation, most agree that a predictable and stable or gently rising price level provides the best climate for healthy economic growth*. 36 In the long-run, economic growth of a country is affected by many other factors and therefore the relation between inflation and growth loses its distinctiveness.

In contrast, a very high rate of inflation of galloping or hyper type causes erosion in real savings and investment and, thereby, decline in real savings and investment. Dornbusch and Fisher quote evidence from Jorret and Selody37 that the output growth in Canada declined by 0.3 per cent for each 1 per cent increase in the inflation rate.38 In their opinion, "... there is no doubt that high inflation is bad for growth."39

31.8.2 Effect on Employment

The rate of employment depends generally on the rate of economic growth. The factors that accelerate the pace of economic growth determine also the course of employment. If inflation rate is such that it affects growth variables—savings, investment and profits—favourably, then it affects employment favourably too. However, growth rate and employment come in conflict at a high rate of inflation. A high rate of inflation increases employment but it affects growth adversely. Besides, inflation as a means to growth and employment involves severe economic and social costs. Evidence shows that a high rate of inflation causes distortions in relative prices, malallocation of resources, and social and political unrest. Policy-makers are therefore often faced with a dilemma: whether or not to control inflation. If inflation is allowed to presist on a high rate, it will affect growth adversely, and if it is controlled, it will affect employment adversely: there may be a high rate of unemployment. The policy-makers are therefore required to find **a trade-off between inflation and unemployment.**

31.8.3 Effect on Income Distribution

Whether inflation affects income distribution depends on how it affects *price received* and *price paid. Prices received are the same as incomes* defined crudely. Incomes are received, for example, in the form of wages and salaries, rents and royalties, dividend, interest, profits and self-employment earnings. Incomes are also received in the form of

34. Shapiro, Edward, *op. cit.*, p. 489.

35. The annual rate of inflation in India during 1950–51 to 1993–94 was 6.5 per cent which is close to the desirable limits of inflation (5–6% p.a.) for developing economies. This is, therefore, a low rate of inflation accompanied by a low rate of economic growth (3.5%) during this period.

36. In *Economics*, *op. cit.*, p. 583.

37. Evidence reviewed in Jack Selody, "The Goal of Price Stability", *Bank of Canada, Technical Report* No. 534, May 1960.

38. Dornbusch Ruddier and Stanley Fischer, *Macroeconomics*, Sixth Edn., *op. cit.*, p. 521.

39. *Ibid*.

old age pensions, unemployment allowances, etc. On the other hand, *prices paid represent expenditures*. Also, prices paid by one person are the prices received by another person and all prices do not change simultaneously and proportionately. *Inflation changes income-distribution-pattern only when it creates a divergence between the total prices received and the total prices paid by different sections of the society.* Let us now see how inflation affects income nad expenditure of different sections of the society.

Effects of Inflation on Different Sections of Society

Inflation has certain definite and predictable effects on the incomes of different sections of society. The effects of inflation on different sections of society are briefly discussed below.

(*i*) Wage Earners. It is a common belief that wage earners are hurt by inflation more than other sections of the society because wage rise lags behind the rise in consumer prices. However, some economists consider this belief as a myth.[40] In fact, whether wage earners lose or gain by inflation is again a matter of labour market conditions. In developed countries, labour is, by and large, organized and labour market is competitive. According to Baumol and Blinder[41], 'the average wage typically rises more or less in step with prices'. This contradicts the 'popular myth' that wage earners are, in general, losers during the period of inflation. Baumol and Blinder have used US data to show that real wage "is not systematically eroded by inflation". They add, "The fact is that in the long run wages tend to outstrip prices as new capital equipment and innovation increase output per worker". The Baumol-Blinder conclusion holds for at least the organized sector in India. In the organized sector, labour is unionized. The organized labour uses its union power to get compensatory increase in their wages. The labour in the organized sector is, therefore, often adequately compensated for the loss of purchasing power due to inflation. According to the official data, the public sector employees—a part of the organized sector—are more than doubly compensated. The per capita annual emoluments had increased by 1326.17 per cent between 1971–72 and 1994–95 whereas the consumer price index (1960 = 100) had gone up by only 630.21 per cent during this period.[42] The annual emoluments in the private organized sector has increased at a faster rate. It may thus be concluded that the wage earners in the organized sector have gained during the period of inflation.

(*ii*) Producers. Whether producers gain or lose due to inflation depends, theoretically, on the rate of increase in product prices and the cost of production. In general, product prices rise first and faster than the cost of production. The product prices rise first due to demand-pull factor, rise in money supply, supply bottlenecks, or a sudden rise in certain input prices (e.g., oil price). Since rise in input prices has a line lag, profit margin increases as a gain to the producers. This creates conditions for additional demand for inputs pushing the input prices up though at different rates and with different time lags. However, it must be borne in mind that wages and salaries do not increase automatically and simultaneously during the period of inflation. There is always a time-lag between the rise in commodity price and wages. The producers gain more than labour during the period of inflation due to wage-lag.

40. See, for example, Baumol, W.J. and Alan S. Blinder, *Economics: Principles and Policy*, *op cit*., p. 100.

41. Baumol, W.J. and A.S. Blinder, *Economics: Principles and Policy*, p. 101.

42. See *Economic Survey,* 1996–97, Government of India, Ministry of Finance, p. S-54.

(*iii*) Fixed Income Class. The fixed income class consists of house-owners getting fixed rent on their property; some category of wage-earners getting fixed wages, and artisans making a limited fixed income during the period of industrialization. The people of the fixed-income category are the net losers during the period of inflation. The reason is, their income does not increase—it remains constant— but the prices of goods and services they consume increase. As a result, the purchasing power of their income is eroded in proportion to the rate of inflation.

(*iv*) Borrowers and Lenders. In general, borrowers gain and lenders lose during the period of inflation. During the period of inflation, borrowers gain because when they pay off their debts, they pay a lower real value. Lenders lose for the same reason. For example, suppose a person borrows `5 million at 12 per cent simple rate of interest for a period of five years to buy a house. Suppose also that escalation in property prices is such that property prices double every five years. After five years, the borrower would pay a total sum of `8 million whereas the price of house rises tò10 million. The borrower thus gains by `2 million. The lender loses by the same amount in the sense that had he bought the house himself, the worth of his money would have risen tò10 million.

(*v*) The Government. The government is a net gainer during the period of inflation. It gains since inflation increases tax yields from personal income tax because (*i*) inflation redistributes income in favour of higher income groups taxable at higher rates of income tax and (*ii*) inflation increases the nominal income at the rate of inflation, real income remaining the same. As a consequence, an income which was non-taxable prior to inflation becomes taxable after inflation. Inflation enhances also the tax base and, therefore, tax revenue increases. In case of corporate income tax, tax-yield increases during the period of inflation on account of (*i*) increase in corporate profits, and (*ii*) increase in depreciation allowance due to increase in the nominal value of firm's assets. Inflation also enhances the revenue from *indirect taxes* because tax yield increases in case commodities are taxed at ad valorem rates. Finally, the government gains as a net borrower.

Conclusion

It may be concluded that inflation affects the earnings of different sections of the society in different ways and to different extent. In general, however, inflation redistribute income in favour of rich sections of the society. As a result, rich get richer and poor get poorer.

31.8.4 Effect on Distribution of Wealth

Here, the term 'wealth' means *net worth* defined as assets *minus* liabilities or debts. For our purpose here, net worth means the net worth of variable-price assets. Variable-price assets are those whose market value keeps changing depending on market conditions. This category of assets includes material assets like land, building, gold, etc., and financial assets like shares, bonds, time deposits, mutual funds, etc. The effect of inflation on the distribution of wealth depends on how inflation affects the net worth of the different classes of the wealth holders. In general, the value of variable-price assets increases at a rate higher than the cost of maintenance. Therefore, sections of the society holding large price-variable assets gain. This increases their returns on wealth and ability to accumulate more assets. Therefore, wealthy

people gain more from wealth holding. The result is redistribution of wealth in favour of rich section of the society, even if low asset holders maintain their assets. However, if low asset holders sell their property under the pressure of inflation, wealth distribution gets more accentuated in favour of the rich. However, the voluminous literature available on the subject and various empirical studies do not produce conclusive evidence on the effect of inflation on the distribution of income and wealth. To quote Samuelson and Nordhaus.43 "The summary wisdom of these studies indicates that the overall impact is highly unpredictable".

31.9 DESIRABILITY OF INFLATION

In the preceding section, we have discussed the *effects of inflation* on economic growth and different sections and sectors of the economy. As noted there, inflation has different kinds of effects – favourable or unfavourable – on different sections and sectors of the economy. However, the most important issue in this regard is the *effect of inflation economic growth*. It is most important issue because if inflation is helpful to a sustainable economic growth over a period of time, then its negative effects are a matter of adjustments. As noted above, the economists in general view a moderate inflation as a positive factor in economic growth, though there is no clear evidence. However, in case inflation is found to have a negative effect on economic growth and its overall effect is negative, the government is required to control inflation. But, if overall impact of inflation is positive on the economy, controlling inflation is considered to be undesirable. Therefore, before we proceed to discuss the policy measures to control inflation, let us have view of desirability of inflation for the benefit for the economy.

31.9.1 Is Inflation Desirable?

A moderate rate of inflation is considered to be desirable for the economy. The limit of desirable inflation varies from country to country and from time to time. The desirable rate of inflation can be determined on the basis of the rise in price that contributes in following ways:

(*i*) Keeping economic outlook optimistic and helping production and employment, and

(*ii*) Promoting mobilization of resources (savings and investment) by what is called inflationary method of financing.

The desirable limit of inflation depends on the need and the absorption capacity of the country. This limit is determined for an economy on the basis of: (*i*) whether inflation makes economy overheated; (*ii*) whether real macro variables like savings, investment and growth of output decline in spite of inflation, (*iii*) whether *BOP* position turns adverse, and (*iv*) whether employment gets adversely affected. These considerations and the absorption capacity, so defined, varies from country to country and from time to time depending on their growth potentials. Therefore, the desirable limit or the moderate limit of inflation has to be determined for each country and for each round of inflation. There is no definite rule in this regard.

What is Desirable Rate of Inflation?

There is no precise or unique answer to this question. However, based on the past experience, it is sometimes suggested that 1–2% inflation in developed countries and 4–6%

43. *Economics, op. cit.,* p. 581.

inflation in less developed countries is the appropriate and desirable limit of moderate inflation.44 Some consider a lower rate of inflation to be desirable. "Some people who regard inflation as an economic evil believe that a price level rising at around 1.5% ... assists in achieving and maintaining full employment and a satisfactory rate of growth".45 These rates of desirable inflation have no theoretical basis. However, these rates have great policy implications. In general opinion of the economists, so long as (*i*) the general level of price rises at an annual average rate of 2 – 3% in developed countries and 4 – 5% in less developed countries, and (*ii*) macro-variables are not adversely affected by the price rise, policy measures to control inflation are not required because controlling inflation under these conditions may distort the price system and disturb employment and growth process.

31.9.2 Every Price Rise is Not Inflation

One may think that *any price rise* in excess of 2 – 3% in the developed and 4 – 5% in the developing economies can be called undesirable inflation. *However, a price rise even in excess of these rates is not considered to be inflationary under the following conditions.*

(*i*) *When prices tend to rise due to change in the composition of GDP, it is not inflationary.* During the period of economic growth, the proportion of low-price goods, e.g., agriculture products decreases and that of high-price goods, e.g., cars, TV sets, computers, high price housing, increases causing a high rise in price index number. This rise in price is not taken as dangerous inflation.

(*ii*) *Price rise due to qualitative change in products is not inflation.* For example, in case of colour TV sets, qualitative improvements have been made in the form of multi-channel and remote control facilities. In case of cars, to consider another example, the brand of car may have qualitative improvement in the form of AC facility, automatic gear system and power brake, etc. Such qualitative changes involve increase in cost of production and therefore cause a rise in the price. But, it is not inflation.

(*iii*) *Short-run rise in price due to sudden increase in demand and/or decrease in supply is not inflation.* Sometimes prices rise because of sudden increase in market demand and/or sudden decrease in supply for such reasons as crop failure, strikes and lock-outs, pre-budget speculations, disruption of foreign supply due to war, etc. The price rise under such conditions is not supposed to be a persistent increase in the price level.

(*iv*) *Price rise after depression or recession is not inflationary.* Prices tend to rise during the phase of recovery after a short-run depression or recession to reach their normal level. Inflation under this condition is not considered to be undesirable inflation.

31.10 POLICY MEASURES TO CONTROL INFLATION

Economists agree that inflation beyond a moderate rate is bad and can often prove disastrous and, therefore, it must be kept under control. The various policy measures

44. Johnson, Harry G., "Is Inflation the Inevitable Price of Rapid Development or Retarding Factor in Economic Growth?", *Malayan Economic Review,* Vol. 11, No. 1, April 1966, partly reproduced in *Leading Issues in Economic Development,* (ed.) by Gerald M. Meier Delhi (Oxford University Press), Sixth Edition quoted from p. 179.

45. Shapiro, Edward, *Macroeconomic Analysis,* (Galgotia Publishers, New Delhi, 1994), p. 468.

generally adopted for controlling inflation can be classified as follows.

(*i*) monetary policy measures,

(*ii*) fiscal policy measures,

(*iii*) price and wage control measures, and

(*iv*) indexation.

These measures of controlling inflation are discussed here briefly.

31.10.1 Monetary Measures

Monetary measures which are generally used to control inflation include (*i*) bank rate policy, (*ii*) variable reserve ratio and (*iii*) open market operation.

(*i*) Bank Rate Policy. Bank rate more appropriately called, 'Central Bank rediscount rate', is the rate at which central bank buys or rediscounts the eligible bills of exchange and other approved commercial papers presented by the commercial banks for building reserves.46 The central bank performs this function as the 'lender of the last resort.' In general, bank rate policy is used even during the period of inflation as the main instrument of monetary control. The use of bank rate policy forms the basis of 'dear money' or 'tight money' policy depending on the need of the country. When the central bank raises the bank rate, if adopts a ‘dear money policy’.

The bank rate as a measure of inflation control works in two ways. ***One,*** during the period of inflation, the central bank raises the bank rate. This increases the cost of borrowing which reduces commercial banks’ borrowing from the central bank. This reduces banks’ ability to create credit through the process of credit creation. As a result, flow of money from the commercial banks to the public gets reduced. Therefore, inflation is controlled to the extent it is caused by the bank credit. ***Two,*** bank rate sets the trend for the general market rate of interest, particularly in the short-term money market. When bank rate is increased with a view to controlling money supply and, thereby, inflation, commercial banks increase their lending rates and other market rates follow suit. In general, the cost of borrowing goes up. This slows down the monetary flows to the society.

This method, however, does not work effectively, if (*i*) commercial banks have excess liquidity, (*ii*) they have alternative sources of creating reserves, (*iii*) they are free not to raise their lending rates following the increase in the bank rate, and (*iv*) future expectations regarding the market prospects is optimistic.

In India, the RBI is constrained to make full use of the bank rate policy as an instrument of monetary control for the fear (*i*) that it might raise the interest rate in the gilt-edged market and thereby increase government’s cost of borrowing and (*ii*) that it might result in capital loss to the bond-holders, the financial institutions. “The role of the bank rate as an instrument of monetary policy has been very limited in India, because of a number of factors like the administered structure of interest rates, sector specific refinance facilities for commercial

46. Since bill market in India is underdeveloped, the Reserve Bank of India (RBI) advances money to the commercial banks in two forms: (*i*) in the form of advances mostly against the government securities and (*ii*) rediscounting facility for eligible usance bill and other approved securities. As mentioned above, RBI makes this kind of lending under its traditional function as ‘lender of the last resort’ to help the commercial banks in the period of liquidity crisis. Since 26 October 2005, the RBI has been using ‘repo rate’ for short-term lending to commerical banks.

banks and underdeveloped bill market” (*Economic Survey, 1994–95*, Government of India, p. 43). Besides, a conflict has arisen in recent years between the RBI and GOI. The RBI is peeking freedom from the constraints imposed on the RBI in making policy decisions whereas GOI wants RBI to take government opinion on policy decisions.

(*ii*) Variable Reserve Ratio. Commercial banks are required to maintain a certain proportion of their total demand and time deposits in the form of *cash reserves,* called **cash reserve ratio (CRR).** The cash reserve ratio (*CRR*) is generally determined and imposed by the central bank which it uses as a weapon to control money supply. To control inflation, the central bank raises the *CRR*. When central bank raises the *CRR*, it reduces the lending capacity of the commercial banks. As a result, flow of money from the commercial banks to the public decreases. In the process, it halts the rise in prices to the extent it is caused by the banks credits to the public. This measure of controlling inflation has the same limitations as the bank rate policy.

(*iii*) Open Market Operations. Open market operations refer to sale and purchase of the government securities and bonds by the central bank. To control inflation, central bank sells the government securities to the public through the authorized commercial banks. The sale of government bonds through the banks results in the transfer of a part of bank deposits to central bank account. This reduces credit creation capacity of the commercial banks. Therefore, money supply with public is reduced by a multiple of sale proceeds of the treasury bills. As a result, inflation is controlled to the extent money supply with public decreases.

Open market operation is regarded as an efficient instrument of monetary control in the developed countries like the USA and the UK. This method is more effective than other methods of monetary control. In the developing countries like India, however, open market operation has not proved very successful mainly because treasury bill market is not developed and well organized. It is largely a captive-market confined to the financial institutions such as scheduled commercial banks, life insurance and general insurance companies, and the government financial corporations. These institutions are required by law to invest a certain proportion of their total deposits in the government bonds and securities. The treasury bill market in India is, therefore, said to be a ‘captive market’.

31.10.2 Fiscal Measures

Fiscal measures to control inflation include *taxation, government expenditure* and *public* *borrowings*. Keynesian economists, also called 'fiscalists', argue that demand-pull inflation is caused by excess of aggregate demand over aggregate supply. The excess demand is the result of increase in expenditure by the households, firms and the government, particularly by excessive spending by government. Excess demand, be it due to household or government expenditure, can be effectively controlled by fiscal measures. Therefore, fiscal policy measures are a more powerful and effective weapon to control demand-pull inflation.

In case *government expenditure* is the cause of excess demand, it can be controlled straightaway by cutting down public expenditure. A cut in public expenditure reduces government demand for goods and services and also the private income and consumption expenditure arising out of government expenditure multiplier. Therefore, the excess demand decreases more than a given cut in public expenditure. And, in case excess demand

is caused by rise in private expenditure, increasing income tax rate is a more appropriate measure to control inflation. Taxation of incomes reduces the disposable income of the people and thereby consumer demand.

In case of a very high rate of persistent inflation, the government may adopt both the measures simultaneously. It should cut down its own spending, on the one hand, and increase the rate of income taxation, on the other. This kind of policy is known as a **policy of surplus budgeting,** i.e., the government should spend less than its tax revenue.

Fiscal Vs Monetary Policy in Controlling Demand-pull Inflation?

Whether fiscal or monetary policy is more effective in controlling demand-pull inflation or any kind of inflation has been a matter of controversy. There has been a prolonged debate between the fiscalists and monetarists on this issue. Fiscalists argue that fiscal policy is more effective in controlling inflation whereas monetarists argue that monetary policy is a more effective weapon than the fiscal measures to control inflation. Economists called 'rationalists' and 'supply-siders' hold a different view from those of the monetarists and fiscalists on the measures to control inflation. Empirical evidence on the relative effectiveness of monetary and fiscal policies are not conclusive either. Some economists, viz., Andersen and Jordan, find in their researches monetary policy to be more effective than the fiscal policy in controlling inflation and promoting employment. Some other economists, viz., Leeuw and Kalchbrenner, find that fiscal policy is more effective than the monetary policy. Findings of Gary Fomm and R. Klien support the view that fiscal policy is more effective. Most researches on this issue find that fiscal policy is more effective, but they do not prove conclusively that monetary policy is not effective.

However, it may be suggested that if inflation is caused by the monetary factors, say, due to excess money supply, then monetary policy would be more effective. And, if inflation is caused by real factors, e.g., due to rise in the household demand and public expenditure, then fiscal policy would be more effective. In fact, an appropriate combination of fiscal and monetary policy is more successful in controlling inflation.

31.10.3 Price and Wage Control

In case monetary and fiscal measures prove ineffective in controlling inflation, direct control measures are adopted to control inflation. Direct measures consist mainly of price and wage controls. The price and wage controls go together because price-push and cost-push inflation go hand in hand whatever may be the cause of initial inflation.

Under **price control method,** a maximum retail price of goods and services is fixed. Price control may be general, applicable to all goods and services or it may be partial, confined to only scarce and essential goods and services. The primary objective of price control is to prevent the price rise of scarce goods and to ration the use of the commodity. Whether price control works effectively and efficiently is also a controversial matter. It is a general experience that price controls lead to black-marketing of goods and unfair distribution of scarce goods and services, especially where administrative machinery is corrupt and inefficient.

Wage control is used where inflation is of *cost-push* or of *wage push nature*. Under this method, rise in wage rate is prevented directly by imposing a ceiling on the wage incomes

in both private and public sectors. Often 'wage-freeze' is applied to control inflation. In case trade unions are powerful and oppose 'wage-freeze' effectively, government uses a weaker method called '**jawboning'**. A more sensible and effective method of containing wage-push inflation is known as '**wage guideposts'**— a plan of action against inflation prepared by common consent and mutual agreement between the representatives of the government, trade unions and the businessmen, for a disciplined and controlled upward movement in the wages and prices. Under this scheme, wage and price rise are monitored by a board of the representatives. However, this method does not work for long, especially when prices continue to rise. Friedman comments, "Guideposts and pleas for voluntary compliance are a halfway [measure] whose only merit is that they can more readily be abandoned than legally imposed controls. They are not an alternative to other effective measures to stem inflation, but at most a smoke-screen to conceal lack of action."[47]

31.10.4 Indexation

It should be clear from the foregoing discussion that inflation is an intractable problem. Controlling inflation involves the risk of aggravating unemployment problem. However, as discussed above, a high rate of inflation affects different sections of the society in different ways. Economists argue that if controlling inflation is not advisable, its adverse effects on different sections of society can be minimized by a method called **indexation** . They suggest indexation of prices, wages and contractual obligations with a view to compensating those who lose their real incomes due to inflation. According to Samuelson and Nordhaus, "**Indexing** is a mechanism by which wages, prices and contracts are partially or wholly compensated for changes in the general price level."[48] Thus, indexation is not a method of controlling inflation. It is a method of adjusting monetary incomes so as to minimize the undue gains and losses in real incomes of the different sections of the society due to inflation. Its main objective is to manage social discontent and to make inflation easier to live with.

Indexation of wages is most important and a common practice in many countries where wage contracts are long-term contracts and inflation continues to persist. In such cases, compensating workers for the loss of their real income due to inflation becomes unavoidable. Two systems are adopted in general: **one** is to tie wages to cost-of-living index (*CPI*) and **second** is to make a periodic scheduling of wage rise after *CPI* goes up by a certain percentage point. For instance, in the USA, wages are linked to cost-of-living index and cost-of-living adjustment (*COLA*) with inflation is made automatically. Here, wages increase automatically following the increase in the cost-of-living index. In India, wage compensation takes the form of 'dearness allowance', i.e., compensation for loss of purchasing power of the nominal wages due to inflation. Dearness allowance to the public sector employees in India is linked to consumer price index (*CPI*) and dearness allowance is granted after every eight percentage point increase in the *CPI*.

Although economists strongly recommend indexing of wages, debts, taxes, and all other long-term contractual payments, the governments doubt the *feasibility and practicability* of indexation method for three reasons.[49] ***First,*** adjustment in indexation is impracticable in case

47. Friedman Milton, "What Price Guidepost?", in Arthur M. Okun (ed.), *op. cit.*, p. 211.

48. Samuelson, P.A., *Economics,* 15th Edn., p. 596.

49. Dornbusch R and Jiseler, S., *op. cit.*, p. 525.

of frequent supply shocks of great amplitude. ***Second,*** economy being a complex system with interlinked and interrelated prices, a reasonable indexing of all prices to the satisfaction of all concerned is an extremely difficult task. ***Third,*** the government finds indexation politically undesirable because it does not control inflation, it rather creates a base for its perpetuation.

SUMMARY

- Although different economists have defined it differently, inflation generally defined as the considerable and persistent increase in the general price level.
- Inflation is measured by two methods: (*i*) percentage change wholesale *PIN* and (*ii*) percentage change *GNP* deflator, both over a period of two years. By using wholesale *PIN*, inflation is measured as $[(PIN_t - PIN_{t-1})/PIN_{t-1}] \times 100$. And, in case of *GNP* deflator method, *GNP* deflator is measured as *Nominal GNP/Real GNP* or as *GNP* at current price/ *GNP* at base year price, and inflation is measured as the percentage change in *GNP* deflator over a period of two years.
- On the basis of its rate, inflation is classified: (*i*) moderate inflation – one-digit price

rise, i.e., from 1-9 %, (*ii*) galloping inflation – two-digit inflation, i.e., from 10–99 percent, and (*iii*) hyper inflation – three-digit rate of inflation.

• Theories of inflation are generally classified under two categories: (*i*) Classical theory – the monetarist view, and (*ii*) modern theories.

• According to monetarist theory, inflation is anywhere anytime a monetary factor. In simple words, inflation is caused by increase in money supply at a rate higher than the rate of increase in output.

• Modern economists have offered four different theories of inflation: (*i*) demand-pull theory, (*ii*) cost-push theory, and (*iii*) demand-pull and cost-push interaction theory. In LDCs, inflation is caused also by structural imbalance, i.e., sectoral imbalance.

• According to the demand-pull theory, inflation is caused by increase in demand at a rate higher than the rate increase in supply of goods and services. Demand may increase because of (*i*) increase in money supply, (*ii*) increase in government spending, (*iii*) increase in consumer demand, and (*iv*) increase in investment, etc.

• According to cost-push theory, inflation is caused by increase in cost production. Cost of production may increase because of increase in wage rate, price of raw materials, and supply shocks.

• Economists, in general, hold the view that neither demand-pull nor cost-push alone causes inflation. It is, in fact, the interaction of both demand-pull and cost-push factors cause persistent inflation though it may begin with any of the factors.

• According to structuralist theory, inflation in LDCs is the result of their ambitious development plans which are generally accompanied by the structural imbalance reflected by food scarcity, input imbalance, foreign trade bottlenecks, etc.

• Inflation has both positive and negative effects on the economy. It has been found empirically that a moderate rate of inflation is good for economy. Inflation affects the different sections of the society in different ways. Fixed income groups like wage earners are affected aversely; producers gain from a moderate inflation as their profit margin increases; and lenders lose and borrowers gain. Inflation redistributes national income in favour of high-income groups.

• When inflation rate crosses the desirable limits, it affects the economic activities adversely. Therefore, the government is required to control inflation. In general, there are three methods of controlling inflation: (*i*) fiscal measures, (*ii*) monetary measures,

and (*iii*) direct price control. Under fiscal measures, the government cut down its expenditure and increases direct tax rate. These fiscal measures are intended to reduce aggregate demand. Monetary measures include (*a*) increasing bank rate, (*b*) open market operation, and (*c*) increasing CRR. These monetary measures are intended to reduce money supply from the banks.

REVIEW QUESTIONS

1. How is inflation defined? Can any rise in prices be considered as inflation? What is the acceptable or desirable limit of inflation?
2. What are the methods of measuring inflation? Why is national income deflator considered as a more reliable method of measuring inflation?
3. Explain the various kinds of inflation? How do they differ from one another?
4. Explain and distinguish between moderate, galloping and hyper inflation. How do these kinds of inflation affect economic growth of a country?
5. ‘A moderate degree of inflation is the logical concomitant of efficient mobilization’. Explain and examine the validity of the statement.
6. What are the effects of inflation on wage-earners, fixed income people, debtors and creditors, producers and the government? Give the reasons for the effects of inflation.
7. In what way does inflation contribute to economic growth? What kind of inflation affects economic growth adversely?
8. Explain the relationship between inflation and employment. Is achieving a high rate of employment by means of inflation always desirable?
9. What is monetarists’ explanation for inflation? Is inflation always and everywhere a monetary phenomenon?
10. “Inflation is always and everywhere a monetary phenomenon … and can be produced only by a more rapid increase in the quantity of money than in output”. Who said it? Do you agree with

this statement? Give reasons for your answer.

11. Explain how demand factors cause demand-pull inflation. What are the major weaknesses of the demand-pull theory of inflation?

12. What are the factors behind cost-push inflation? Is there any link between cost-push and demand-pull inflation?

13. Distinguish between demand-pull and cost-push inflation. Can the two types of inflation go hand-in-hand ? Explain in this regard the 'wage price spiral'.

14. "The distinction between demand-pull and cost-push inflation is unworkable, irrelevant or even meaningless". Who gave this statement? Do you agree with this statement? Why?

15. Theories of inflation based on conditions in and experience of developed countries cannot be and should not be applied straightaway to explain inflation in less developed countries. Do you agree with this statement? If not, why?

16. What is the 'structuralist view' on inflation? Explain the structural bottlenecks that are supposed to cause inflation in the developing countries.

17. Combating inflation has been one of the most intractable economic problems faced by the developed and underdeveloped countries. Comment.

18. What are the traditional monetary measures to control inflation? Explain how these measures work to control inflation.

19. Distinguish between fiscal measures and monetary measures of controlling inflation. What of the two policy measures are more effective under what conditions?

20. Explain the working of the monetary weapons of inflation control. Which of these weapons is more effective under what conditions?

21. What are the fiscal measures of controlling inflation? Are they more effective than monetary measures in controlling inflation?

FURTHER READING

Ackley, G., *Macroeconomics: Theory and Policy* (Macmillan, New York, 1978), Ch. 15.

BAumol, W. J. and A.S. Blinder, *Economics: Principles and Policy* (Harcourt Brace Jovannovich, Publishers, New York, 1988), 4th Ed., Ch. 6.

BronfenBrenner mArtin and frAnklin D. HolzmAn, "A Survey of Inflation Theory", in *Surveys of Monetary Theory,* Vol. I, 1965.

duesenBerry, J., "Mechanics of Inflation", *Review of Economics and Statistics,* May 1950.

dwivedi, D.N., *Macroeconomics: Theory and Policy* New Delhi (Tata McGraw-Hill), 3rd Ed., 2010, Chs. 21 and 22.

JoHnson, Harry G., "A Survey of Theory of Inflation", *Ind. Eco. Rev.*, Vol. VI, No. 4, August 1963.

lAidler, D.E.W. and J.M. PArkin, "Inflation—a Survey", *Economic Journal,* Dec. 1975.

APPENDIX

Annual Average Rate of Inflation in Different Countries

1970-80 and 1980-91

(*Percentage*)

Countries	*1970-80*	*1980-91*
High-Income Countries		
Australia	11.8	7.0
Canada	8.7	4.3
France	10.2	5.7
Germany	5.1	2.8
Italy		

15.6

9.5

Japan

8.5

2.5

Netherlands

7.9

1.8

Spain

16.1

8.9

UK

14.5

5.8

USA

7.5

4.2

Middle-Income Countries

Algeria

14.5

10.1

Argentina

133.9

416.9

Brazil

38.6

327.6

Colombia

22.3

25.0

Iran

22.4

13.8

Korea, South

20.1

5.6

Mexico

18.1

66.6

Peru

30.1

287.3

Philippines

13.3

14.6

Turkey

29.4

44.7

Venezuela

14.0

21.2

Yugoslavia (former)

18.4

123.0

Low-Income Countries

Bangladesh

20.8

9.3

Egypt	9.6	12.5
Ghana	35.2	40.9
India	**8.4**	**8.2**
Indonesia	21.5	8.5
Kenya	10.1	9.2
Nepal	8.5	9.1
Nigeria	15.2	18.1
Pakistan	13.4	7.0
Sri Lanka	12.3	11.2
Tanzania	14.1	25.7

Source: *World Economy and India's Place in It.,* Centre for Monitoring Indian Economy (CMIE) October 1993, Table 11.6.

Part VII

INTERNATIONAL ASPECTS OF BUSINESS MANAGEMENT

We have been so far concerned with managerial aspects of the internal, business issues and domestic business environment. In reality, however, most countries and many of their business organizations are involved in international business - foreign trade and investment. The managerial issues of international business are different and more complex compared to those in internal business management. The major issues related to international business are discussed in this Part of the book.

CHAPTER 32 Changing International Business Environment

CHAPTER OBJECTIVES

The objective of this Chapter is to have a look at the nature and extent of globalization of the world economy and its managerial implication. Specifically, by going through this chapter you will know the following aspects of international business environment.

- Meaning and extent of globalization of the world economy
- Consequences of globalization for an economy with large foreign trade
- Factors behind globalization and
- Managerial implications of globalization

32.1 INTRODUCTION

Till chapter 31, we have been concerned with the managerial problems of domestic business and have discussed how economic theories can be applied to find solution to domestic business problems. In modern times, however, most economies are open economies as they have a large proportion of their *GDP* consisting of international economic transactions.

A large number of multinational corporations have emerged in recent times. World trade, foreign investment, international flows of manpower have increased tremendously over the past three decades so much so that world economy is said to be *globalized*. The globalization affects the working and level of both industries dealing with domestic demand and industries having foreign links. This has changed the business environment of the world economy. Therefore, business corporations, especially those having foreign transactions, have to take into account the changing global business environment while making business decisions. In this part of the book, we move on to discuss some basic aspects of international economic transactions and their managerial implications. We begin our discussion by having a glance at the globalization of the world economy.

32.2 GLOBALIZATION OF THE WORLD ECONOMY

Globalization has been defined by the social scientists in many ways by taking in view the international changes in views, thoughts, ideas, products, culture, language, and social and political systems all over the world over a period of time. In economic sense, however,

globalization refers to growing economic interdependence and economic integration of the nations across the world due to a high rate of increase in international flows of goods and services, capital and labour, and technology and knowledge. Economic globalization comprises of integration of production, markets,' competition, technology, foreign investment, business corporations and industries. The two major factors that accelerated the process of globalization are (*i*) the advancements in the means of transportation, and (*ii*) telecommunication facilities – telephone and internet.

From our point of view, the most important aspects of globalizations are: (*i*) growth in world trade, (*ii*) growth in foreign investment, and (*iii*) international flow of manpower. These aspects are discussed here briefly.

32.2.1 Growth of World Trade

The high growth rate of international trade since 1980s till 2007 has been the main cause and speed of globalization. Although foreign trade has been going on since the time immemorial, it increased rapidly after the World War II. During the period from 1950 to 1990, world exports (at nominal prices) had increased at an annual average growth rate of 10.4 percent. During this period India's export had increased at an annual average growth rate of 7.1 percent. World trade increased more rapidly after the disintegration of the Soviet Union and economic liberalization in China in the early 1990s, though the growth of world trade remained limited to a small number of countries. In 1992, only 40 nations accounted for 94.4 percent of world export of which only 5 nations had a share of about 45 percent. The individual shares of countries having major share in world export in 1992 are as follows.

USA

– 12.1%

West Germany – 11.6 %

Japan

– 9.2%

France

– 6.4% and

UK

– 5.2%

Surprisingly, India with a share of 0.5 percent belonged to the 40 top nations.

Although most of world trade has remained confined to a small number of nations, there has been a high rate of globalization of the world economy. According to the WTO Report-2007, the volume of world trade increased twenty-seven fold from $296 million in 1950 to $ 7 trillion (i.e., $ 7,000,000 million) in 2005. According to GNI and foreign-trade data published in WB Development Report (2010), in year 2007, global merchandise exports alone accounted for 26.6 percent of the global GNI. Although the growth rate of the world trade started declining since 2008-09 due to global recession, it started reviving since 2010–11. According to the WTO economists, the world trade is estimated to increase by 2.5 per cent in 2013.

32.2.2 International Flow of Capital and Investment

Apart from tremendous increase in foreign trade, the *international flow of capital and investment* has contributed greatly to globalization. The international flow of private capital to developing countries increased rapidly during the 1990s mainly due to slowdown of growth in developed countries and growing investment opportunities in developing countries. The increase in investment by developed countries in developing economies

was, in fact, replacement of development aids and financial assistance into investment. In recent years, China and India are the two largest recipients of the foreign investment. Foreign investment takes forms of Direct Foreign Investment (FDI), Financial Institutional Investment (FII), banking deals, investment in share and bonds.

32.2.3 Growth in International Flow of Manpower

International flow of manpower is another contributor to the process of globalization. *Labour migration* has also increased unprecedentedly. By 2014, 232 million labour accounting for 3.1 percent of the global population had migrated internationally. Nearly half of the labour migrants (48%) consist of women seeking employment abroad. In 2013, migration of Asian labour accounted for the largest share of the international labour migrants'1.

32.3 FACTORS BEHIND GLOBALIZATION

As noted above, globalization is a multi-dimensional phenomenon. However, we point out here only economic factors that have facilitated the process of globalization.

1. Formation of International Organizations. After the Second World War, several international organizations were formed by different group of nations, which removed many trade barriers and created conditions for a freer trade between the nations of the group. The international organizations that were formed after the Second World War include (*i*) General Agreements on Tariffs and Trade (GATT), (*ii*) World Trade Organisation (WTO), (*iii*) International Monetary Funds (IMF), and United Nations Conference on Trade and Development (UNCTAD). The GATT, WTO and IMF not only minimized the trade barriers but also provided financial assistance to countries facing foreign exchange problems. These organizations created conditions for rapid increase in foreign trade. World exports had almost 'doubled from 8.5 % of total gross world product in 1970 to 16.2% in 2001'. Thus, by eliminating trade restrictions and barriers, international organizations contributed to globalization.

2. Formation of Custom Unions and Trade Associations. While GATT, WTO, IMF, and UNCTAD had a worldwide perspective, several regional and country-group organizations were formed over time to facilitate trade within the group members. The country-group unions include such trade unions as (*i*) European Union, (*ii*) North American Free Trade Agreements (NAFTA), (*iii*) Association of South-East Asian Nations (ASEAN), (*iv*) South Asian Association for Regional Cooperation (SAARC), (*v*) Preferential Trading Clubs, (*vi*) Free Trade Areas, (*vii*) Customs Unions, (*viii*) Common Markets, and (*ix*)) Economic Unions. These kinds of custom unions and trade associations removed barriers on the trade among themselves. Thus, formations of international organizations accelerated the process of growth in foreign trade and globalization.

3. Establishment of Free Trade Zones. In addition to formation of *trade unions* and *trade associations,* some other kinds of free trade areas were established during the last quarter of the 20th century. A free trade zone is a free trade block

1. Incidentally, the largest labour migration takes place currently from Mexico to the US.

consisting of some member countries which have signed the free trade agreements with agreement to eliminate tariffs and import quotas. The establishments of this category include Special Economic Zones (SEZ), Free Trade Zones (FTZ), Export Processing Zones (EPZ), Free Port, Urban Enterprise Zones, etc. The basic purpose of these zones is to increase trade and direct foreign investment within the nations of the zones. Specifically, the purpose is to promote international business via multinational corporations

32.4 CONSEQUENCES OF GLOBALIZATION

There have been certain far-reaching consequences of globalization. Some of the major economic consequences are described here briefly.

(*i*) *Growth of Foreign Trade*. Due to removal of restrictions on foreign trade or, what is called, foreign trade liberalization, world trade has expanded at a very high rate in terms of both volume and direction of trade. In terms of volume, the value of merchandise export increased from $ 3311 billion in 1990 to $ 6350 million in 2000 and to $ 9124 million in 2004. The annual average growth rate of merchandise export has risen from 9.2 per cent during 1990–2000 to 10.2 per cent during 2000–2004. Merchandise export had been increasing in the past also, almost at the same rate. But what makes the difference is the actual volume of exports. What is more important in assessing the growth in the world trade is that the percentage of world merchandise exports to the world *GDP*, i.e., world export-GDP ratio. The world export-GDP ratio increased from 10 per cent in 1990 to 20 per cent in 2000. In 2001, the trade-GDP ratio was 38 per cent in high income countries and 49 per cent in middle income countries.

(ii) Growth in International Migration of Labour. Until globalization started, labour force was generally im mobile internationally. With globalization of the world economy, the international flow of labour has increased tremendously. Today labour with technical skill and blue-collar job abilities are mobile all over the world. According to the IMF Report, *World Economic Outlook* – 2007, as a result of entry of China and India into the global commodity and labour markets over the last two decades, global labour force has expanded nearly fourfold. In some countries, immigrants are growing at a very fast rate. For example, according to a survey conducted by World Wide Immigration and Consultancy Services (WWICS), Indian workforce in Canada accounted for nearly 30 per cent of the total i mmigration in 2006, which overtakes the Chinese immigrants. The sectors that have attracted Indian labour are IT, medical, engineering and hospitality sectors. Nearly 2.5 lakh immigrants settle in Canada every year.

International migration of labour, especially when it reaches a considerable proportion, causes cultural and social conflicts which create social and political problems also. While some countries are able to accommodate and manage the migrant population, some others are finding it difficult to control and manage the problems created by the migrant population. For example, Canada, where nearly one-quarter of the population consists of migrants is managing migrant population fairly well, but in Britain and the US, it is becoming a problem now.

(iii) Growth in International Flow of Capital. As regards the *international flow of capital,* foreign investment is moving freely across the boundaries of the nations, with some constraints,

of course, in the form of Foreign Direct Investment (FDI) and Foreign Institutional Investment (FII). The period from 2001 to 2006 has witnessed a high growth of FDI, though there was a considerable decline in FDI in 2003. The world FDI increased from about $ 1 trillion in 2001 to $ 1.2 trillion in 2006, i.e., a 20% growth over a period of five years. The post-2003 upsurge in FDI was in the emerging markets. The prominent ones of the emerging markets include China, India, Russia and East Asian Countries (including Thailand, Indonesia, Malaysia and Singapore). According to forecasts made by Colombia Center for International Investment: *World Investment Prospects – Boom or Backlash (2006)*, the global foreign investment is expected to rise to $ 1.4 trillion in 2010. However, most of the increase in FDI is expected to flow to the developed countries, mainly because of cross-border mergers and acquisitions (M&A)2. An important reason for FDI flowing to developed countries and also to China may be the high rate of profitability achieved by the firms of the developed countries.

Just to have an idea of importance of foreign investment, look at the percentage share of the total revenue and percentage share of the total profit of some major US firms in 1997. It can be seen from the data given in Table 32.1 that most of the major US firms (listed) have about 50 per cent sales abroad and a major part of their total profit comes from foreign countries.

Table 32.1 *The Share of Foreign Sales and Profits of Some Top US Firms*

Firms	*% of total sales*	*% of total profit*
Exxon	76.9	60.0
IBM	58.4	

61.0

Mobil

59.4

59.6

Texaco

55.6

50.4

Hewlett-Packard

55.5

64.3

Citicorp

62.2

69.6

Proctor and Gamble

48.8

36.3

American Intel Group

53.8

68.8

Intel

55.9

32.5

Motorola

45.2

93.9

Xerox

57.4

55.8

Coca Cola

65.5

75.1

Dow Chemical

56.3

40.5

Compaq Computer

45.4

38.9

Source: Forbes, July 27, 1998.

Comparable data are not readily available for the Indian firms and other countries. However, there are indications that flow of capital abroad in search for higher profit continues to take place. For example, foreign investment increased in India from $ 2 billion in 2000–01 to $ 4.7 billion in 2005–06.

2. Summary of the study reported in *TOI*, 18/09/2006.

32.5 MANAGERIAL IMPLICATIONS OF GLOBALIZATION

Globalization of the world economy has serious implications for managerial decision-making. The recent and on-going changes taking place in the world economy have important repercussions for business managers irrespective of whether their own organization is or is not involved in foreign trade or foreign investment. Some major areas of managerial problems that business managers are facing and are likely to face in future too are pointed out here.

(i) Facing International Competition. The most challenging task that managers, especially the managers of MNCs, are currently facing and would continue to face in future too is international competition. The most important factor that figures in competing with rival firms in both domestic and foreign markets is *price competition*. One of the methods being used to avoid competition is merger and acquisition. Although mergers and acquisitions result in reducing the number of competitors in the international market, the market structure changes and tends to take the form of monopolistic or oligopolistic competition which is tougher than perfect competition. Not only that, the units of the same MNCs in different countries would come under country-wise competition. Facing competition has become and would continue to remain a challenging task for the managers. What strategy is formulated to face competition successfully is a persistent issue for business managers. One of the common strategies to face competition is to cut cost. For instance, Nissan Motor Company of Japan has decided to cut 1,500 jobs for two reasons: (*i*) shrinking market share due to growing tough competition and (*ii*) cutting down the cost to be able retain the market share (*Times of India*, 30 April, 2007).

(ii) Formulation of New and Effective Business Strategy. With globalization making business conditions more and more complex and competition tougher, survival of an inefficient organization is always at risk. Therefore, formulation of a new business strategy in the fast changing world becomes one of the prime tasks of the business managers. Strategy formulation is a challenging task in the sense that it involves various parameters which vary from country to country and from time to time and go beyond the control of managers. The parameters include (*a*) political environment of a country, (*b*) economic policies of the country, particularly policies pertaining to foreign companies, and (*c*) social and cultural environment of the country. For example, in a country dominated by corrupt politicians and bureaucracy, operating business is a costly affair; food items of beef would not be accepted by Hindu community in India; McDonald has done better business by introducing many vegetarian preparations in its India retail chains because majority of population is vegetarian and vegetarianism is a growing trend.

(iii) Efficient Labour Management. Globalization has made labour laws flexible. But facing competition successfully, as noted above, is an essential condition for the survival of the firm. One of the methods of facing competition is keeping cost of production at competitive level, if not lower. Where labour accounts for 30-40 per cent of the total cost, keeping cost low means keeping labour cost low. Labour supply has two components – domestic and foreign. Where sufficient domestic labour supply is not available, labour has to be outsourced from abroad – most developed countries facing labour shortage

are outsourcing labour from developing countries. Besides, domestic labour cost is often influenced by the foreign labour cost. For example, high salaries offered by the MNCs to MBAs of IIMs of India have raised the managers' salary for the Indian firms. Indian firms have to increase the salary of managers just to retain their managerial manpower. Not only retaining manpower, maintaining and increasing labour efficiency to international level is an equally important managerial task. Under the conditions of globalization, therefore, efficient labour management becomes an important decision area.

(iv) Managing Technology Transfers. An important byproduct of globalization is transfer of technology from one country to another—from rich to rich country, from poor to poor country, from rich to poor and from poor to rich, depending on the need of a country. There was a time when transfer of technology meant technology transfer from developed to less developed countries. That is no more the case—it may be either way. Most technology transfers are, however, taking place from the countries making investment in foreign countries. Managerial problems that are related to technology transfer inclulde (*a*) choice of technology, (*b*) acquiring chosen technology, (*c*) determining terms of technology transfer, and (*d*) creating conditions for social and political acceptability of the technology. Specifically, the basic problem is choosing and adopting a technology which is cost effective and suitable for quality of labour available in the country of investment.

(v) Understanding the Basics of International Trade. For the firms involved in foreign trade, it is essential to understand the basics of international trade and related issues. Some of the important areas in which business managers need to have a clear understanding and operational ability are following:

- Basis of international trade: The theory of foreign trade
- Scope and prospects of international trade with different nations
- Foreign trade policy of different countries, restrictions and control measures
- Customs policy of different countries—import duty and export duties, if any, imposed by the government on traded goods
- Subsidies and direct and indirect financial or other supports provided by the governments of the competing nations

• Supply position of the traded goods and services
• Economic strength of the country and future prospects of the economy
• Balance of payments position of the nation.
• Foreign exchange market and foreign exchange rate policy of the country
• Country's cooperation with the international organizations like IMF, WTO, World Bank, etc. and following the rules and regulations set by them

In this part of the book, we shall be concerned with some of these issues in detail. We begin will with the basis of foreign trade discussed under the title 'Theories of International Trade'.

SUMMARY

• The world economy is globalized today. Globalization of the world economy has certain serious managerial implications for the managers of the economies having large scale international economic transactions.
• In economic sense, ***globalization*** means a high rate of international flows of goods and services, capital and labour, and technology and knowledge across the world. Economic globalization comprises of integration of production, markets, competition, technology, foreign investment, multi-national corporations and industries.
• Globalization of the world economy is reflected by (*i*) a high growth rate of foreign trade, (*ii*) a high international flow of capital and investment, and (*iv*) a high rate of international flow of labour. The two major factors that accelerated the process of globalization are (*i*) the advancements in means of transportation, and (*ii*) telecommunication facilities – telephone and internet.
• The environment for the globalization of the world economy was created after the Second World War by the establishment of such international organizations as GATT, WTO, World Bank, IMF. These organizations eliminated the trade barriers and restrictions and created conditions for free flow goods and services, capital and investment and also provided financial help for countries facing foreign financial problems.
• The other kinds of regional and inter-country blocks were formed to promote trade among the member countries. Such regional and block or zones include European Unions, NAFTA, SAARC, Custom Unions, SEZ, Free Trade Areas, common markets, etc.
• Globalization has some important managerial implications, especially for large corporations: (*i*) facing effectively the international competition, (*ii*) formulating some new business strategy, (*iii*) efficient resource management, particularly labour management, and (*iv*) adoption and management of new technology, and (*v*) adopting business strategy according to international business environment.
• Managerial implications of globalization necessitate having a clear understanding of logistics of foreign trade, foreign trade policy of own and different countries, market determination of exchange rate, and balance-of-payments accounting.

REVIEW QUESTIONS

1. In what sense is the global business environment changing? What factors are responsible for the change in global business environment?
2. What is meant by globalization? What factors have contributed to globalization? Has India benefited or lost due to globalization?
3. What consequences of globalization are obvious in the world economy? Is India not losing its managerial manpower to foreign countries?
4. What are the managerial implications of changing business environment of the world for the managers of MNCs?
5. What aspects of foreign trade need to be taken into account by the business managers?

CHAPTER

33 Theories of Foreign Trade

CHAPTER OBJECTIVES

The objective of this Chapter is to begin our discussion on the foreign aspects of business. We begin our discussion with the theories of foreign trade. Theories of foreign trade reveal the basis of foreign *trade. The following theories of foreign trade are discussed in this chapter.*

• The basis of international trade – why countries go foreign trade

• Adam Smith's theory of foreign trade – the absolute advantage theory of trade

• Ricardian theory of foreign trade – the comparative advantage theory of trade advantage

• Heckscher-Ohlin theory of foreign trade and

• Terms of trade – the rate of product exchange

33.1 INTRODUCTION

As noted above, the objective of this chapter is to discuss the various theories of foreign trade. The theories of foreign trade answer a fundamental question: What is the basis of foreign trade? Besides, theories of foreign trade answer also such questions as: Why do nations go for trade with other nations? What determines the volume and direction of trade between the nations? How are the gains from trade measured? The economists of different generations - right from Adam Smith, the father of economics, to modern economists - have developed different theories of foreign trade. However, given our limited objective of this book, we will confine our discussion to two classical theories and Heckscher-Ohlin theory, the so called modern theory of foreign trade. Finally, we will discuss the determination of the terms of trade, i.e., how product exchange rate between the trading partners is determined.

33.2 BASIS OF FOREIGN TRADE

Before we discuss the theories of foreign trade, let us first have a glance at the general *basis of trade*. In general, the basic factors that create conditions for foreign trade are the same as the basis of general economic behaviour of the people. The basic factors that create the conditions for foreign trade are as follows.

1. *Human wants, needs and requirements keep increasing endlessly.* People go for foreign trade because their wants and needs increase endlessly which they are not

able to meet at minimum cost with the resources available within their country. Therefore, they look for foreign goods and services.

2. *Resource endowments are insufficient and variant.* Both natural and man-made resources available to different countries are limited and variant in terms of productivity. While some countries, like India and China, are better endowed with manpower, natural resources, e.g., vast area of fertile and cultivable land, some countries like Arab countries, with large mineral deposits, some other countries are better endowed have capital and technology. This kind of resource endowment differentials cause cost differentials. But most countries aim at cost minimization.

3. *Resource endowment differentials cause cost differentials.* Given the resource endowments, production cost of a commodity varies from country to country. For example, cost of agricultural products in Saudi Arabia is much higher than that in India, and cost of petroleum production is much higher in India than that is Saudi Arabia. This kind of resource endowment differentabls cause cost differentials. Therefore, with the objective of cost minimization of economic welfare, these countries go for foreign trade.

4. *People of all countries aim at and work for maximizing their economic welfare.* The ultimate objective of all economic activities is the maximization of economic welfare. So is the case for foreign trade. In order to maximize economic growth and economic welfare, nations go for foreign trade.

Having described the basic purpose of foreign trade, we proceed now to discuss briefly the theories of foreign trade as propounded by the classical and modern economists. The foundation of foreign trade theories was laid down by Adam Smith. We begin our discussion on the theories of foreign trade with Adam Smith's *theory of absolute advantage*.

33.3 ADAM SMITH'S THEORY OF FOREIGN TRADE:

THEORY OF ABSOLUTE ADVANTAGE

According to Adam Smith, foreign trade take place because some countries have absolute cost advantage in production of some goods and absolute cost disadvantage in production of other goods. In simple words, according to Adam Smith, while cost of production of some goods in some countries is very low and in some countries, it is very high. The country having lower cost of production in a commodity has *absolute advantage* in that commodity. According to Adam Smith's theory of foreign trade, a country exports a commodity in which it has absolute cost advantage and it imports a commodity in which it has absolute disadvantage.

Adam Smith's theory of absolute advantage can be illustrated through a simple, hypothetical two-country and two-commodity example. Let us suppose that per quintal labour-cost of production of rice and jute in India and Bangladesh is given as in Table 33.1. As the table shows, India needs 30 man-hours to produce one quintal of rice whereas Bangladesh needs 50 man-hours. The cost of rice production in India is thus much lower than that in Bangladesh. India, therefore, has an absolute advantage in rice production. In case of jute production, while Bangladesh needs 20 man-hours to produce one quintal of jute, India needs 60 man-hours. Thus, Bangladesh has absolute advantage in jute production. Going by the theory of absolute advantage, India would specialize in rice production and import jute, and Bangladesh would specialize in jute production and import rice.

Table 33.1 *Per Quintal Labour Cost (Man-hour)*

Country	*Rice*	*Jute*
India	30	60
Bangladesh	50	20

According to Adam Smith, trade between the two countries will prove advantageous to both since both of them can avail, given their labour force, a larger quantity of both the commodities and at a lower cost. For example, let us suppose that India produces both rice and jute. She will have to sacrifice 2 quintals of rice to produce 1 quintal of jute and, in her domestic trade, the exchange rate will be 2 quintals of rice for 1 quintal of jute. If India specializes in rice production, she can get 1.67 quintals of jute from Bangladesh in exchange for 1 quintal of rice. Similarly, in their domestic trade, people in Bangladesh get only 0.4 quintal of rice in exchange for 1 quintal of jute whereas they can import 0.67 quintal of rice against 1 quintal of jute. These gains to both the countries will be available only if India specializes in rice production and Bangladesh in jute production and there is trade between them.

33.4 RICARDIAN THEORY OF COMPARATIVE ADVANTAGE

The theory of *absolute advantage* gives the impression that trade between any two countries can be possible and mutually gainful only if both the countries have *absolute advantage* in the production of at least one commodity and *absolute disadvantage* in the production of at least one commodity. That is, if a country has an absolute advantage in the production of both the commodities and the other country has an absolute disadvantage in the production of both the commodities, there is no basis for trade between them. However, David Ricardo formulated a trade theory, known as the theory of comparative advantage, to prove that foreign trade can take place under these conditions also. The *Ricardian theory of comparative advantage* [1] states that there is a possibility of gainful trade between two countries even if one has absolute advantage in the production of both the commodities and the other country has absolute disadvantage in production of both the commodities provided both the countries have comparative cost advantage. His theory provides a powerful argument that so long as countries have comparative advantage in the production of commodities, specialization and trade between them would always be possible and advantageous to all of them. For example, let us suppose that in our two-country-two-commodity model, India is more efficient in producing both the goods, rice and jute, as shown in Table 33.2.

Table 33.2 *Per Quintal Labour Cost (Man-hour)*

Country	*Rice*	*Jute*
India	30	60
Bangladesh		

50

80

As the table shows, India can produce both the goods more efficiently, i.e., at a lower cost, compared to Bangladesh. But, her own relative efficiency is evidently

1. According to some economists, the idea of comparative advantage was originated by R. Terrens in his *Essay on the External Corn Trade* (J. Hanchard, London, 1815).

greater in rice production because her cost of rice production is just half her cost of jute production. India has *comparative advantage* in rice production because she needs only (30/50) 100 = 60 per cent of the cost of rice production in Bangladesh. But India has *comparative disadvantage* in jute production because her cost of jute production is twice her cost of rice production. On the other hand, Bangladesh has comparative advantage in jute production because her relative cost of jute production (i.e., 80/50) is less than India's (i.e., 60/30). In other words, if India shifts 30 man-hours from rice to jute production, she gains only 30/60 = 0.5 quintal of jute, whereas if Bangladesh shifts 50 man-hours from rice to jute production she gains 50/80 = 0.625 quintal of jute. Thus, India has comparative advantage in rice production and Bangladesh in jute production. Therefore, if India specializes in rice production and Bangladesh in jute production and they trade their surplus, both stand to gain.

33.4.1 Gain from Foreign Trade

To show the gains from trade, let us begin by assuming that there is no trade between India and Bangladesh. Both countries produce both the goods—rice and jute—and consume their total produce. In that case, given the cost of production in Table 33.2 the domestic exchange or barter rate in India and Bangladesh will be as follows:

(*i*) In India

30

1 Qtl. of rice = 60 = 0.5 Qtl. of jute

or

1 Qtl. of jute = 60 = 2.0 Qtl. of rice

30

(*ii*) In Bangladesh

50

1 Qtl. of rice =

= 0.625 Qtl. of jute

80

50

or

1 Qtl. of jute =

= 1.6 Qtl. of rice

80

These internal exchange rates are summarized in Table 33.3.

Table 33.3 *Internal Exchange Rates*

(*Quintals*)

India	*Bangladesh*
Rice = Jute	Rice = Jute
1 = 0.5	1.0 = 0.626
2 = 1.0	1.6 = 1.000

If there is no trade between India and Bangladesh, internal trade will be carried out on the basis of the exchange rates given in Table 33.3.

Let us now suppose that India specializes in rice production and Bangladesh in jute production, and they trade their surplus produce between them. Now, whether the two countries gain from foreign trade or not depends on whether or not their external exchange rate, i.e., terms of trade2, is higher than the internal exchange rate. If external

2. The concept and determination of 'terms of trade' will be discussed in the following section.

exchange rate is higher than the internal exchange rate, the two countries definitely gain from foreign trade and *vice versa*.

Given this criterion, both the countries gain through specialization and foreign trade. This can be proved. In the internal market of India, 1 quintal of rice is bartered for only 0.5 quintal (500 kg) of jute whereas by exporting 1 quintal of rice to Bangladesh, India can get 625 kg of jute. Similarly, in the internal market of Bangladesh, 1 quintal of jute is exchanged for 1.6 quintals of rice whereas by exporting 1 quintal of jute to India, she can obtain 2 quintals of rice. Thus, both the countries gain from specialization and trade between them. Going by the same logic, if India specializes in jute and Bangladesh in rice, and they exchange their produce, both countries will be worse off. The obvious conclusion, therefore, is that India would specialize in rice production and import jute, and Bangladesh would specialize in jute production and import rice.

Distribution of Gains Which country will gain more from foreign trade—India or Bangladesh[3]? The answer to this question depends on the determination of commodity exchange rate between the two countries. The gainful exchange rate for India ranges between 500 kg and 625 kg of jute for 1 quintal of rice, and for Bangladesh, it ranges between 1.6 quintals and 2 quintals of rice for 1 quintal of jute. Given the lower and upper limits of possible exchange or barter rates, a country's gain depends on the exchange rate in the range of exchange ratios. For example, if India exports 1 quintal of rice for 625 kg of jute, or imports 1 quintal of jute for less than 2 quintals of rice from Bangladesh, India gains more. And, if Bangladesh exports 1 quintal of jute for 2 quintals of rice and imports 1 quintal of rice for less than 625 kg of the jute, Bangladesh gains more. If exchange rates in foreign trade are the same as the internal exchange rate, both the countries gain fully from foreign trade. Gains are distributed between them on the basis of their production efficiency. No country gains at the cost of the other.

33.4.2 Critical Appraisal of Ricardian Theory

Ricardo's theory of comparative advantage was based on certain simplifying assumptions. Although it has still retained its importance as the basic theory of trade, it has been severely criticized over time for its simplifying assumptions and the consequent shortcomings. In the process of criticism, many improvements have also been suggested. The main points of criticism of Ricardian theory of comparative advantage are mentioned here briefly.

1. *Labour is not homogenous.* Ricardo assumed homogeneity of labour and he considered it as an 'approximation of reality'. But this is an unrealistic assumption. Labour is not homogenous throughout the world. It varies in its skill and productivity. Due to a high degree of specialization, labour is not interchangeable. Therefore, wage differentials are quite likely in the short-run, which affect both domestic and external rate of exchange between commodities.

2. *Labour is not the only factor.* Even if one assumes that labour is homogenous and wages are uniform, labour is not the only factor of production. There are other factors of production, viz., land and capital. The factor combination varies from industry to indsustry depending on the state of technology in a country. The production of some goods may be highly capital-intensive resulting in larger capital content in its cost. On the other hand, production of some goods may be labour-intensive, with larger labour content in the cost. Under such conditions, the labour theory of comparative advantage would not apply.

3. This question will be taken up again, in its general form in the next section.

3. *Demand-side ignored.* The theory of comparative advantage concentrates only on the supply side of the trade. It suggests that so long as there is comparative advantage in trade, two countries will trade their product between them. It also suggests that exchange rates, i.e., price ratios, based on comparative advantage would be beneficial to the trading partners. This theory, however, does not answer two important questions: (*i*) how are the price ratios determined between trading partners, and (*ii*) what quantities would be traded.

4. *Other criticisms.* The theory of comparative advantage has been criticized on many other grounds also: (*i*) its assumption that factors are perfectly mobile is wrong; (*ii*) it assumes away the cost of transportation, and (*iii*) its assumption of existence of perfect competition is unrealistic.

33.4.3 Improvements Suggested in Ricardian Trade Theory

As noted above, economists criticised Ricardian theory of trade on several grounds. Neo-classical economists suggested the following improvements in Ricardo's theory.

1. Introduction of Opportunity Cost. Ricardian theory has been criticised on the ground that it is based on only one production factor (labour) whereas production of a good involves both labour and capital and that labour cost alone does not represent the actual cost of production. Marshall attempted to resolve this problem of Ricardian theory by expressing the product value in terms of 'bales', i.e., in terms of basket of goods and services in exchange. But it did not prove to be a good improvement. Haberler suggested another modification by introducing the concept of *opportunity cost* to the Ricardian theory of trade. According to him, cost of production must be expressed in terms of opportunity cost, i.e., in terms of loss of one product due to production of another product. But this modification too did not complete the Ricardian Theory.

2. Introduction of Reciprocal Demand and Offer Curves. Ricardian theory has also been criticised for not taking into account the demand side of the trade. To complete this deficiency, J.S. Mill introduced the concept of ***reciprocal demand,*** i.e., demand for the product of other country in exchange for domestic product. Later on, Edgeworth and Marshall converted the concept of reciprocal demand into 'offer curve'. The *offer curve* s of the two countries are used to determine the quantum of demand from the two countries and the prices of the products. These contributions remain an important aspect of the theory of comparative advantage.

33.5 HECKSCHER-OHLIN THEORY OF TRADE

The Ricardian theory of comparative advantage states that the basis of international trades is the comparative advantage in cost of production. This theory, however, does not answer the question: Why does a nation have comparative advantage in the production of a commodity and comparative disadvantage in the production of another? It does not also say why the production possibility curves of any two nations differ. Heckscher4, a Swedish economist, attempted to explain why some countries enjoy comparative advantage

4. Heckscher Eli, "The Effect of Foreign Trade on the Distribution of Income". *Economisk Jidskrift*, 1919, reprinted in Hollis S. Ellis and Lloyd. A. Metzler (eds.), *Readings in the Theory of International Trade* (Blakiston Co., Philadelphia, 21 (1959), Ch. XIII.

in the production of some goods and how it forms the basis of international trade. His student Bertil Ohlin5, another Swedish economist, developed Heckscher's theory further in 1933. The theory of trade expounded by Heckscher and Ohlin is popularly known as the *Heckscher-Ohlin theory of trade,* also referred to as *factor-endowment theory of trade* or *the modern theory of trade*.

The Heckscher-Ohlin theory of trade states that comparative advantage in the cost of production arises exclusively by the differences in the *factor endowment* of the nations. In a general sense of the term, *factor endowment* refers to the overall availability of usable resources including both natureal resources and man-made means of production. However, in the exposition of the trade theory, only the two most important factors—labour and capital—are taken into account. The factor endowments of different countries are different. While some countries have abundance of labour, some countries have abundance of capital. The labour-abundant countries have comparative cost advantage in the production of goods which require a labour-intensive technology and capital-abundant countries have comparative cost advantage in the production of goods that need capital-intensive technology. For example, India and China, the two labour-abundant countries produce and export huge quantities of shoes and garments because these goods are generally produced with labour-intensive technology. Capital-abundant countries like the United States and Japan produce and export capital intensive goods like automobiles.

The Heckscher-Ohlin theory of trade is based on the comparative cost advantage due to factor-endowment differentials. This theory of trade has been developed in the form of two theorems, which are discussed below.

33.5.1 Heckscher-Ohlin Theorems

Theorem I – Heckscher-Ohlin Trade Theorem : A country tends to specialize in the export of a commodity whose production requires intensive use of its abundant resources and imports a commodity whose production requires intensive use of its scarce resources.

Theorem II – Factor-Price Equalization Theorem : The international trade equalizes the factor prices between the trading nations. In the absence of foreign trade, it is quite likely that factor prices are different in different countries. The Heckscher-Ohlin theorem

II postulates that foreign trade eliminates the factor price differentials.

The construction of the Heckscher-Ohlin theory and theoretical proofs of the two theorems are based on the following model and its assumptions.

33.5.2 The Model and the Assumptions

The Heckscher-Ohlin theory assumes a model of two countries (*A* and *B*), two commodities (*X* and *Y*) and two factors, labour (*L*) and capital (*K*) with the following assumptions:

1. There is perfect competition in both product and factor markets in both the countries.
2. Factors—labour and capital—are fully mobile between the industries *X* and *Y* within a country but completely immobile between the countries.
3. Factors (*L* and *K*) of both the countries (*A* and *B*) are homogenous.

5. Ohlin Bertil, *Interregional and International trade* (Harvard University Press, Cambridge, 1933). Incidentally, Ohlin was awarded the Nobel Prize in Economics in 1977.

4. Factor supply is given and factors are fully employed.
5. Production functions for both the goods and in both the countries are homogenous of degree one, i.e., production function gives constant returns to scale.
6. Production technology for commodity *X* is labour-intensive and for *Y* it is capital-intensive.
7. Production possibility curves are concave to the origin.
8. The demand conditions for goods *X* and *Y* are identical in countries *A* and *B*, i.e., the two countries have identical indifference curves.
9. There is no transportation cost, nor is there any trade barrier.

33.5.3 Some Definitions

Before we explain the Heckscher-Ohlin theory of trade, let us know the meaning of *factor abundance* and the *production possibility curve*.

Factor Abundance *Factor abundance* is defined in terms of (*i*) factor-ratio criterion, i.e., labour-capital (*L*/ *K*) ratio or capital-labour (*K*/ *L*) ratio, and (*ii*) factor-price ratio criterion, i.e., *PL*/ *PK* ratio or *PK*/ *PL* ratio (where *PL* = price of labour and *PK* = price of capital, the capital rental).

According to the ***factor ratio criterion,*** country *A* is labour-abundant compared to country *B* if the ratio of total labour supply to total capital supply (*L*/ *K*) in country *A* is greater than that in country *B*. That is, country *A* is labour-abundat if

$$\frac{L^A}{K^A} > \frac{L^B}{K^B}$$

According to ***factor-price ratio criterion,*** country *A* is labour-abundant compared to country *B* if the ratio of labour price to capital price (*PL*/ *PK*) in country *A* is lower than that in country *B*. That is, country *A* is labour-abundat if

$$\frac{P_L^A}{P_K^A} < \frac{P_L^B}{P_K^B}$$

where *PK* = rental price of capital, *PL* = price of labour (wages), and superscripts *A* and

B refer to countries *A* and *B*, respectively.

Having defined factor-abundance, let us now look at the effect of factor abundance on the production possibility curves of the two countries with respect to two commodities.

Factor Endowment and Production Possibility Curve Production possibility frontier (*PPF*) of a country is determined on the basis of its factor endowment. To illustrate how factor-abundance determines the production possibility curves (*PPCs*) of the two countries, we assume:

(*i*) country *A* is labour-abundant and country *B* is capital-abundant,

(*ii*) production of commodity *X* is labour-intensive and that of commodity *Y* is capital-intensive and

(*iii*) production function for both the goods (*X* and *Y*) reveals diminishing returns to both the inputs (*L* and *K*).

Given these conditions, the *PPCs* of the two countries with respect to commodities *X* and *Y* are illustrated in Fig. 33.1. The curve *AA¢* is the *PPC* of labour-abundant country *A*. The shape of

the curve *AA¢* shows that country *A* can produce a larger quantity of labour-intensive commodity *X* than of capital-intensive commodity *Y*.

Similarly, the curve *BB¢* is the *PPC* of capital-abundant country *B*. The shape of the curve *BB¢* shows that country *B* can produce a larger quantity of capital-intensive commodity *Y* than that of labour-intensive commodity *X*.

33.5.4 Production and Consumption in Autarky

Before we explain the Heckscher-Ohlin theory of trade, let us take a view of production and consumption in a country under the

Fig. 33.1 *Production Possibility Curves*

condition of autarky – the country having no foreign trade. The production and consumption patterns in the two countries under the condition of autarky are illustrated in Fig. 33.2. *Autarky* is a situation of self-sufficiency in production and consumption without trade. To find the equilibrium level of product-mix and consumption under autarky, let us introduce the indifference curves and factor-price lines in Fig. 33.2. The curves *IC* 1 and *IC* 2 represent the identical indifference curves of the two countries (assumption 8). The lines *JK* and *LM* are the factor-price lines of countries *A* and *B*, respectively, assuming constant factor prices. The equilibrium level of product-mix and consumption under autarky in both the countries can be found by reading *PPCs* in conjunction with the indifference curves and factor-prices lines.

Fig. 33.2 *Production Possibility Curves and Equilibrium under Autarky*

As the figure shows, the indifference curve *IC* 1 and *A*'s factor-price line *LM* are tangent to its production possibility curve *AA¢* at point *EA*. The slope of the *PPC* curve *AA¢* gives the *marginal rate of transformation* between *X* and *Y* (i.e., *MRTx,y*); the slope of the indifference curve *IC* 1 gives the *marginal rate of substitution* (i.e., *MRSx,y*); and the slope of the line *JK* gives the relative factor price (i.e., *P A*

A

K / PL)6. At the point of tangency, therefore,

MRT A

A

A

x, y = MRS Ax,y = PK / PL

Therefore, point *EA* is the point of equilibrium for country *A* under autarky. At autarky equilibrium, country *A* produces and consumes *OX* 2 of commodity *X* and *OY* 1 of commodity *Y*.

Similarly, indifference curve *IC* 1 and *B*'s factor-price line *JK* are tangent to its production possibility curve *BB¢* at point *EB*. At point *EB*, therefore,

MRTB

B

B

x,y = MRSBx,y = PK / PL

Therefore, point *EB* is the point of equilibrium for country *B* under autarky. At autarky equilibrium, country *B* produces and consumes *OX* 1 of commodity *X* and *OY* 2 of commodity *Y*.

The conclusions that emerge from Fig. 33.2 can now be summed up as follows:

(*i*) Under autarky, country *A* produces and consumes *OX* 2 of commodity *X* and *OY* 1 of commodity *Y* and country *B* produces and consumes *OX* 1 of commodity *X* and *OY* 2 of commodity *Y*.

(*ii*) A labour-abundant country *A* produces and consumes more of labour-intensive commodity *X* and a capital-abundant country *B* produces and consumes more of a capital-intensive commodity *Y*.

(*iii*) Both the countries maximize their social welfare at indifference *IC* 1 in the absence of trade to trading countries.

Having described the equilibrium conditions under autarky, let us now explain the Heckscher-Ohlin theory of trade in terms of the two Heckscher-Ohlin theorems and show the gains from trade.

33.5.5 Heckscher-Ohlin Theorem–I: Trade Leads to Specialization

As noted above, the Heckscher-Ohlin Trade Theorem I states that *a country specializes in the export of a commodity whose production requires intensive use of its abundant (low-cost) resources and imports a commodity whose production requires intensive use of its scarce (costly) resources.* This theorem is illustrated in Fig. 33.3. To begin with, let us assume that there is no trade between countries *A* and *B* and that country *A* is in equilibrium at point *EA* and country *B* at point *EB*. When trade commences between them, the labour-abundant country *A* tends to specialize in the production and export of labour-intensive commodity *X* because it can produce this commodity at a relatively lower cost. As a result, demand for labour increases and demand for capital decreases. This kind of change in factor demand causes a change in the relative factor prices. The price of labour (*PL*) increases and price of capital (*PK*) decreases. Consequently, the factor-price line *LM* rotates clockwise: it becomes steeper. As a result, *A*'s equilibrium point *EA* shifts rightwards till it reaches oint *T*.

6. The relative factor price can also be expressed as *r*/ *w* where *r* is capital rental and *w* is wage rate.

+

Fig. 33.3 *Heckscher-Ohlin Theorems: Specialization and Gains from Trade*

Similarly, when there is trade between the two nations, the capital-abundant country *B* specializes in the production of the capital-intensive commodity *Y*. As a result, demand for capital increases and demand for labour decreases. Consequently, price of capital (*PK*) increases and price of labour (*PL*) decreases. This change in relative factor price changes the slope of the factor-price line *JK* and brings it to the position of line *JK*. As a result, *B*' s equilibrium point *EB* shifts upwards along the *PPC* curve *BB*¢ till it reaches point *R*.

This analysis establishes the Heckscher-Ohlin Theorem I that *a capital-abundant country specializes in capital-intensive goods and a labour abundant country specializes in labour-intensive goods.*

The greater the degree of specialization, the larger the volume of trade and gains from trade to the two countries. Therefore, specialization will continue to increase until the rates of transformation (i.e., slopes of possibility curves) are the same in both the countries. This position is shown by the line *DN* which is tangent to country *A*' s production possibility curve *AA*¢ at point *T* and at point *R* to the production possibility curve *BB*¢ of country *B*. At points *T* and *R*, slopes of the two production possibility curves are equal. It means that at points *R* and *T*, rates of transformation are the same in both the countries.7

Gains from Trade

As regards the gains from the trade, both the countries – *A* and *B* – gain from trade between them. The gain from trade is indicated by the movement of both the countries

7. This analysis however holds only if community indifference curves for both the countries are assumed to be identical. If indifference curves are assumed to the different or intersecting, specialization and trade may not be possible.

to a higher indifference curve *IC* 2. Thus, trade increases the availability of both the goods (*X* and *Y*) to both the countries, though the demand pattern changes.

The trade pattern is shown by the triangles *EFT* and *RCE*. Country *A* will export *FT* of *X* to country *B* and import *EF* of *Y*. Country *B* will export *RC* of *Y* and import *CE* of *X*. *Under the simplifying assumptions made above, production and trade will be in the state of a general equilibrium.*

33.5.6 Factor Price Equalization Theorem: Theorem II

Heckscher-Ohlin-Samuelson Theorem. The factor price equalization theorem is in fact a corollary to the Hecksher-Ohlin trade theory. It holds only so long as Heckscher-Ohlin theorem holds. It was Paul Samuelson who provided a rigorous proof of the Heckcher-Ohlin theorem of factor-price equalization. For this reason, the above theorem is also referred to as the *Heckscher-Ohlin-Samuleson theorem*. In this section, we have presented very briefly presented the basic idea of the theorem8.

The factor price equalization theorem states that international trade brings about equalisation in the absolute and relative returns to homogeneous factors of production and of their prices. In other words, the wages of homogeneous labour and rentals to homogeneous capital will be the same in all the nations trading between themselves.

Figure 33.4 illustrates the factor-price equalization theorem in our 2-country (*A* and *B*), 2-commodity (*X* and *Y*) and 2-factor (*K* and *L*) model with our earlier assumptions that country *A* is labour-abundant and country *B* is capital-abundant and commodity *X* is labour-intensive and commodity *Y* is capital-intensive. Lines *KALA* and *KBLB* are pre-trade isocosts of countries *A* and *B,* respectively. The slope of the isocost lines gives the factor price ratio (w/ *r*). Since isocost *KALA* and *KBLB* have different slopes, the factor-price ratios are different in countries *A* and *B*.

Fig 33.4 *Factor-Price Equalization*

8. For details and rigorous proof, see, P.A. Samuelson, "International Trade and Equalisation of Factor Prices, *Eco*., *JL*., June 1948, and his "International Factor Price Equalization—Once Again" *Eco. JL*., June 1949, Reprinted in Jagdish Bhagwati (ed.), 1981, *International Trade: Selected Readings,* (Cambridge, Massachusets, MIT Press). For a good summary see, Sodersten, B., *International Economics*, Chapter 4.

Let us now show that factor-prices tend to be equal in country *A* and *B* with trade and specialisation. Suppose that before trade, country *A* produced at point *E* and country *B* at point *C*. When country *A* increases its specialisation in production of the labour-intensive commodity *X*, its demand for labour increases, making price of labour (*w*) to increase in relation to capital cost (*r*). As a result, point *KA* shifts towards point *KC* and point *LA* shifts towards point *LC*. This process continues until *KA LA* reaches the position of isocost *KC LC*. Similarly, when country *B* increases its specialization in the production of capital-intensive commodity *Y,* its demand for capital increases, causing wage rates(*w*) to decrease in relation to *r*, the interst rate. Consequently, point *LB* shifts towards point *LC* and point *KB* shifts up towards point *KC*. This process continues until *KBLB* reaches the position of *KCLC*.

Note that both the nations *A* and *B* reach finally the same isocost *KCLC* which is tangent to isoquant *X* at point *F* and to isoquant *Y* at point *D*. At these points, the slope of *KCLC* is the same. Therefore, factor-price is the same in both the nations. It means that the factor prices are equalized.

International trade not only reduces the international differences in the factor prices, but also it brings about a complete equalization in relative factor prices. This happens because so long as there is difference in the relative factor prices, relative prices of commodities will differ and trade between countries will continue to increase. As trade increases, the difference in relative factor prices and, therefore, in relative commodity prices will decrease. When relative commodity prices tend to equalise between the nations, the comparative advantage tends to disappear. Ultimately, both relative factor prices and commodity prices come in complete equality between the nations.

In the final analysis, when general equilibrium is attained across the countries, then the following conditions hold.

(*i*) Factor-ratios are equalsed when

L

L

A

B

=

K

K

A

B

This condition is fulfilled at points *D* and *F* simultaneously,

(*ii*) Factor-price ratios are equalied whe

A

B

P

P

K

K

=

A

B

P

P

L

L

r

r

or

in *A* =

w

w in *B*, and

This condition is also fulfilled where condition (*i*) is fulfilled.

(*iii*) Relative commodity-prices are equalsed when

A

B

P

P

X

X

=

A

B

P

P

Y

Y

Condition (*iii*) is satisfied when condition (*i*) and (*ii*) are satisfied.

These conclusions hold under the following assumptions:

(*a*) there is perfect competition in both factor and commodity markets of both the countries;

(*b*) there is free trade between the nations; and

(*c*) all the nations use the same technology with constant returns to scale.

If these conditions hold, then international trade makes not only the relative factor prices and commodity prices to equalise, but also the absolute prices of factors and commodities, between the trading nations.

International trade thus serves as a substitute for the international mobility of factors of production. The factors of production may move physically from one nation to another. But trade will bring their returns in equality with one another as if factors of production were physically and freely mobile between the nations. In this lies the main significance of factor-price equalization theorem.

33.5.7 Criticism of Heckscher-Ohlin Theory

Heckscher-Ohlin theory has been criticized and challenged on conceptual, methodological and empirical grounds. Here we will summarize the criticism of Heckscher-Ohlin theory based on conceptual and methodological grounds.

First, Heckscher-Ohlin theory assumes identical quality of factors such that factor endowments of different countries may be measured in homogeneous units. In reality, however, *factors of production are not homogeneous or of identical quality.* Factor endowments vastly differ in quality and variety. This causes a real problem in the measurement and comparison of factor endowments of the various Countries.

Secondly, Heckscher-Ohlin model relies on the assumption that difference in relative factor prices reflects the difference in relative factor endowments, i.e., relative supply position of factors. This means that, in the determination of factor prices, supply outweighs the demand. That is, factor-abundance in physical terms of the same as factor-abundance in the economic sense. But, if demand for factor to be equally important in factor price determination, then the relative factor prices would not reflect the relative factor endowments. For, interest and wages would tend to equalise only under the conditions of perfect competition (which is one of the assumptions of Heckscher-Ohlin model). If it is so, labour-abundance in the physical sense would not mean lower wages in relation to interest, and scarcity of capital would not mean higher interest relative to wages.

Thirdly, Heckscher-Ohlin model assumes also that production technology for both goods is the same in the two countries. That is, isoquant for a given commodity is the same in both the countries. This is, however, not necessary. Besides, even if this assumption is dropped, it can be shown that comparative advantage exists and trade takes place, but not in correspondence to Heckscher-Ohlin proposition.9

Fourthly, it has also been shown that if production technique is such that isoquants intersect twice, then one good will be capital-intensive over certain range of factor prices, and labour-intensive over another. That is, factor intensity is reversed. This is *known as*

9. For proof, see, Ford, J., The *Ohlin-Heckshcer Theory of Commodity Trade* (Asia Publishing House, London, 1965). A summary of his argument is available in David Young, *International Economics* (Intertext Books, London, 1969), pp. 30–34.

factor intensity reversal argument. The factor intensity reversal argument states that once the Heckscher-Ohlin assumption regarding the relation between 'factor-price structures and commodity factor intensities are reversed,' the probable trade pattern need not follow the basic theorem of Heckscher-Ohlin.

Fifthly, Heckscher-Ohlin theory assumes constant returns to scale in both the countries of the model and that commodity-prices reflect the factor-prices. But, if an industry in a country enjoys economies of scale, the commodity prices will be higher than the actual cost. It means that commodity prices do not reflect the cost of production.

Sixthly, Heckscher-Ohlin assumption of perfect competition is highly questionable. Nearly 50 per cent of the world trade in manufactured goods among the industrially advanced countries comprises differentiated products which is, of course, a feature of monopolistic competition.

Seventhly, it is also argued that, contrary to the assumption of no-transportation cost, prohibitive transportation cost has been found to affect the volume of trade considerably.

Finally, in an earlier criticism of Ohlin's theory, "Viner unleashed a barrage of citations to show that, at one time or another, most of Ohlin's choicest points have been made by the older writers." Donaldson's review concluded that Ohlin had achieved only an elaboration upon the classical theory: there is nothing new.

33.6 TERMS OF TRADE

The theory of comparative advantage, both its original or modified forms, suggests that if countries specialize in the production of the commodity in which they have comparative advantage and trade their surplus produce with the other countries, they will have the gains from foreign trade. But, the distribution of gains from trade depends on the 'terms of trade'. In other words, whether a country gains from trade more or less than the other country depends on the terms of trade. *In a general sense, terms of trade is defined as the quantity of domestic goods that must be given in exchange for one unit of imported goods.* In simple words, *terms of trade* refers to rate of exchange of traded goods.

A number of measures of terms of trade have, however, been suggested by the economists, viz., Mill, Marshall, Taussing, Ohlin, Haberler, Viner and Kindleberger. The multiplicity of measures (or formula) of terms of trade indicates the lack of unanimity among the economists on the measures of terms of trade. However, we discuss here the various concepts of terms of trade.

(a) *Net Barter Terms of Trade* It is the most common and widely used measure of terms of trade. The *net barter terms of trade*, (*Tn*) has been defined by Taussing and Vineras,

$$T_n = \frac{X_p}{M_p}$$

where *Xp* and *Mp* are prices of exports and imports, respectively. By comparing the *Xp*/ *Mp* in two different periods, one can find the trends in terms of trade.

(b) *The Gross Barter Term of Trade* The gross *barter terms of trade* was defined by Taussing and Viner as *Mq*/ *Xq*, i.e., the ratio of quantity of imports (*Mq*) to the quantity of exports (*Xq*).

When trade between two countries is balanced, the net barter terms of trade is equal to the gross barter terms of trade, i.e., when trade is alnced,

$$\frac{X_p}{M_p} = \frac{M_q}{X_q} \quad \text{or} \quad X_p \times X_q = M_p \times M_q$$

This equation has the following implications.

(*i*) If *Mp* and *Mq* are constant and *Xp* increases, terms of trade will be favourable to the exporting country since a smaller *Xq* will be required to balance the trade.

(*ii*) If *Mp* and *Mq* are constant and *Xp* decreases, terms of trade becomes unfavourable, since a larger *Xq* will be required to balance the trade.

(*iii*) If *Xp* and *Xq* remain constant and *Mp* increases, terms of trade become unfavourable, since larger imports will be required to balance the trade.

(*iv*) If *Xp* and *Xq* remain constant and *Mp* decreases, it is a favourable development in the terms of trade.

Kindleberger has, however, commented that gross barter terms of trade *Mq*/ *Xq*, is least informative "since they reflect less price movements than the changes in the balance of payments, and capital movements." He also suggests that when reference to 'terms of trade' is made without qualification, it should be taken for the net barter of trade, i.e., *Xp*/ *Mp*.

(c) *Income Terms of Trade* A serious limitation of net barter terms of trade is that its application is limited to the conditions of perfectly competitive system of international pricing system. Otherwise, if export prices are held constant, the volume of exports will fall. It will, no doubt, reflect a favourable terms of trade but at the cost of export earnings. To overcome this anomaly, at least at the theoretical level, another concept of terms of trade, called *income terms of trade*, has been suggested. The income terms of trade is defind as,

$$X_q \cdot \frac{X_p}{M_p} \quad \text{or} \quad \frac{X_p \cdot X_q}{M_p}$$

Since in case of balanced trade *Xp* · *Xq* = *Mp* · *Mq*, the income terms of trade (*Xq* · *Xp*)/ *Mp*, determines the volume of imports, *Mq*. This concept is also called the 'capacity to import'.

The income terms of trade is a concept superior to the net barter terms of trade for the purpose of less developed countries, though the latter is more frequently used despite its shortcomings. When the concept of net barter terms of trade is applied to more than one commodity, the ratio of *Xp* to *Mp* in the current year and the same in the base year are used. Th formula is

X 1 p X p 0

÷

X 1

p

X p 0

where subscripts 1 and 0 refer to the reference year and base year respectively.

(d) ***Single-Factor Terms of Trade*** Yet another concept of terms of trade suggested to modify the major defect of net barter terms of trade (*Xp*/ *Mp*), is *single-factor terms of trade*. Kindleberger calls it another 'monstrous piece of jargon.' It has been pointed out that *Xp*/ *Mp* does not take into account the change in efficiency and, hence, ignores its

consequence on the welfare of country to the extent it is based on foreign trade. For example, if export prices fall by 10 per cent due to a 15 per cent fall in cost of production on account of improvement in the efficiency, the exporting country is still better off. Therefore, an adjustment in *Xp*/ *Mp* should be made to account for improved efficiency. The formula uggested is

X p E

M

x

p

where *Ex* denotes the improved efficiency in the export sector, measured over a suitable base year.

(e) ***Double-Factor Terms of Trade*** The single-factor terms of trade is further modified to include the improved efficiency of factors in the country from which it imports. The concept of *double-factor terms of trade* i defined as

X p Ex

.

M p Em

The single-factor terms of trade is considered to be superior to the double-factor terms of trade for the purpose of a single country. But, in the analysis of overall gains from trade to the trading partners, the double-factor terms of trade would be preferable. The reader might have noted that the basic concept of terms of trade is the net barter terms of trade. All other concepts are only modifications in it to account for other factors affecting the terms of trade.

SUMMARY

- In general, the bases of foreign trade are (*i*) human wants and needs are growing endlessly, (*ii*) resource endowments are inadequate and variant, (*iii*) cost of production varies from country to country, and (*iv*) people of all countries aim at and work for maximization of their economic welfare. However, economists have built their trade theories on the basis of only cost differentials.
- The earliest theory of foreign trade was offered by Adam Smith, his theory known as 'theory of absolute advantage'. According to his theory, foreign trade between any two nations takes place because one nation has absolute cost advantage in producing one good and absolute cost disadvantage in another product.
- David Ricardo offered another theory of trade, known as 'comparative advantage theory of trade'. According to Ricardo, foreign trade can take place between any two nations even if one of them has absolute advantage in both the goods and the other country has absolute disadvantage in both the goods, provided they have 'comparative advantage' in cost of production. Comparative advantage means the domestic exchange rate is higher in terms of one good comparative to foreign exchange rate.
- Heckscher-Ohlin constructed another theory of trade, regarded as the modern theory of trade. According to Heckscher-Ohlin theory, trade takes place between nations because of difference in their resource endowments. While some nations are capital-abundant, some others are labour-abundant. As a result, capital abundant countries specialise in capital-intensive products and produce the goods at lower cost and labour-abundant countries specialize in labour-intensive goods with lower cost of

production. Thus, resource-endowment differentials cause cost differentials. This is the basis of gainful trade between the nations.

• The Heckscher-Ohlin theory of trade is stated in two theorems: (*i*) Theorem I: A country specializes in production and export of a commodity which requires intensive use of its abundant resources and vice versa; and (*ii*) Theorem II: Foreign trade equalizes the factor prices.

• As regards the 'terms of trade', it refers to the rate of exchange between the domestic and foreign goods. Economists have suggested different measures of terms of trade: (*i*) net barter terms of trade, (*ii*) gross barter terms of trade, (*iii*) income terms of trade, (*iv*) single terms of trade, and (*v*) double-terms of trade.

REVIEW QUESTIONS

1. What is the basis of international trade according to the classical theory of trade? Can there be trade between nations even if one of them has absolute advantage in all the goods?
2. Explain the comparative advantage theory of trade. How does it show an improvement over the absolute advantage theory? What are its shortcomings?
3. State and explain the Heckscher-Ohlin theorems. Under what conditions do they hold?
4. What is the factor price equalization theorem? What are the necessary conditions for this theorem to be valid?
5. Compare the classical and general equilibrium theory of international trade.
6. How does international trade result in commodity and factor-price equalization in the trading countries?
7. (*a*) Show that in a two country, two commodity and two factor world, trade can take place even if factor endowments in the two countries are identical.

(*b*) Under what conditions are factor prices equalized even if the factor endowments are different?
8. Show that international trade extends a country's consumption possibility beyond its production possibility frontiers.
9. Using the Heckscher-Ohlin model of trade, show that free trade is better than no trade. In this context, derive also a graphical measure of gains from trade.
10. Compare and contrast the meaning of comparative advantage in the Ricardian and Heckscher-Ohlin Trade Models. Give your comments on attempts made at empirically testing of the two theories.
11. How far does free trade lead to complete equalization of factor rewards? How do you then explain wide divergences in incomes in the world?
12. Using diagrams, write short notes on any two of the following:

(*i*) The possibility of trade between countries with same factor endowments and different tastes.

(*ii*) Gains from international trade.

(*iii*) The basis of trade between countries with identical factor endowments, identical tastes and increasing returns.
13. (*a*) What do you mean by factor reversibility? What effect will it have on factor price equalization?

(*b*) Give a geometric exposition of gains from trade.
14. State and explain the conditions under which the factor price equalization theorem of international trade holds. What would be the effect on tendency towards factor price equalization when there is complete specialization in the production of a particular commodity by one of the trading countries?

15. Given two countries and two goods, show that as long as both countries produce some amount of both goods, free trade will lead to complete factor price equalization. What happens when one country completely specializes in the production of one good?
16. Do you agree with the view that differences in factor endowments are neither a necessary nor a sufficient basis for international trade?
17. Assume that by using all its resources, nation *A* can produce 80 units of *X* and 40 units of *Y*. Comparable figures for nation *B* are 30 units of *X* and 30 units of *Y*. Will the countries specialize? Explain your answer and state your assumptions.
18. What are the necessary and sufficient conditions for Hickscher-Ohlin Samuelson's factor price equalization theorem to hold?
19. Country *A* is capital abundant comparedto contry *B* hen

L

L

L

L

(a) A

B

$>$

$<$

K

K

(b) A

B

A

B

K

K

A

B

(c) K

K

A

B

$<$

(d) None of the above

L

L

A

B

20. Country A is capital-rich if capital-labour price ratio in A compared to that in country B is

(a) smaller

(b) greater

(c) equal

(d) none of the above

21. Which of the following conditions must be fulfilled, according to Hekscher-Ohlin-Samuelson theorem, for there to be general equilibrium, across ountries A and B?

L

L

r

r

(a) A

B

$=$

(factor ratio)

(b) in A = in B (factor price ratio)

K

K

w

w

A

B

A

B

P

P

(c) $x = x$

A

B

P

P (commodity price ratio)

(d) All the above

y

y

[**Ans.** 19(*d*), 20(*b*), 21(*d*)]

FURTHER READING

Bhagwati, J., "The Pure Theory of International Trade: A Survey," *Eco. Jl.* , 1964.

ChaCholiades, M., *International Economics*, (McGraw Hill, New York, 1990), Chs. 2 and 4.

ChipMan, J.s., "A Survey of the Theory of International Trade," *Econometrica*, 1965.

dwivedi, d.n., *International Economics:Theory and Policy* (Vikas Publishing House, New Delhi, 2013), chs – 3 and 4.

heCksCher, e.F., "The Effect of Foreign Trade on the Distribution of Income," *Economisk Tidskrift*, 1919, Reprinted in H.S. Ellis and L.M. Metzler (eds.), *Readings in the Theory of International Trade,* (AEA, Harvard University Press, Cambridge, Mass.), 1933.

heller, roBert h., *International Trade: Theory and Empirical Evidence,* (Englewood Cliffs, Prentice Hall, 1973), Ch. 4.

Johnson, harry g., "Factor Endowments, International Trade and Factor Prices," *Manchester School of Economics and Social Studies,* September, 1957.

Jones, r.w., "*Factor Proportions and the Heckscher-Ohlin Theorem*", *Rev. of Eco. Stud.,* January 1956.

lanCaster, k., "The Heckscher-Ohlin Trade Model: A Geometric Treatment", Economica, February 1957, Reprinted in Jagdish Bhagwati (ed.), *International Trade: Selected Readings*, (Penguin, Baltimore, 1969).

leontieF, w.w., "Factors Proportions and the Structure of American Trade: Further Theoretical and Empirical Analysis," *Rev. of Eco. and Stat.,* November 1956.

leontieF, w.w., "Domestic Production and Foreign Trade: The American Position Re-examined",

Proceedings of the Am. Phil. Society, September 1953. Reprinted in Jagdish Bhagwati (ed.), *International Trade: Selected Readings*, (Penguin, Baltmore, 1969).

salvatore, d., *International Economics,* (Macmillan Publishing Co., New York, and Collier Macmillan, London, 1983), Chs. 5 and 6.

saMuelson, P.A., "International Trade and Equalization of Factor Prices", *Eco. Jl.* , June 1948.

saMuelson, P.A., "International Factor-Price Equalization-Once Again," *Eco. Jl.* , June 1949. Reprinted in Jagdish Bhagwati (ed.), *International Trade* (University Press, Cambridge, Mass, 1981).

sodersten, Bo., *International Economics,* (Second Edition, Macmillan 1980), Chs. 3 and 4. (A good text for beginners).

young, d., "*International Economics,* " (Intertext Books, London, 1969), Chs. 3 and 4.

CHAPTER

34 Foreign Trade Policy: Free Trade *vs*. Trade Control

CHAPTER OBJECTIVES

Having discussed trade theories in the preceding chapter, we move on to discuss the trade policy. The objective of this chapter is to discuss the purpose and consequences of two basic kinds of trade policies – the free trade policy and trade control policy. By going through this chapter, you will learn the following aspects of these kinds of trade policies.

• The meaning and purpose of free trade policy

• Arguments for and against free trade policy

• The meaning and objectives of protected trade policy

• Arguments for and limitations of trade control and

• Trade policy adopted by India

34.1 INTRODUCTION

In this chapter, we move on from *trade theory* to *trade policy*. Foreign trade policy refers to the directive principles and control measures that a country adopts to control and regulate its exports and imports. A foreign trade policy is a widely known controversial issue of trade policy, i.e., whether trade between countries should be free or should the state control and regulate foreign trade. *Free trade vs. trade protection* has been a controversial policy issue, perhaps ever since trade between two nations took place. Economists have debated on this issue over centuries, building up arguments in favour of and against both free trade and controlled trade policies. In reality, however, most countries have adopted some or the other kinds of trade restrictions—direct or indirect in spite of efforts made earlier by GATT or now by WTO. We will describe here the meaning of free trade and trade control policies, and present the arguments in favour of and against free trade and protected trade. Here our discussion focuses on conceptual and theoretical aspects of trade policy.

34.2 FREE TRADE POLICY

Free trade policy is based on the principle of non-interference by the government in the foreign trade. The government neither helps nor hinders the sale and purchase of goods and services between the countries. Goods and services can be freely imported from and

exported to the rest of the world. Under free trade policy, the distinction between domestic trade and international trade disappears.

The free trade policy was first advocated by Adam Smith. Later, economists built up rigorous arguments against controlled trade and offered very powerful arguments in favour of free trade. In this section, we briefly describe the case for free multilateral trade, i.e., *arguments* in favour of free trade.

34.2.1 Arguments for Free Trade

Most arguments for free trade have been built on the grounds of *efficiency, economic growth,* and *welfare*. According to Samuelson, there is essentially only one but 'exceedingly powerful' argument for free trade that "*trade promotes a mutually profitable regional division of labour, greatly enhances the potential real national product of all nations, and makes possible higher standard of living all over the globe*".1 This statement sums up all the major arguments in favour of free trade policy. Let us now look at the basis of there arguments.

(*i*) Free Trade Adds to Efficiency and Growth. Haberler, elaborating on the case for free trade, said that a free trade policy is economically advantageous to the participating countries since it maximizes their social product. Free trade is supposed to be carried out under the conditions of free competition in which price mechanism "...automatically ensures that each country, specializes in the production of those goods, and only those goods which it can produce more cheaply, taking account of transport (cost)". Given the real resources of a country, if it specializes in the production of goods in which it is relatively more efficient, or its cost of production is comparatively lower, its total national product will be much larger than if it spreads its limited resources over the production of all the goods, irrespective of the cost of production. With specialization in the efficient sectors, a larger national product can be achieved; a larger exportable surplus can be generated; and a larger volume of goods and services of the country's requirement can be imported from other countries at lower prices. This increases total availability of goods and services and raises the standard of living of the people. How it all happens technically has already been explained while dealing with the theory of comparative advantage.

(*ii*) Free Trade Keeps Prices Low at Competitive Level. Another most powerful argument for free trade is that 'free import lowers the prices of imported goods,' or at least, prevents increase in prices, by preventing the growth of monopolies. Under free trade, each country specializes in a few products, and tries, under the force of competition, to minimize the cost of production, and keeps the price at the competitive level. Besides, free competition in the international market has an 'educative effect' in the sense that it compels countries to enhance their efficiency through better management of resources and for quick adoption of improved and more efficient techniques of production to reduce cost of production. But, if trade is restricted and friction in free competition is created, monopolistic tendencies in the international market might grow. Monopolies generally lead to lower production and higher costs. Consequently, the total world supply of goods and services is reduced and prices are enhanced. It may, however, be mentioned that free trade does not necessarily provide a complete safeguard against international monopolies.

1. Samuelson, P.A., *Economics* (McGraw Hill-Kogakusha Ltd., Tokyo, 1976), 10th Edn., p. 692, (emphasis added).

(*iii*) Free Trade Promoters Economic Welfare. Finally, an argument for free trade with reference to economic welfare, Meade2 argues that free trade helps to optimize world trade and to maximize world production and 'thus maximize world economic efficiency' on the pattern of what he calls 'modified laissez-faire policy'3 for domestic trade.

34.2.2 Arguments Against Free Trade

The case for free trade, built mostly on theoretical grounds, has been challenged on empirical grounds. The assumptions of the free trade policy and the gains from it have been strongly questioned. According to Kindleberger, "Modern theorists … make a very weak case for free trade, arguing only that free trade is better than no trade, and even some trade is better than no trade but being unwilling to say anything as strong as that free trade is better than restricted trade."[4] Besides, the available empirical evidence[5] too does not lend support to the proposition that 'free trade is better than some trade'. Attempts have been made by the economists to estimate the gains from the abolition of tariffs. The gains have been estimated in terms of increase in national income or the cost of maintaining tariff. The estimates show only insignificant gain (or even loss) against highly optimistic gains of free trade. For instance, P.J. Verdoorn has estimated the gains to the six European Common Market countries (viz., France, Italy, Germany, Holland, Belgium, and Luxemburg) from the abolition of 20 per cent trade tariffs to be something in the order of 1/20th of 1 per cent of their national income. Harry G. Johnson estimated the maximum cost of tariffs for Britain (i.e., its cost of not joining the Common Market) to be approximately one per cent of her national income. According to the estimates made by Welmesfelder, Germany gained by less than 1 per cent of her national income from a major reduction in tariff. Thus, the much acclaimed gains from free trade are not tenable against empirical test, despite its powerful logical foundation. That is perhaps, why, Kindleberger, remarks, "It is…reasonable to believe in the doctrine of free trade but unreasonable to be doctrinaire about it."

34.3 TRADE PROTECTION POLICY

34.3.1 Historical Perspective

The much advocated free trade policy had hardly ever existed in its form and content on a universal scale. For some political and economic reasons, most countries had adopted a protectionist policy, at some time or the other in the history of their foreign trade. Perhaps, the world market never provided the perfect market conditions required for free foreign trade on a global scale. One could, however, say that the international economy which 'reasonably approximated' to the world of theoretical model existed between 1860[6]

2. Meade, J.E., *The Theory of International Economic Policy,* Vol. II; *Trade and Welfare* (Oxford University Press, ELBS, Edn., 1970), p. 154.

3. Under laissez faire policy, marginal conditions of free competition are maintained through state interference by offsetting the external economies through fiscal measure.

4. Kindleberger. C.P., *International Economics* (Richard D. Irwin. Inc., Illinois, 1973, and D.B. Taraporevala & Sons, Bombay, 1976), p. 189.

5. Quoted in Lipsey, R.G., *An Introduction to Positive Economics* (Weidenfield and Necolson, 1975, ELBS, 4th Edn.), p. 678.

6. This was the year in which British parliament had abolished duties on all but 60 of 400 taxed commodities.

and the outbreak of the First World War.[7] Even during this period tariffs existed in many countries (e.g., in France, Holland, Belgium, the United States, and Germany) though at a relatively lower rate.

In fact, the growth of free trade itself created conditions for rivalry among countries for market share. This along with aftermath of World War I revived economic nationalism, on the one hand, and protectionism on the other. Consequently, the trend in tariff-protection was revived, and has stayed as a fact of the world economy. For the purpose of regulating foreign trade or protecting a country's interest in foreign trade, tariff is one of the most important tools. By imposing tariffs, i.e., levy on imports and exports, a country can influence the pattern, volume and direction of its trade with the rest of the world. The imposition of tariffs changes the relative price structure which, in its turn, changes the structure and pattern of trade between a country and the rest of the world. For example, if a country wants to curb the imports of luxury goods which are cheaper even on comparable quality basis, it will impose tariffs or custom duties on imported luxury goods at a rate which can more than neutralize the price differences. This would divert consumer demand from foreign to domestic goods. Consequently, import of foreign luxury goods would decrease.

34.3.2 Arguments for Tariffs Protection

1. *Infant-industry argument.* The oldest argument in favour of controlled trade was advanced by Alexander Hamilton in 1790. It was developed further by Friedrick List in the 1840s as an argument for tariff protection to German industries which were then in a developing stage. The argument was later supported by other orthodox economists like Mill, Marshall, Taussing, etc. Incidentally, the economists who put forward the case for tariff protection to the infant industries were mostly 'Free traders,'8 the economists who supported free trade policy.

The core of the argument is that in the initial stage of industrial growth of an economy, most industries are in their infancy, not mature enough to withstand competition from highly developed and well established industries of developed countries. If infant industries are exposed to competition with the industries of highly developed nations which have achieved a high level of technical efficiency, economies of scale and financial strength, they would run the risk of dying out in their infancy. The infant industries of the developing economies, therefore, need tariff protection until they achieve competitive strength. Tariff imposition on competing imported goods makes their prices comparable with or even higher than those of products of infant industries in the domestic market. This change in relative prices, shifts demand from imported goods to domestic goods and helps, thereby, the domestic industries survive and grow. The tariff protection was, however, recommended only for the period of infancy and was to be withdrawn when the infant industry achieved competitive strength.

The infant industry argument has a great deal of plausibility in it. History provides evidence that tariff protection has helped industries in various countries to grow enough to face international competition. But, there is large evidence to show that industries provided with tariff protection could never come up and tariff shelter became a permanent feature. It is, therefore, suggested that tariff protection is warranted only, "if the industry in question

7. Meade, J.E., *op. cit.* , p. 221.

8. Ellsworth, P.T., *op. cit.*, p. 221.

is clearly suited to the country's factor endowment, market prospects, and facilities for obtaining raw materials and has enough potential to grow within a reasonable short period."9

2. *Diversification of industrial gorwth.* One of the earliest and famous arguments for tariff protection is based on the need for diversification of industries and for balanced growth of the economy in the long run. The logic behind the diversification and balanced growth argument is that the theory of comparative advantage recommends countries to specialize in the industries for which they have suitable and adequate resource endowments. Naturally, most countries would specialize in a narrow range of industries and depend on foreign countries for the rest of their requirements. However, such a trade policy can be successful and desirable only under the following conditions.

(*i*) Peace prevails in the world and no major wars take place;

(*ii*) The world economy remains stable, i.e., economic upheavals like the Great Depression of 1930s do not take place;

(*iii*) There is no political or economic rivalry between the nations except a fair economic competition, and

(*iv*) There is no economic or trade groupism.

It is only under these conditions that the specialization theory would probably work in the best interest of the world economy.

But what if these conditions do not exist? Suppose a major world war breaks out and economic relations are disrupted, what could be the fate of countries depending on other countries for their supplies? Such economies will be completely shattered.10 It is, therefore, in the best interest of the nations to secure their economy against world economic disruption by setting out for a fairly balanced growth and for a reasonable degree of self-reliance. This can be achieved by providing tariff protection to the sectors which have potential to achieve a reasonable level of growth under protection. Some Latin American economists suggest the use of tariffs also against the growth of 'monoculture.' Many Canadian economists plead for autarky even at the cost of 5–10 per cent reduction in their standard of living.

Although the diversification argument has a strong popular appeal, it does not make a general case for tariff protection. It applies only to the highly specialized economies dependent on other countries for their essential supplies. It also applies to the cases where the cost of protection is not prohibitive. In other cases, the problem of securing the economy against war and depression should be solved by some other measures.

3. *Promotion of employment.* Tariff protection is also suggested as an effective remedy to the serious unemployment problem like the one the world experienced during the depression of the 1930s. Imposition of tariff on imports directly competing with domestic products helps to expand employment opportunities in the import-competing industries by securing the domestic market. The employment effect of tariffs does not remain confined only to the protected industries. It spreads also to their ancillary industries and to other industries through a general growth effect.

9. Ellsworth, *op. cit.* , p. 221

10. The world has had such an experience during the period between the two World Wars. Countries like Australia, Argentina, Mexico and Chile, having specialized only in foodstuffs and raw materials suffered a great deal for lack of supplies of other goods.

4. *Tariffs for improving terms of trade.* Tariffs have been suggested also as a means to force down the prices in the exporting country. Tariff imposition reduces the demand for foreign goods in the domestic market. If it disturbs the demand and supply conditions in the exporting countries, prices there go down. Price reduction abroad turns the terms of trade in favour of the tariff imposing country, since it can exchange a given quantity of exports with a larger quantity of imports than earlier.

But the effectiveness of tariffs in improving the terms of trade depends on the elasticity of supply in the exporting country. If supply is highly elastic, the relative price may not be affected at all. Besides, such a measure has always the possibility of retaliation by the affected countries. In case of retaliation, the terms of trade revert back to the earlier level. The net result may be a loss in consumer's satisfaction and worsened allocation of resources. Therefore, it is suggested that if at all necessary, an optimal tariff should be used instead of a prohibitive one.

5. *Anti-dumping tariff.* There is a strong case for tariff against dumping. Dumping means setting a price for the foreign markets much lower than those in the domestic market with a view to destroying the potential competition. in modern times, China is said to be dumping its products at a large and wide scale. In case of dumping, the importing countries may take recourse to tariff protection. The validity of this argument is, however, limited to intermittent and short-run dumping. However, short-run anti-dumping tariffs can prove very disturbing since they cause short-term fluctuations, malallocation of resources, and may also sometimes prove disastrous. This argument does not apply to the persistent, continuing dumping based on discriminating monopoly pricing in which case the exporting country is capable of charging different prices because of different elasticities of demand at home and in foreign markets. It is also argued that it is rather in the interest of the importing countries not to impose tariffs, as it would be able to consume larger quantity of goods at lower prices.

6. *Other arguments for tariff protection.* There are other cases cited in economic literature. Some of these arguments are: (*i*) tariffs for keeping money in the country, (*ii*) tariffs for higher money wages; (*iii*) tariffs against cheap labour arguments; (*iv*) tariffs protection of the home market and (*v*) tariffs for equalizing cost of production. These arguments have, however, been put under the category of 'nonsense' or 'fallacious' arguments.

34.3.3 Arguments for Protected Trade Are Questionable

The arguments for tariff protection as a means of industrial diversification and employment generation have, however, been questioned on the basis of their **limitations,** especially its employment effect.

First, employment creation through protection is based on an implicit assumption that the exports of the country will not be affected, or that other countries will continue to import as much as they used to before the tariff imposition. So long as this assumption holds, the cost of employment generation in tariff-imposing country falls on its trading partners. The affected countries, therefore, impose retaliatory tariffs on their imports, which neutralizes the employment effect of tariff protection.

Secondly, even if other countries do not take retaliatory steps, employment effect of tariff protection is still doubtful. Tariffs on imports reduce the exports of the tariff affected

countries to the extent of price elasticity of imports, and thereby generate unemployment there. Fall in employment and hence fall in incomes, reduces their importing capacity. It ultimately reduces the exports, and thereby employment, in the tariff imposing country. So the employment generation through tariff protection may get counter-balanced with unemployment generation in the export sector.

Thirdly, employment creation through tariff protection may involve a permanent reallocation of resources which may result in a temporary gain but permanent malallocation of resources and reduction in employment.

Fourthly, if growth potential of the protected industries is not fairly large and self-sustaining, tariffs may not only become a permanent feature but would also yield only a marginal gain in terms of additional employment.

Finally, if the unemployment problem is confined only to a few countries, tariffs may be an effective solution, but not if the problem is global.

34.4 INDIA'S FOREIGN TRADE POLICY — A HISTORICAL VIEW: 1950–2005

Foreign trade policy of a country refers to import and export policy and control of foreign exchange. India's foreign trade policy consisting of its export and import policies and exchange control has undergone several changes since 1951, the year of the commencement of the First Five Year Plan. The history of India's trade policy can be divided into the following ***three phases.***

(*i*) Phase I: from 1951 to 1985—the phase of restrictive trade policy, which is also known as *protective trade policy* and *inward looking trade policy.*

(*ii*) Phase II: from 1980 to 1991—the phase of gradual liberalization of restrictive trade policy.

(*iii*) Phase III: 1991 onwards—the phase of post-economic reform foreign trade policy, also known as drastic reform of trade policy leading to freer trade.

The basic characteristics of the three phases of India's trade policy are briefly described below.

Phase-I: Restrictive Foreign Trade Policy Import policy in phase-I was formulated and adopted by the government according to the need of the country at that time. All the arguments given in favour of protective trade policy applied to the Indian economy. Essentially, under her industrialization policy, India wanted to protect domestic industries from foreign competition. In phase-I, India had adopted a highly restrictive and protective trade policy with different measures for restricting imports and promoting exports.

Import policy of the country was characterized by (*i*) import restriction and (*ii*) import substitution—helping the growth of industries capable of producing importable commodities. Import was restricted to a few essential commodities and import licence holders only. Under Import policy goods were classified under three categories: (*a*) imports prohibited, (*b*) controlled imports with quantitative restriction and with high import duties varying between 150–300 per cent, and (*c*) imports by the government agencies like State Trading Corporation (STC). Import restriction had two broad objectives: (*i*) saving foreign

exchange for importing essential goods and (*ii*) to achieve self-reliance in the production of goods and services by widening the scope of domestic industries.

As regards ***export policy,*** export of certain essential goods was allowed only under the government permission; export of some goods was freely allowed; and export of certain category of goods was facilitated with subsidy and lower bank rate.

Phase-II: Liberalization of Foreign Trade By the end of Fifth Five Year Plan (1974–79), Indian economy had gained some economic strength; growth rate had risen to 5%; inflation was under control by and large; and India had achieved some level of industrial growth. It was realized by the economists and the policy-makers that Indian industry was being provided over-protection; over-protection was affecting industrial growth adversely, restrictive trade policy was also hampering the overall economic growth of the country; and that economy needed some openness. The export-import policy of 1980s was guided by the recommendations of three committees, viz., Alexander Committee (1978), Tandon Committee (1982), and Hussain Committee (1985). These committees made recommendations of trade liberalization. As a result, Rajiv Gandhi, the then Prime Minister of the country, adopted a policy of drastic liberalization of foreign trade of the country with the objective of enhancing export competitiveness of Indian export. As a policy of trade liberalization, the list of prohibited imports was reduced; import duties on many import items were reduced considerably; import licensing was restricted to fewer goods; foreign exchange control measures were relaxed; and COFEPOSA was also relaxed.

Phase-III: Post-Economic Reform Foreign Trade Policy The post-reform trade policy is marked by substantial liberalization of imports and exports. In 1996, import of 6161 items were made free from import duty. By March 2000, the list was increased to 8066. Export- import (EXIM) policy of 2000–01 removed quantitative restriction on imports from 714 commodities. More important change was made in regard to tariff rationalization. On the recommendation of Chelliah Committee, the Finance Ministry reduced import duty from 110 per cent to 85 per cent in 1993–94 Budget, from 85 per cent to 65 per cent in 1994–95 Budget, from 65 per cent to 50 per cent in 1995–96 Budget, and then from 50 per cent to 35 per cent in 2000–01 Budget.

Besides, in order to facilitate desirable export and import, Indian currency which was non- convertible was made convertible on capital account in 1993. Now the government has made the rupee partially convertible at capital account also.

Furthermore, the Government of India announced on March 31, 2002, a new EXIM policy (2002–07) keeping in view the objectives of the Tenth Five Year Plan. In the new policy, the government removed quantitative restrictions on all imports and on all exports. However, depending on supply-demand conditions, the government kept imposing temporary restrictions from time to time on exports and imports, as it did on export of jute and onion in 2002–03 and on wheat export in 2007. Various measures were being adopted to boost agri-exports. Also Special Economic Zones were being set up to facilitate tariff free exports and imports by increasing tax exemptions. Other tax benefits, including central sales tax exemptions, etc., are being offered to encourage exports.

34.5 CURRENT FOREIGN TRADE POLICY OF INDIA (2009–14)

The trade policy reforms made during the 1990s had aimed at export promotion by liberalization, openness, transparency and removal of quantitative restrictions. The trade-promotion oriented policy was adopted in 2000-01 which continued till 2008–09. However, the government felt forces to make some important changes in its foreign trade policy during the period 2009–14 because of changing world economic conditions. Let us now look at the background in which the trade policy for 2009–14 was formulated.

The world economy was facing the second worst economic depression since 2007. The economic depression experienced in 2009 faced by the world economy was the worst of period. The severe economic depression was reflected by a sharp decline in production, trade, capital flows, employment, per capita income, per capita saving and investment. The WTO had projected a 9% decline and IMF had projected 11% decline in global trade. Indian economy was affected but not as severely as other economies. Nevertheless, economic growth of the Indian economy, especially its exports, was seriously constrained. It was under this condition that Indian government had to formulate its foreign trade policy for 2009–14. It was a daunting task.

The ***main objectives of the new foreign trade policy*** were following.

(*i*) To reverse the declining trend of exports by providing additional supports to the export industries hit by global recession,

(*ii*) To provide facilities to increase exports at an annual rate of 15 %, and further the export growth rate to 25% by 2014, and

(*iii*) To achieve long-term objective of doubling India's share in world exports by 2020 .

Broadly speaking, the government planned to adopt a policy-mix to achieve these objectives. The policy-mix included (*a*) fiscal incentives to exporters, (*b*) institutional changes wherever required, (*c*) rationalization of export procedure, (*d*) widening access to export market, and (*e*) diversification of export products. The major policy measures announced by the government in new foreign trade policy to promote exports include the following measures.

1. Expansion of market area. India's 70% foreign trade was concentrated in in few countries Euro Zone (36%), the USA (18%) and Japan (16%). But since market was shrinking due to great depression, India planned to expand its foreign trade market to Latin America, South Africa, Brazil and Mexico.

2. Increase in incentives for FMS and FPS. Incentive available under the Focus Market Scheme (FMS) has been raised from 2.5% to 3% and under Focus Product Scheme (FPS) has been raised from 1.25% to 2%.

3. Reduction of duty on import of capital goods. New foreign trade policy allows zero duty import of capital goods including engineering and electronic products, basic chemicals and pharmaceuticals, apparels and textiles, plastics, handicrafts, and leather.

SUMMARY

- There have been two kinds of trade policies adopted by different countries over a period of time: (i) free trade policy, and (ii) protected trade or controlled trade policy. Both these policies have their own merits and demerits and advantages and disadvantages.
- Free trade policy is the policy of non-interference by the government in foreign trade. Under free trade policy goods and services are freely exported and imported by the people as per need of the country.
- Advantages of free trade policy as pointed out by the economists are: (a) it adds efficiency to production and economic growth, (ii) it keeps prices low at competitive level, and (iii) it promotes economic welfare of trading countries. However, modern economists claim that free trade policy has not been found empirically to be as advantageous as claimed by a section of economists.
- Protected trade policy is essentially controlled trade policy. Under controlled trade policy government controls and regulated exports and imports through different kinds of policy measures including (i) imposition of tariffs – the custom duties, and (ii) direct controls of exports and imports.
- The arguments for adopting protected trade policy are (i) development of infant industries, (ii) diversification of industrial growth, (iii) promotion of labour employment, (iv) tariff imposition to improve terms of trade, and (v) controlling dumping by anti-dumping tariffs.
- As noted above, both free trade and trade control policies have their own merits and demerits. Whether free trade policy or trade control policy is more is more beneficial for trading countries has been a controversial issue.
- India being a developing economy has been following trade control policy. Different methods of import controls and export promotion measures have adapted right from 1950 to 2014.

REVIEW QUESTIONS

1. "Free trade promotes a mutually profitable regional division of labour, greatly enhances the potential of real national product of all nations and makes possible higher standard of living all over the globe." Explain and critically examine the statement.
2. Critically examine the arguments for trade protection. Why is trade protection desirable for less developed countries?
3. What is the 'infant-industry' argument? Why is this argument discredited with inefficiency?
4. Why do most countries adopt regulatory trade policy while free trade policy looks theoretically more sound?
5. What are the major arguments for trade protection? Do these arguments hold under all conditions?
6. What trade policy has India adopted during the past six decades? How has India's foreign trade policy changed over time?

FURTHER READING

ChaCholiades, M., *International Economics* (McGraw-Hill, New York, 1990), Chs. 7–9.

dwivedi, d.N., *International Economics: Theory and Policy* (Vikas Publishing House, Noida, 2013), ch. 10.

ellsworth, P.t. and J. Clark leith, *International Economy* (Macmillan Company, New York, 1975), Ch. 14.

helleiNer, G.k., *International Trade and Economic Development* (Penguin Books Ltd., England, 1972), Ch. 7.

Meade, J.e., The Theory of International Economic Policy, Vol. II, *Trade and Welfare* (The English Language Book Society, London, 1970), Part II.

salvatore, sChiavo–CaMPo, *International Economics* (Winthrop Publishing Inc., Cambridge, 1978), Chs. 4–7.

sodersteN, Bo, *International Economics* (Macmillan, New York, 1970), Ch. 19.

CHAPTER

35 Determination of

Foreign Exchange Rate

CHAPTER OBJECTIVES

The objective of this chapter is to discuss how foreign exchange rate – the rate of exchange between two foreign currencies – is determined under different market conditions. The content of this chapter includes the following aspects of exchange rate determination.

- The need for exchange rate determination
- Foreign exchange market and its functions
- Kinds of foreign exchange transactions
- Theory of exchange rate determination in free market
- Purchasing power parity theory of exchange rate determination and
- Controversy on fixed vs. flexible exchange rates

35.1 INTRODUCTION

In foreign trade, goods and services are traded across national boundaries but the currency of one country is not acceptable as legal tender in other countries. This creates problem of payments. For instance, Indian rupee is not acceptable as a medium of exchange in other countries (except in Nepal and Bhutan), nor is the currency of any other country acceptable as a general medium of exchange in India. The payment for imports have, therefore, to be made in terms the currency of the exporting country. For foreign payments, therefore, importing countries have to buy the currency of the exporting country in the foreign exchange market. For instance, if India imports fighter planes from the USA, India will have to buy US dollars in the exchange market (assuming India does not have dollar reserves) to make the payment to the USA. And if the US importers import garments from India, they will have to buy Indian rupee in the foreign exchange market (assuming they do not possess rupee reserves) to make payment to the Indian exporters. Foreign currencies are bought and sold in the foreign exchange market. The price of one currency in terms of another is called the *exchange rate*.

Now the question arises: how is the price of one currency determined in terms of another in the foreign exchange market? In other words, how is the rate of exchange between any two currencies determined in the exchange market? This question is answered by the theroy of exchagne rate determination which we will discuss in this chapter.

It may be added at the outset that there are two kinds of market conditions under which the rate of exchange is determined: (*i*) free foreign exchange market, and (*ii*) free foreign exchange market controlled and regulated by the central bank. We will discuss the exchange rate determination under these two different foreign exchange market conditions.1 Let us first get an idea about the organs of the foreign exchange market and their role in exchange rate determination.

35.2 FOREIGN EXCHANGE MARKET AND ITS FUNCTIONS

35.2.1 Foreign Exchange Market

The foreign exchange market refers to the buyers and sellers involved in sale and purchase of foreign currencies. Foreign exchange market is, thus, the market in which currencies of different countries are bought and sold. This market is constituted of the central banks, brokers, commercial banks, exporters and importers, investors, tourists and immigrants. These main players of the market make the structure of the foreign exchange market. Their place and position is shown in Fig. 35.1. As the figure shows, at the bottom of the foreign exchange market are the actual buyers and sellers of the foreign currencies— *exporters, importers, tourists, investors* and *immigrants*. They are the actual users of the foreign exchange. Those who need foreign currency approach commercial

Fig. 35.1 *Structure of the Foreign*

banks to buy the currency.

Exchange Market

Commercial banks make the second and the most important organ of the foreign exchange market. Banks dealing in foreign exchange play the role of 'market makers' in the sense that they quote the daily exchange rates for buying and selling the foreign currencies. They also work as the clearing houses. They help in wiping out the difference between the demand for and supply of a currency. They buy foreign currency from the brokers and sell to the buyers.

Foreign exchanges brokers make the third layer of the market. Brokers work as a link between the central bank and the commercial banks and also between the banks. They are the major source of market information and they work also as a link between the actual buyers and the banks. They themselves do not buy or sell the foreign currency—they only strike the deal between the buyers and sellers on a commission basis.

The *central banks* of different countries are the apex body in the organization of the foreign exchange market. They work as the 'custodians of the foreign exchange of the country' and 'lenders of the last resort'. They have the power to control and regulate the foreign exchange market with a view to make sure that the foreign exchange market works in an orderly fashion. One of their main functions is to prevent violent fluctuations in the

1. We will ignore the outdated mint parity theory of exchange rate.

exchange rate, if necessary, by direct intervention. The main form of intervention is to sell a foreign currency when it gets overvalued and to buy it when it tends to be undervalued.

Foreign exchange market is the biggest market in the world today. According to a recent estimate, daily transaction exceeding US $100 billion. Most major countries have foreign exchange market centres. London, New York, Paris, Tokyo, Zurich, Frankfurt, Singapore are some famous foreign exchange market centres for the US dollar. The foreign exchange market centres are very well connected with each other by telecommunication systems and they work 24 hours.

35.2.2 Functions of the Foreign Exchange Market

The foreign exchange market performs the following functions which are, in fact, the outcome of its working.

(*i*) The foreign exchange market transfers funds (foreign currency) from one country to another where they are needed in the settlement of payments;

(*ii*) It provides short-term credit to the importers and, thereby, facilitates the smooth flow of goods and services from one country to another, and

(*iii*) The spot and forward markets work in such a way that it helps often in stabilizing the foreign exchange rate.

35.2.3 Spot and Forward Markets

The foreign exchange market is classified on the basis of the nature of transactions in two categories: (*i*) *spot market* and (*ii*) *forward market*. Accordingly, there are two kinds of exchange rates in the foreign exchange market: *spot rate* and *forward rate*. The nature of spot and forward markets are described here briefly.

(*i*) *Spot Market.* The spot market refers to that segment of the foreign exchange market in which sale and purchase transactions are settled within *two days* of the deal. That is, when buyers and sellers of a currency settle their transaction within two days of the deal, it is called *spot transaction*. The spot sale and purchase of foreign exchange makes the *spot market*. The rate at which foreign currency is bought and sold in the spot market is called the *spot exchange rate*. For all practical purposes, the *spot rate* is the prevailing exchange rate.

(*ii*) *Forward Market.* The forward exchange market refers to foreign exchange deals for sale and purchase of foreign currency at some future date, normally after 90 days of the deal. When buyers and sellers enter an agreement to buy and sell a foreign currency after 90 days of the deal at the agreed rate of exchange, it is called *forward transaction*. The forward transactions in foreign exchange make the *forward market*. The exchange rate settled between the buyers and sellers for forward sale and purchase of currencies is called *forward exchange rate*.

35.2.4 Nature of Foreign Exchange Transactions

There are three kinds of players and three kinds of transactions in the foreign exchange market. These are: **(1) hedging, (2) arbitrage and (3) speculation.** We will now briefly discuss the nature of deals in these kinds of foreign exchange transactions.

1. *Hedging.* An important feature of the forward exchange market is hedging (or risk covering). Hedging is settling the exchange rate in advance for future transaction with a view to avoiding the loss that might arise due to exchange depreciation in future. Hedging is essentially covering the risk arising out of exchange rate fluctuation. The contract between exporters and importers to sell and buy goods at some future date takes place at the current prices and the current exchange rate. There is always a time-lag between the sale-purchase agreement and the final delivery of goods. Under a flexible exchange rate system, it is quite possible that exchange rate fluctuates up and down which may reduce the profit or may altogether wipe it out. The country whose currency depreciates due to exchange rate fluctuation suffers losses. Such losses can be avoided provided there exists a market for forward sale and purchase of foreign exchange, which can assure the supply of foreign exchange at a future date at the exchange rate prevailing at the time of bargain. The existence of a forward exchange market enables the exporter to hedge against the probable risk due to currency depreciation. Through hedging, the exporter is assured of the value of his exports at the current exchange rate. Similarly, an importer who enters into an agreement to import goods from abroad secures his interest against possible increases in the cost of imports due to exchange rate fluctuation, by buying the foreign exchange for import payments in advance.

The banks dealing in forward purchase and sale of foreign exchange and exporters and importers make the forward market. Banks play a very important role in the forward foreign exchange market as they help exporters and importers in securing their interest against the risk involved in exchange rate fluctuations and, thereby, enable them to concentrate on their pure trading operations.

2. *Arbitrage.* *Arbitrage is an act of simultaneous purchase and sale of different currencies in two or more exchange markets.* The objective of such operations is to make profit, taking the advantage of exchange-rate differentials in the various markets.

The significance of arbitraging lies in its ability to equate the foreign exchange rates in all the major foreign exchange markets. Arbitrage operations lead to transfer of foreign exchange from the markets where exchange rate is low to the markets where exchange rate is higher. Thus, arbitraging equates the demand for foreign exchange with its supply. It works as a stabilizing factor in the foreign exchange markets.

Arbitrage is, however, possible only where foreign exchange is free from controls, and if any, controls are of limited significance. When purchase and sale of foreign exchange are subject to severe controls and regulations, arbitrage becomes impossible.

3. *Speculations.* Broadly speaking, speculation in the context of foreign exchange, is an act of buying and selling a currency under the condition of uncertainty with a view to making profit. Speculators buy a currency when it is weak and sell it when it is strong. Also, speculators who expect the spot rate to increase in future, buy forward and sell ‘on the spot’ the currency that they have bought. On the contrary, the speculators who anticipate a fall in the exchange rate ‘sell forward’ at the current rate and buy spot when they need the currency for delivery.

The speculative transactions are said to have both stabilizing and destabilizing impact on the exchange rate. If speculators buy a currency when it is cheap and sell when it is dear, it has a stabilizing effect on the exchange rate. There is, however, controversy about the stabilizing and destabilizing effects of speculative transactions.[2]

2. See also Kindleberger, *op. cit.* , pp. 409–12.

One of the many controversial conditions for destabilizing speculation is “selling when a currency is weak in the expectation that it will be weaker, or buying it when it rises in price, believing that it will rise still more.” Milton Friedman has, however, pointed out that “if the exchange rate were highly overvalued or undervalued, speculation which drove it towards equilibrium might properly be regarded as stabilizing if it reinforced market movements.”[3] According to Robert Aliber, “if speculation drives the spot and forward markets in the same direction rather than in the opposite direction, it is destabilizing”[4]. A more general view is that if speculation pushes the exchange rate beyond or below critical level from where return is impossible or disadvantageous, it is destabilizing. On the other hand, advocates of flexible exchange rate believe that speculation cannot be destabilizing. It may be concluded from above that if speculators buy when a currency is weak and sell when it is strong, it will be stabilizing.

35.3 DETERMINATION OF EXCHANGE RATE IN A FREE MARKET

35.3.1 Free Exchange Market

A free exchange market is one in which there is no restriction on the foreign exchange transactions. The government does not intervene with the process of exchange rate determination: rather, it creates conditions for the exchange rate to be determined by the market forces by allowing free sale and purchase of foreign exchange. In a free market setting, *foreign exchange market does not refer to a particular place*. *Infact, it refers to the facilities provided by the bankers, brokers, and other specialized institutions which deal in foreign exchange*. The foreign exchange market for a currency may exist anywhere in the world. Let us now see how exchange rate is determined in a free market.

35.3.2 Exchange Rate Determination in Free Market: The Market Theory

As mentioned above, *the exchange rate is the price of one currency in terms of another. Or, the rate of exchange is the rate at which one currency is exchanged for another.* Since rate of exchange is a price, it is determined by the market forces, viz., demand and supply. The demand for and supply of foreign exchange are composite demand and supply. The demand for foreign exchange is derived from the demand for foreign goods, services and securities. In addition, the other kind of demand comes from the speculators and the monetary authorities willing to build their foreign exchange reserves. Similarly, the supply of foreign exchange is derived from a composite supply of a variety of goods, services, and securities. In addition, foreign exchange supply also comes from speculators, foreign exchange dealers and the monetary authorities trying to get rid of their excess foreign exchange reserves. The exchange rate determination by market forces of demand and supply is called as the ***market theory of exchange rate determination.***

To explain the market theory of exchange rate determination, let us first explain the derivation of demand and supply curves for foreign exchange. The demand curve for foreign exchange is derived from the demand for foreign goods and services. There is an inverse relationship between exchange rate and demand for foreign exchange as shown by *DD* 1 in

3. Milton Friedman, “The case for Flexible Exchange Rate” in his *Essays in Positive Economics* (Chicago: University Press of Chicago, 1953).

4. R.Z. Aliker, “Speculation in Foreign Exchange: The European Experiment 1919–1926) *Yale Economic Essays*, Vol. 2, 1962, p. 234.

Fig. 35.2. In fact, a lower exchange rate implies lower prices of foreign goods, and *vice versa*. As the law of demand states, the lower the exchange rate, the higher the demand for foreign goods (given the domestic comparable prices) and the greater the demand for foreign exchange. As regards the supply of foreign exchange, it is derived from the demand for *foreign exchange*. Any country which buys a foreign currency by paying in its own currency, simultaneously supplies the foreign exchange (its own currency) for other nations. ‘A country’s supply schedule of its own currency in the foreign currency market is, therefore, best viewed as an inverted demand schedule for foreign exchange’.5 There is, thus, a positive relationship between exchange rate and supply of foreign exchange. The supply curve of foreign exchange is then the usual supply curve sloping upward to the right as shown by *SS*¢ in Fig. 35.2.

To illustrate the determination of exchange rate in a free market, let us suppose a two-country (India and the USA) and two-currency (rupee and dollar) model. Let us also suppose that hypothetical demand and supply curves for dollar are given by *DD* 1 curve and *SS*¢ curves, respectively, as shown in Fig. 35.2. The demand and supply curves intersect at point *P*, where demand for dollar (by India’s importers) and supply

Fig. 35.2 *Determination of Exchange Rate*

of dollars are equal. At this point,
exchange rate is determined at $ 1 = `45, the demand for and supply of dollars being equal at $ 10 million. This exchange rate clears the foreign exchange market. It is, therefore, an equilibrium rate of exchange. This rate would keep in balance the autonomous transactions between India and the USA, and would keep balance of payments between the two countries in equilibrium. In case India's demand for dollars increases, given the exchange rate, her demand curve *DD* 1 will shift upwards to *DD* 2, other things remaining the same. The exchange rate will then rise to $ 1 = `48. This implies depreciation of rupee and appreciation of dollar.
Let us now describe the factors which cause shifts in the demand curve for foreign exchange and change in exchange rate.

35.3.3 Causes of Change in Exchange Rate

The exchange rate determined by the market forces changes due to changes in the factors which determine the demand for foreign exchange. The major factors which cause change in demand for foreign exchange and force a change in the exchange rate are described here briefly.

First, a rise in the domestic prices and cost of production (foreign prices and costs remaining the same) increases the demand for foreign goods, and also the demand for

5. Sodersten, Bo, *op. cit.*, (1980), p. 320.

foreign exchange. Consequently, the demand curve shifts upwards causing a rise in the exchange rate. On the other hand, a fall in the exports causes a leftward shift in the supply of national currency. The ultimate result of the upward shifts in demand curve and backward shift in the supply curve is a rise in the exchange rate. Frequant rise in demand and/or fall in supply of foreign exchange cause fluctuation is the exchange rate.

Secondly, a rise in the real income of a country, other factors remaining the same, causes an increase in its imports depending on income-elasticities of imports. Increase in imports causes an increase in demand for foreign currency causing a rise in the exchange rate. Reverse will be the impact if real income of a country decreases.

Thirdly, domestic real income remaining the same, a rise in the real income abroad would tend to increase the domestic exports and a simultaneous increase in the foreign demand for domestic currency leading to a rise in the price of the domestic currency. This causes a change in the exchange rate.

Fourthly, if the rate of interest in the domestic economy increases in relation to interest rates in the foreign countries, the capital inflow increases and outflow decreases. Consequently, foreign demand for domestic currency increases in relation to domestic demand for foreign currency. This increases the rate of exchange. A fall in the domestic rate of interest will have a reverse impact on the exchange rate.

Finally, the speculative purchases and sales of foreign exchange in the exchange market also affect the demand for and supply of foreign exchange. This causes a change in the exchange rate. The change in the exchange rate depends on the effect of speculators' behaviour on the supply and demand conditions in the foreign exchange market.

35.3.4 Critical Evaluation

The *market theory of exchange rate* has been criticized on the grounds of its limited applicability. Applicability of market theory of exchange rate determination has been questioned on the following grounds.

First, the market theory of exchange rate assumes that domestic price levels are independent of exchange rate whereas there is a close link between them. A continuous rise in the general price level leads to exchange depreciation.

Secondly, this theory is applicable only where the exchange rate responds to the changes in conditions of demand for and supply of foreign exchange.

Thirdly, market theory works better only if exports of a country do not have a perfect substitute available in the domestic market of the importing country, otherwise rise in exchange rate shifts demand from foreign goods to domestic goods.

Finally, the equilibrium exchange rate determined by market forces is supposed to be stable and to bring automatic adjustments in the balance of payments. But experience shows that the market mechanism has not been as automatic as it is expected to be.
Inspite of its limitations, the market theory of exchange is considered to be better than other theories of exchange rate determination: it explains better the state of affairs

in the foreign exchange market and fluctuations in the exchange rate.

35.4 PURCHASING POWER PARITY (PPP) THEORY OF EXCHANGE RATE

A Swedish economist, Gustav Cassel, developed the concept of equilibrium rate of exchange6, after the First World War. This theory asserts that the relative value of different currencies correspond to the relation between the real purchasing power of each currency in its own country.7 In other words, under inconvertible paper currency system, the rate of exchange between any two currencies is determined on the basis of their purchasing power in their respective countries. For example, if a basket of goods and services can be bought for `100 in India, and for $ 2 in the USA and if both countries are on inconvertible paper currency, the exchange rate between the two currencies will be determined as follows.

$ 2 = `100, or $ 1 = `50

This theory, however cannot be applied to absolute levels of prices which depend on a number of factors operating under divergent conditions. It is rather impossible to measure absolute prices. The theory can nevertheless be used to explain the changes in the international levels of prices and for determining the changes in the exchange rate. For example, let P_{a0} and P_{b0} represent the price levels in countries *A* and *B*, respectively, in base (*o*) year, and R_0 be the exchange rate between their respective currencies. The rate of exchange (R_1) between the two currencies in some future year may be calculatd as folows:

$$R_1 = R_0\left(\frac{P_{a1}/P_{a0}}{P_{b1}/P_{b0}}\right) \text{ and } R_1 = R_0\left(\frac{P_{a1}/P_{b1}}{P_{a0}/P_{b0}}\right)$$

where P_{a1} amd P_{b1} denote price levels (price index number) for countries *A* and *B*, respectively, in the *reference year*.

Criticism of PPP Theory

The purchasing power parity theory too has been *criticized* on the following grounds.

First, the *PPP* theory helps only in determining the change in exchange rate, not the absolute exchange rate. Hence, purchasing power parity theory cannot be applied to the absolute levels of prices.

Secondly, the wholesale price index number, on which purchasing power parity theory is based, does not give an accurate measure of change in purchasing power.

Thirdly, while purchasing power of a currency in terms of only internationally traded goods and services is relevant for its external value, the wholesale price index number includes prices of all the commodities.

Fourthly, apart from goods, many other items of services (e.g., banking, insurance, consultancy, etc.), enter the foreign transactions and a large amount of capital transfers take place between the nations which do not enter the wholesale price index number.

Fifthly, as Haberler has pointed out, the imposition of embargo and tariffs, and provision of export subsidy often cause deviations in the purchasing power which are not accounted for by the *PPP* theory.

6. Haberler G. has however quoted Prof. Angell to show that the purchasing power parity theory was earlier formulated by Wheatley in 1802 and by William Black in 1810, *op. cit.*, p. 32.

7. Haberler, *op. cit.*, p. 32.

Sixthly, the change in exchange rate depends, by and large, on the elasticities of reciprocal demand whereas purchasing power parity theory recognizes the changes in the exchange rates due to changes only in the relative prices.

Finally, the *PPP* theory assumes that the relative commodity prices are the sole determinants of international transactions and the change in the relative prices is the sole determinant of exchange rate. In reality, however, changes in exchange rate also take place due to disequilibrium caused by capital transfers, service payments and change in real income.

Despite these shortcomings, however, this theory may be used as the first approximation to an equilibrium rate of exchange during the periods of high and frequent price changes. For a more accurate measure in the exchange rate, other factors like capital flows, changes in tastes and real incomes would have to be taken into account.

35.5 FIXED EXCHANGE RATE SYSTEM

In general, most countries do not leave the foreign exchange rate to be determined by the market forces. Instead, they adopt a fixed exchange rate policy. In this section, we discuss the determination of the fixed exchange rate and its advantages.

35.5.1 Method of Fixing the Exchange Rate

Under exchange control system, the rate of exchange is fixed by the central bank of the country. Flexibility of exchange rate is allowed within certain limits, usually 10 per cent under the IMF arrangements. When the exchange rate is fixed, the central bank issues an official price of its currency in terms of a reserve currency (usually US dollar or

D

S

D

D

per $)`

P

T

F

B

P

Exchange rate (

D

D

S

D

O

M

N

Q

R

Amount of Dollars

Fig. 35.3 *Exchange Rate under Exchange Control*

alternatively a basket of key currencies). The exchange rate thus fixed is known as the currency's *par value*. The exchange rate fixed in this manner is known as *fixed* or *official exchange rate*. After the rate is fixed, the central bank undertakes to buy and sell the foreign exchange in the exchange market, and private purchases and sales are suspended. Any change in the exchange rate, if necessary, is made by the central bank. A reduction in currency's par value is known as *devaluation*, and a rise, as *revaluation.*

Control and Regulation of Fixed Rate

The fixed exchange rate is today a rule rather than an exception. Most countries have adopted fixed exchange rate systems. This rate is managed by the central bank of the country in the following manner.

Let us suppose (*i*) that free market demand and supply curves for dollar in India are given by *DD* and *SS¢* curves, respectively, as in Fig. 35.3; (*ii*) that RBI fixes the exchange rate at *OF* and allows exchange rate variation between *OB* and *OT*, and (*iii*) that demand for foreign exchange is controlled through rationing and quota, and demand for dollar is allowed to fluctuate between *ON* and *OQ*, i.e., dollar demand curve may shift up and down between points *P* and *P¢* on the *SS¢* curve. So long as demand variations remain confined within these limits, the RBI need not take measures to control demand. But, if for some reasons, the demand curve for dollars shifts up or down beyond these limits, the RBI will intervene to curb the demand fluctuation—it will sell or buy dollars to prevent any major fluctuation in its prices. For example, if the demand curve shifts upward to *D¢ D¢*, the central bank must sell dollars to the extent of *QR* to prevent the rise in dollar price beyond *OT*. This is how central banks try to stabilize the fixed exchange rate. Similarly, if demand for US dollar decreases and demand curve shifts downward to D^2 D^2, the RBI would buy *MN* dollars to prevent the fall in exchange rate below *OB*.

35.5.2 Exchange Rate under the IMF System

The exchange rate system under Bretten Woods Agreement is the system of fixed exchange rate. Although the exchange rates are, in principle, fixed (or pegged), the member nations can vary their exchange rates to the extent of 10 per cent in consultation with the *IMF*. But in case of fundamental disequilibrium member countries are allowed to devalue their currency by more than 10 per cent. Since 1950, many countries devalued their currency heavily. For instance, Britain, France and many other West European countries devalued their currencies by about 30 per cent in the early 1950s. In 1958, France devalued its currency again by 18 per cent. Britain devalued the pound sterling by 15 per cent in 1967. India devalued the rupee by about 38 per cent in 1969. Between 1990-91 and March 1992–93, India had depreciated her currency against the US dollar by about 75 per cent. Most governments had maintained adjustable fixed exchange rate till 1973 under the *IMF* system. But the *IMF* system failed to provide an adequate solution to three major problems causing exchange instability. There problems are:

(*i*) providing sufficient reserves to mitigate the short-term fluctuations in the balance of payments while maintaining the fixed exchange rate system;

(*ii*) problems of long-term adjustments in the balance of payments, and

(*iii*) crisis generated by speculative transactions.

For these reasons, the currencies of many countries, specially the reserve currencies, were subject to frequent devaluation in the early 1970s. This raised doubts about the plausibility of the Bretten Woods System, and also about the viability of the fixed exchange rate system. Consequently the Britten Woods system broke down in 1970s.

35.6 FIXED *vs*. FLEXIBLE EXCHANGE RATE

The break down of the Bretten Woods system generated a debate by the economists on whether fixed or flexible exchange rate system is desirable for the world economy. The arguments offered by the economists in favour of fixed and flexible exchange rate systems are mentioned here briefly.

35.6.1 Arguments for Fixed Rate

First, the argument in favour of fixed exchange rate is that it provides stability in the foreign exchange markets and certainty about the future course of exchange rate and it eliminates risk caused by uncertainty due to fluctuations in the exchange rate. The stability of exchange rate encourages international trade. On the contrary, flexible exchange rate system causes uncertainty and might also often lead to violent fluctuations in international trade.

Secondly, the fixed exchange rate system creates conditions for *smooth flow of international capital* simply because it ensures a certain return on the foreign investment, while in the case of flexible exchange rate, capital flows are constrained because of uncertainty about expected rate of return.

Thirdly, the fixed rate *eliminates the possibility of speculation*, whereby it removes the dangers of speculative activities in the foreign exchange market. On the contrary, flexible exchange rates encourage speculation. As mentioned earlier in this chapter, there is controversy about the destabilizing effects of speculation. But if speculators buy a currency when it is strong and sell it when it is weak, speculation will be destabilizing.

Fourthly, the fixed exchange rate system reduces the possibility of competitive depreciation of currencies, as it happened during the 1930s. The possibility has been further strengthened by the *IMF* rule for the member nations. Also, deviations from the fixed rates are easily adjustable.

Finally, a case is also made in favour of fixed exchange rate on the basis of existence of currency area. The flexible exchange rate is said to be unsuitable between the nations constituting a currency area, since it leads to a chaotic situation and hence hampers trade between them.

35.6.2 Arguments for Flexible Rate

The advocates of flexible or floating exchange rate system have put forward no less convincing arguments in its favour. As mentioned above, the major arguments placed against flexible exchange rate are that it is destabilizing; it causes uncertainty and risk; and, it encourages speculation. All these charges have been denied by the proponents of flexible exchange rates. As Sodersten puts it, "It is difficult to argue on a *priori* grounds that flexible exchange rates would either decrease or increase uncertainty in an economy. This depends to a large degree on which factors produce the uncertainty."8 Besides, many other arguments have been put forward in favour of flexible exchange rate.

8. B. Soderston, *International Economic*, *op. cit.*, p. 273.

First, a flexible exchange rate system provides larger degree of autonomy in respect of domestic economic policies. For, under flexible exchange rate system, it is not obligatory for the countries to tune their domestic economic policies to the fixed foreign exchange rate.

Secondly, a great advantage of flexible exchange rate is that it is self-adjusting and, therefore, it does not devolve on the government to maintain an adequate foreign exchange reserve.

Thirdly, the flexible exchange rate, which is determined by market forces, has a theory behind it and has the quality of predictability.

Fourthly, flexible exchange rates serve as a barometer of the actual purchasing power and strength of a currency in the foreign exchange market. It serves as a useful parameter in the formulation of the domestic economic policies.

Finally, economists have also argued that the most serious charge against fluctuating exchange rate, i.e., uncertainty, is not tenable because speculators themselves create conditions for exchange rate stability. Also, the degree of uncertainty associated with flexible exchange rates would not be much greater than what the world has experienced with adjustable fixed exchange rate under the Bretten Woods System.

35.6.3 Debate Inconclusive

The *debate* on fixed *vs*. flexible exchange rate is inconclusive. In fact, both systems have their own merits and demerits. The empirical evidence does not provide sufficient proof to the proclaimed advantages of flexible exchange rate. The period between the two World Wars was the period of experiments with the flexible exchange rate and the world monetary system was under severe strain during this period. On the other hand, the world experienced exchange instability, particularly during the early 1970s, under the fixed exchange rate system. It was equally disconcerting.

The destabilizing effects of speculation have been experienced under both the systems. There is ample evidence to show that the fixed rate system is subject to "periodic bouts of very heavy speculation". On the contrary, the experiments with flexible rates by Canada in the 1950s and in 1969, and by Britain and Germany in 1971 have not produced evidence of any serious fluctuation in the exchange rate caused by speculatory transactions.

For these reasons, it has not been possible for the economists to reach an agreement on the relative merits and demerits of both the systems. The counter-arguments keep coming from both the sides. In fact, the majority of central bankers and policy-makers favour the fixed exchange rate, and theoreticians mostly favour the flexible exchange rate, possibly because of their faith in the efficacy of the free market system and price-mechanism.

SUMMARY

• Foreign exchange rate is the rate at which currency of one country can be exchanged for currency of another country. For example, the market exchange rate between Indian currency (INR) and US dollar on 12th April 2014 was determined as US$ 1 = `60.18.

• Exchange rate is determined in the foreign exchange market. The dealers of the foreign exchange market are (*i*) the central bank of the country – the apex body controlling and regulating exchange rate, (*ii*) commercial banks, (*iii*) foreign exchange brokers – coordinators between the central bank and commercial banks, and (*iv*) importers, exporters, investors, tourists and immigrants.

• Based on kinds of the nature of transaction, the foreign exchange market is classified as (*i*) spot market where foreign exchange transaction is settled within two days, and (*ii*) forward market where foreign exchange transaction is settled in 90 days. Foreign exchange transactions are classified also as (*a*) hedging – a risk covering transaction, (*b*) arbitrage – simultaneous sale and purchase of two currencies in two different markets, and (*c*) speculations.

• There are two theories of foreign exchange rate determination: (*i*) free market theory, and (*ii*) purchasing power parity theory. Under free market system, exchange rate is determined by the market forces – the demand for and supply of foreign exchange. According to purchasing power theory, exchange rate between any two currencies is determined on the basis of their comparable purchasing power in terms of the same basket of products.

• In practice, however, there are two systems of exchange rate determination, viz., (*i*) exchange rate determined by the market, and (*ii*) exchange rate determined by the monetary authority of the country – the central bank. In free market system, exchange rate is determined by the market forces, i.e., sale and purchase of foreign exchange. This is flexible exchange rate system. Also, while free market system works through the foreign exchange dealers, the exchange rate for official transactions is determined by the monetary authority of the country. This is known as the fixed exchange rate system.

• Whether fixed or flexible exchange rate should by adopted by the countries of the world has been a controversial issue. While some economists offer arguments in favour of fixed exchange rate, others offer the arguments in favour of flexible exchange rate. In fact, both fixed and flexible exchange system have been found to have their advantages and disadvantages.

• The issue whether fixed or flexible exchange rate is desirable or preferable remains an unsettled issue. In practice, however, both the systems – the fixed and flexible exchange rate systems – work simultaneously.

REVIEW QUESTIONS

1. How is exchange rate determined in a free market? What are the factors responsible for exchange rate fluctuation?
2. Show that the demand for foreign exchange is a derived demand.
3. Explain the purchasing power parity theory of exchange rate determination. What is the utility of this theory in the modern context?
4. What is the role of the various organs of the foreign exchange market? Do speculators' activities necessarily cause exchange rate instability?

5. What are the various kinds of transactions in foreign exchange? How do various foreign exchange transactions affect foreign exchange?
6. How is the exchange rate fixed under the fixed exchange rate system? How is the fixed exchange rate prevented from fluctuating?
7. Examine the arguments for and against the fixed and flexible exchange rates. Which of the two is suitable for less developed countries?

FURTHER READING

ChaCholiades, M., *International Economics* (New York, McGraw-Hill, 1990), Ch.18.

CleMent, M.o., R.l. PfisteR, K.J. Rothwell, *Theoretical Issues in International Economics* (London: Constable and Company Ltd., 1967), Ch. 6.

dwivedi, d.n., *International Economics: Theory and Policy* (New Delhi: Vikas Publishing House, 2013), Chs. 19 and 20.

ellswoRth, P.t., *International Economy* (New York: Macmillan Company, 1964), Chs. 17 and 18.

KReinin, MoRdeChai, e., *International Economics: A Policy Approach*, (New York: Harcourt Brace Jovanovich, Inc., 1975), Ch. 5.
leighton, ReChaRd J., *Economics of International Trade* (New York: McGraw Hill Book Company, 1970), Ch. 9.
salvatoRe, doMiniCK, *International Economics,* 3rd Edn. (Macmillan and Collier Macmillan, 1990), Ch. 19.
salvatoRe sChiavo-CaMPo, *International Economics* (Cambridge: Winthrop Publishing Inc., 1978), Ch. 12.
sodeRsten, Bo., *International Economics* (New York: Macmillan, 1970), Ch. 17.

CHAPTER

36 Balance of Payments: Meaning, Measurement and Adjustment

CHAPTER OBJECTIVES

The objective of this chapter is to discuss how annual account of foreign transactions – balance of payments – is prepared by the trading countries and how imbalance in balance of payments is corrected. The following aspects of balance of payments have been discussed in this chapter.

• Meaning of balance of payments (BOP)
• Purpose of preparing balance of payments
• Accounting system of balance of payments
• Meaning and measurement of BOP imbalance – the BOP disequilibrium
• Causes and consequences of BOP disequilibrium and
• Methods of BOP adjustment

36.1 INTRODUCTION

In Chapter 34, we have discussed the theories of foreign trade involving export and imports of goods and services only. In reality, however, there are two other kinds of important international transactions: (*i*) foreign investment, i.e., inflow and outflow of capital, and (*ii*) income remittances by the migrant labour. All these kinds of international transactions involve inflow and outflow of incomes and capital. For example, exports cause inflow of income and imports outflow of income. Similarly, foreign investment in a country increases its stock of capital and *vice versa*. It may add to national income but it may cause also the outflow of income. So is the case with remittances of labour income. There is no reason to believe that total receipts and payments will always be in balance. Inflow and outflow of remittances may increase or decrease the national income. In case inflows of income and capital exceed outflows, it adds to economic growth of the country. And, in case outflows of income and capital exceed inflows thereof, it causes decline in the domestic economy. Therefore, countries having international transactions make periodic accounting of their foreign transactions to assess the final outcome of the international transactions. The process of accounting is known as the Balance of Payments (BOP) accounting. In this chapter, we discuss the various aspects of BOP and related issues, as pointed out in 'Chapter Objective'.

36.2 MEANING AND PURPOSE OF BALANCE OF PAYMENTS (BOP)

We being our discussion on the *balance of payments* by looking at what it means nad what is its purpose.

36.2.1 Meaning of Balance of Payments

The BOP of a country is a systematic accounting of all economic transactions between the residents of the country and the residents of the rest of the world over a period of time, usually a year. In simple words, the BOP of a country is an account of all international receipts and payments for the accounting year. The accounting year may be a calendar year or a financial year.1 In India, the BOP is prepared for every financial year, although it is also prepared and reported on a quarterly basis.

Although the definition of BOP is quite clear, some terms used in the definition need some clarification.

The term 'systematic accounting' does not refer to any particular system. But the system generally adopted is *double entry book-keeping system*. Under this system, both receipts and payments of an economic transaction—sale and purchase—are recorded in one entry. We know that each economic transaction involves two flows—inflow and outflow. What comes in, i.e., inflow, is treated as a *debit item* and what goes out, i.e., the outflow, is treated as a *credit item*. Under double entry book-keeping system, the values of debit and credit items are recorded simultaneously. For example, in case of exports, goods and services go out of the country and payments received come to the country. What goes out (export) is recorded as a credit item and what comes in (payments received), as a debit item. Both the value of exports and the payments received are recorded in the same accounting entry of the transaction. Similarly, in the case of imports, the value of imports of goods and services is recorded as a debit item and payments for imports are recorded as a credit item. Both the value of imports and the payments made are recorded in the same accounting entry. This is the double entry book-keeping system.

'Economic transactions' include all such transactions that involve the transfer of money (foreign exchange) along with the right of ownership of the transacted goods, money, and assets. Specifically, exports and imports of goods and services, foreign investment, and borrowing from and lending to a foreign country, are all economic transactions. It is important to note here that while in some transactions, there is a physical transfer of goods, services, assets, and money along with the transfer of ownership, in others, no physical transfer is required. For example, if a US resident imports garments from India, they have to be transferred physically along their ownership. But if a US multinational corporation operating in India reinvests its profits in India, there is no physical transfer, i.e., transfer of money.

The term 'resident' refers to the nationality of the persons or organizations carrying out the transaction, not their physical presence in their own country. For example, foreign tourists, diplomats, military personnel, migrant labour and branches of companies in a country are treated as residents of their own nation, not as residents of the country in which they are physically present.

1. In India, the annual balance of payments is prepared for every financial year beginning on 1 April of a year and ending on 31 March of the next year.

36.2.2 Purpose of Balance of Payments

The purpose of preparing balance of payments (BOP) is similar to that of the balance sheet of business firms. We know that all business firms prepare their annual (or periodic) balance sheet of their transactions with the rest of society with the purpose of taking stock of their profit and loss and their assets and liabilities. Likewise, all nations having economic transactions with the rest of the world prepare their periodic BOP accounts just to take stock of their receipts and payments, and their assets and liabilities resulting from their international economic transactions. However, the purpose of BOP goes far beyond this limited purpose as it reveals the overall implications of international economic transactions for the economy of a country. The BOP accounts serve the following purposes.

First, the international economic transactions of a country—its exports, imports, and foreign investments—and inflows and outflows of incomes (e.g., remittances) have a serious impact on the levels of its consumption and the stock of capital. The change in the levels of consumption and of capital accumulation affects the national income of the country and its economic growth over time.

Therefore, in order to find out the impact of the monetary inflows and outflows of on the economy, a country has to prepare its balance of payments.

Second, in case a country has a large and persistent trade deficit, and capital inflows are not sufficient to make up the trade deficit, it faces a serious foreign payment problem. A country having a large and persistent trade deficit has to borrow from abroad to settle its payments. When the country makes a large borrowing from abroad over a long period of time, it tends to fall into international indebtedness. This can have serious adverse repercussions for the of the economy of the country. This is what had happend an India in 1990–91.2

In fact, India's foreign exchange reserves had fallen heavily from $7 billion in 1979–80 to $750 million in 1990–91, i.e., foreign exchange resources had declined form $7 billion to less than $1 billion. The reserves started declining sharply in September 1990, going from $3.11 billion at the end of August 1990 to $896 million on 16 January 1991. 'By June 1991, the BOP crisis had become overwhelmingly a crisis of confidence—of confidence in the government's ability to manage the balance of payment.'3 To tide over the crisis India had to borrow heavily from the IMF.

Third, the balance of payments provides useful data for analysing a country's weaknesses and strengths with respect to foreign trade and its gains and losses from international transactions. The balance of payments contains data on exports and imports, capital inflows and outflows, borrowings and lendings, and foreign exchange reserves for successive years. By analysing the data one can find out whether the economic condition of a country in relation to the world economy is improving or deteriorating. If it is deteriorating, the government will have to formulate appropriate corrective policy measures and implement them on time. 'One of the basic problems of international economic policy is to find effective means of restoring external balance to the country whose balance of

2. For further details, see *Economic Survey 1990–91: Part I: General Review*, Ministry of Finance, Government of India, Section II: The Payments Crisis, pp. 4–10.

3. *Economic Survey*: 1990–91, Part I: General Review, Ministry of Finance, Government of India, Section II: The Payments Crisis, p. 10.

payments is seriously in surplus or in deficit.'4 The balance of payments provides useful data and guidelines for the formulation a suitable foreign trade and exchange rate policy.

36.3 BALANCE OF PAYMENTS ACCOUNTS

As noted above, the purpose of preparing balance of payments is to have an overall picture of international economic transactions of the country with rest of the world over a year or so. Since different kinds of international economic transactions have different kinds of economic repercussions for the economy, for the purpose of BOP accounting, economic transactions between a country and the rest of the world during a specific time period are grouped under two broad categories:

- Current transactions
- Capital transactions

Current transactions include transactions that affect the current level of income and consumption of the country during the current year–the year for which balance of payments is prepared. Current transactions include the following transactions.

(*i*) Exports and imports of goods and services.

(*ii*) Unilateral or unrequited receipts and payments like foreign aid, gifts, donations, etc.

These transactions are treated as current transactions because they have to be settled by the end of that accounting year.

Capital transactions, on the other hand, are transactions that result in inflows and outflows of capital. Capital transactions broadly include the following:

(*i*) Foreign investment—inflows and outflows

(*ii*) Foreign borrowings and lendings

(*iii*) Banking transactions—receipts and payments

Capital transactions change the stock of capital of the country and thereby the production capacity of the country. Therefore, capital transactions have long-term repercussions for the economy. Also, the settlement of capital transactions is spread over a long period of time. Therefore, capital transactions are recorded separately. Besides, another important distinction between current and capital transactions is that current transactions are *flow transactions*, while capital transactions are *stock transactions*.

The items of the *current account* and *capital account* are discussed below in detail.

It may be noted here that the methods of classifying current and capital transactions vary from country to country. We will, however, use the classification of current and capital account transactions followed in India.

36.3.1 Current Account

The items of international transactions that are recorded in the current account of the balance of payments are classified under the following categories:

4. J E Meade, *The Theory of International Economic Policy, Vol. I: The Balance of Payments*, 2nd edn., (London: The English Language Book Society, and Oxford University Press, 1972), p. 3.

(*i*) Merchandise trade—exports and imports

(*ii*) Invisibles, including receipts and payments on account of foreign travels, transportation, insurance, investment income, official and private transfers

(*iii*) Unilateral items or 'unrequited items', including gifts, donations, military aid, technical assistance, etc.

The net balance of merchandise trade, i.e., the excess of exports over imports, is called *balance of trade*. If exports are greater than imports, it shows a *surplus trade balance* and if exports are less than imports, it shows a *deficit trade balance*. Similarly, the net balance of 'invisibles' may show a surplus or a deficit.

The net of trade balance of receipts and payments on account of 'invisibles' and 'unilateral transfers' gives the overall *balance of the current account*, called the *current account balance*. The current account balance may show a surplus or deficit.

36.3.2 Capital Account

As mentioned earlier, the capital account is a record of capital transactions with the rest of the world. It is, in fact, an account of the inflows and outflows of capital. Capital transactions are classified under the following three broad categories:

- Short-term capital movements
- Long-term capital movements
- Change in the stock of gold and foreign exchange reserves

Short-run capital transactions generally include (*i*) the sale and purchase of short-term securities such as treasury bills, commercial bills, and acceptance bills, etc. (*ii*) speculative purchase of foreign currency, and (*iii*) cash balance held by foreigners.

Long-term capital transactions include (*i*) direct investment in shares and bonds, (*ii*) direct investments in real estate and corporate investment in plant, building, machinery, and equipment with investors holding the power to control the investment, (*iii*) portfolio investment, including investments in government bonds and securities, and in private bonds and securities, over which the investor has no controlling power, (*iv*) amortization of capital, i.e., repurchase and resale of securities and bonds earlier sold to, or purchased from, foreigners.

It is important to note here that, unlike the *export* and *import* of goods and services, the *export of capital* is treated as a *debit item* and the *import of capital* as a *credit item* in the balance-of- payments accounting. This kind of treatment of capital export and import is based on the logic that export of capital reduces the stock of foreign exchange and import of capital increases it.

In India's balance-of-payment accounting system, however, *capital account transactions* are classified under the following three categories:

- Foreign investment—net of inflows and outflows of capital
- Net of loans, external assistance, commercial borrowings, rupee debt services, etc.
- Other type of short-term capital transactions—private capital transactions

The sum of the net of all these items gives the *capital account balance*.The capital account balance added to the current account balance gives the *balance of payments*.

36.4 ASSESSMENT OF BALANCE OF PAYMENTS

The balance of payments prepared on the basis of the double-entry book-keeping system shows BOP to be in balance. In reality, however, total foreign receipts and payments are never equal – there is either a surplus or a deficit in the balance of payments. Therefore, a different system is adopted to assess the actual balance of payments at the end of the accounting year. In this section, we proceed to discuss the method of assessment of the final status of the balance of payments of a country. The method of assessing the final status of the balance of payments will be illustrated by detailing the method used in India.

36.4.1 Method of Assessing BOP

Having described the nature and components of the current and capital accounts of the balance of payments, we now show the method of assessing the final status of the balance of payments. This is important from the view point of both BOP analysis and policy formulation. Three kinds of balances are worked out to assess the BOP balances.

- Balance of trade
- Current account balance
- Balance of payments, i.e., the overall balance

The method of measuring these balances is shown below.

1. Current account balances

(*a*) Balance of trade = Exports – Imports = Net merchandise trade

(*b*) Current account balance = Balance of trade + Net transfer payments and receipts

2. Capital account balances

(*a*) Net long-term capital flows = Net foreign investment

= Capital inflows – Capital outflows.

(*b*) Net short-term private capital flows = Capital inflows – Capital outflows

(*c*) Capital account balance = Net foreign investment + Net short-term capital flows

= (*a*) + (*b*)

3. Balance of payments

Balance of payments = Capital account balance + Current account balance

The overall balance of payments may show a positive balance or a negative balance, or in other words, a surplus or a deficit. When BOP shows a surplus, it is considered to be a favourable BOP and if it shows a deficit, the BOP is taken to be unfavourable. A favourable BOP shows a rise in foreign exchange (Forex) reserves, and an unfavourable BOP shows a rise in foreign debt. The method of BOP accounting used is India is described below.

36.5 BOP ACCOUNTING SYSTEM IN INDIA

The accounting system of BOP is illustrated here by the method used in India—first by the current account and then by the capital account. Table 36.1 shows ***India's current account***.

Table 36.1 *India's Current Account: 2009–10 to 2011–12*

(` in crore)

S.N.	*Item*	*2009–10*	*2010–11*	*2011–12 (PR)*
1	**Merchandise Trade**			
2	Exports (f.o.b.)	863300	1165700	1482500
3	Imports (c.i.f.)	1423200	1746200	2394600
4	**Balance of Trade (2 – 3)**	**–559900**	**–580500**	**–912100**
5	**Invisibles:**			
6	Receipts	774600	867200	1053500
7	Payment	394400	506400	517300
8	**Net Invisibles (6 – 7)**	**380200**	360800	536200

9

Current Account Balance (4 + 8)

–179700

–219700

–376000

PR = Provisionally revised.

Source: *Economic Survey–2013–14*, MOF, Government of India, Statistical Appendix, pp. 63–64.

As shown in Table 36.1, in India, current account transactions are of two kinds, viz., merchandise trade, i.e., exports and imports, and invisibles. The net of merchandise trade (exports less imports) gives the *balance of trade*. The difference between receipts and payments on account of invisible transactions give the *net of invisibles*. The sum of the trade balance and net of invisibles gives the *current account balance*. In other words, the current account balance is the sum of trade balance and net invisibles. The current account balance is adjusted in the capital account to obtain the final picture of the balance of payments.

Let us now look at items recorded in India's capital account and the process of finding the final status of the country's balance of payments. The procedure for estimating the ***capital account balance*** is illustrated with India's capital account. The structure of India's capital account is shown in Table 36.2. As shown in the table, India's capital account contains four major items of capital transactions:

• Foreign investments

• Foreign loans

• Banking capital transactions (inflows and outflows)

• Other capital inflows and outflows

As the table shows, the net of inflows and outflows of foreign investment gives the net foreign investment. Similarly, the net of all capital transactions (inflows and outflows) are worked out. The sum of the net of all capital flows gives the capital account balance. Once the current account balance and the capital account balance are worked out, the overall balance of payments can be obtained easily by adding up the capital account balance and current account balance. The sum total of the *capital account balance* and the *current account balance* give the *overall balance of payments* of the county. As shown in Table 36.2, item number 7 gives the capital account balance. The current account balance from Table 36.1 has been transferred to Table 36.3. The sum of these two give India's overall BOP for the financial year 2009–10. This is illustrated in Table 36.3.

Table 36.2 *India's Capital Account and Overall Balance: 2010–11*

(` in crore)

S.N. Item

Amount

Net amount

1

Foreign investment:

(*a*) Net FDI

54100

(*b*) Net-portfolio investment

139400

(*c*) Net Foreign investment

193500

2

Loans:

(*a*) External Assistance (net)

22500

(*b*) Commercial borrowings (net)

110200

(*c*) Net loans [(*a*) + (*b*)]

132700

3

Banking capital:

	(*a*) Receipts	419300
	(*b*) Payments	397300
	(*c*) Net banking capital	22000
4	Rupee debt services	– 300
5	Other capital flows:	
	(*a*) Receipts	45200
	(*b*) Payments	101900
	(*c*) Net other capital flows	– 56700
6	Errors and omissions	–12100
7	**Capital Account Balance (Net of 1 to 6)**	**279100**

Source: *Economic Survey* 2013–14, MOF, Government of India, Statistical Appendix, p. 63.

As shown in Tables 36.2 and 36.3, India had a *surplus balance of payments* in the financial year 2010–11. In year 2008–09, however, India had a high deficit balance of payments of `97,115 crore. It proves the point that the BOP is not always in balance or in equilibrium.

Table 36.3 *The Overall Balance of Payments of India: 2010–11*

Item	` *(crore)*
Capital account balance	279100
Current account balance	–219700
Overall balance of payments	59400

In general, the BOP of most countries shows either a deficit or a surplus. In either case, the BOP of the country is said to be in disequilibrium. A large and persistent BOP deficit or surplus—the BOP disequilibrium—has serious implications for the economy as a whole. Therefore, the concept and adjustment of BOP disequilibrium has become an important issue from both a theoretical and practical point of view.

36.6 BOP EQUILIBRIUM AND DISEQUILIBRIUM

The ultimate objective of BOP accounting is to assess whether BOP is in *balance* or in *imbalance*. In technical terms, the purpose of BOP accounting is to find whether BOP is in ***equilibrium*** or in ***disequilibrium***. In this section, we define the concept of *BOP equilibrium* and *disequilibrium* and the purpose behind the BOP assessment.

As noted in Sec. 36.4, BOP is assessed as *the sum of capital account balance and current account balance*. As such, in final assessment,

BOP = Capital Account Balance + Current Account Balance

Given this method of BOP assessment, the *BOP equilibrium* and *disequilibrium* can be defined as follows. *In case the sum of the capital account balance and current account balance of a country is equal to zero, its BOP is said to be in* ***equilibrium.*** And in case *the sum of the capital account balance and current account balance of a country is greater than zero, its BOP is said to be in* ***disequilibrium.***

As defined above, BOP disequilibrium may be in the form of *surplus* or *deficit*. If BOP disequilibrium shows a surplus or deficit, it involves the problem of BOP adjustment.

Looking at the problem from BOP adjustment point of view, if BOP shows a small and manageable surplus or deficit, the BOP disequilibrium is not a serious economic problem. However, if BOP disequilibrium shows a high level of surplus or deficit persisting over time, it raises a serious problem of BOP adjustment. Nevertheless, if BOP disequilibrium is of ***surplus kind***, it is not a serious problem for the country as it can be sustained for some period. But if BOP disequilibrium is of ***deficit kind*** and deficit is of significantly high level, it becomes a serious problem for the economy and for the government. It creates a serious problem for the government of a country facing deficit kind of BOP disequilibrium. The government is required to devise appropriate policy measures to correct the BOP disequilibrium. From policy formulation point of view, a different method is applied to assess the extent of BOP disequilibrium and to identify the sources thereof. The method of assessing the BOP disequilibrium will be discussed in the next section.

36.7 ASSESSMENT OF BOP DISEQUILIBRIUM

The BOP disequilibrium may be in the form of BOP surplus or BOP deficit. For assessing the BOP position of a country, all international transactions—current and capital transactions—are taken into account. For the purpose of BOP assessment, all international transactions are grouped under the following two categories:

- Autonomous transactions
- Induced or accommodating transactions

36.7.1 Autonomous Transactions

Autonomous transactions are transactions that take place automatically due to natural human need and desire for consuming more and better goods, and to make the best possible use of resources for higher profits and incomes. Autonomous transactions generally appear in the form of exports and imports of goods and services. The autonomous exports and imports are necessitated by the following economic reasons:

- The scarcity of necessities in the domestic market, e.g., food and clothing
- The demand for such inputs as oil, industrial raw materials, etc.
- The need for better and less costly machinery, equipment, and technology to increase the production capacity of the economy
- The desire for more and better goods and services

All these kinds of transactions are recorded in the current account of the balance of payments. In other words, autonomous transactions are essentially current account transactions.

36.7.2 Induced or Accommodating Transactions

The autonomous transactions between a nation and the rest of the world are not always in balance. Exports and imports are not always equal in value terms. It is the general experience of nations that $X \neq M$. Either $X > M$ or $X < M$ in value terms. That is, *receipts from exports* and *payments for imports* are not equal always. In case $X \neq M$, there is a need for payments by the countries with a deficit. The need for payments generates short-term capital movements in the form of borrowing or lending—the deficit country borrows and the surplus country lends in order to settle the payments. Such short-term borrowing and lending is induced by foreign trade—exports and imports. That is why these transactions are called *induced transactions* or *accommodating transactions*.

Autonomous and induced transactions arise also on account of long-term capital transactions. The long-term foreign investments, i.e., exports and imports of capital, are treated as autonomous transactions. Also, short-term overseas investments motivated by the desire for higher returns fall in the category of autonomous transactions. But short-term capital movements in the form of gold movements and accommodating capital movements on account of autonomous transactions are treated as *induced transactions*.

Having explained the concept of *autonomous* and *induced* transactions, let us now look at the method of assessing a disequilibrium in the balance of payments.

36.7.3 Method of Assessing a BOP Disequilibrium

In assessing the BOP equilibrium or disequilibrium, only autonomous transactions of both current and capital accounts are taken into account. If total receipts and payments on account of autonomous transactions in both capital and current accounts are in balance, the BOP is said to be in equilibrium. However, as noted above, the receipts and payments on account of autonomous transactions are hardly ever in balance—they are usually unequal.

Therefore, in most cases, the BOP is in disequilibrium.

The method of assessing the BOP disequilibrium has already been illustrated in Tables 36.1 and 36.2. As shown there, a BOP disequilibrium is assessed by combining the autonomous transactions in the current and capital accounts. In the process of assessing the BOP disequilibrium, first the current account balance of autonomous transactions—the balance of exports, imports, and unilateral payments—is estimated. Then, the capital account balance is worked out on the basis of short-term and long-term capital transactions. Finally, the current account balance (surplus or deficit) and capital account balance are summed up to assess the BOP status of the country. If the sum of the current and capital account balance is negative, it shows a BOP disequilibrium of deficit nature. If the sum of the autonomous transactions of the current and capital account balance is positive, it shows a BOP disequilibrium of surplus nature. Thus, the BOP accounting may result in either a deficit or a surplus. The BOP deficit is financed through foreign borrowing and a BOP surplus goes towards foreign lending. Foreign borrowings and lendings bring the balance of payments to equilibrium. 'A country is said to be in balance of payments equilibrium when the sum of its current, capital, and non-reserve

financial accounts equals zero, so that the current plus capital account balance is financed entirely by international lending without reserve movements.'[5]

As noted above, the BOP disequilibrium may appear in the form of a BOP surplus or a BOP deficit. A surplus BOP disequilibrium is reflected in a substantial rise in the gold and foreign exchange reserves of the country and external lending. And a deficit BOP disequilibrium is reflected in a substantial depletion of gold and foreign exchange reserves of the country and external borrowing. However, a deficit disequilibrium is a matter of great concern for a country. This is because a large and persistent deficit in the BOP affects the economy of the country adversely. Therefore, nations facing a deficit BOP disequilibrium must adopt suitable policy measures to correct it. These measures are discussed in detail in a forthcoming section. Let us now discuss the causes and kinds of BOP disequilibrium.

36.8 CAUSES AND KINDS OF BOP DISEQUILIBRIUM

As noted above, the BOP disequilibrium may be in the form of a surplus or deficit arising out of autonomous current and capital account transactions. The BOP disequilibrium of surplus kind does not create a big problem for the economy. In long run, however, it may cause inflation due to overspending by the government and currency appreciation, which affects exports adversely. Inflation and currency appreciation do affect the economy adversely but these problems are manageable. But BOP disequilibrium of deficit nature creates serious problems for the economy, which are often difficult to manage. Therefore, we will confine our discussion here to the causes and kinds of BOP disequilibrium of deficit nature. Let us now look at the causes and kinds of BOP disequilibrium with reference to deficit in balance of payments.

36.8.1 Causes of Trade Deficits

As noted above, BOP disequilibrium of deficit nature arises when autonomous payments far exceed autonomous receipts. Autonomous payments include (*i*) payments for the import of goods and services, and (*ii*) capital outflows, i.e., investments abroad. Similarly, autonomous receipts include (*i*) export earnings, and (*ii*) capital inflows. If autonomous capital outflows and inflows are in balance or capital outflows and inflows are not significantly different, the BOP disequilibrium is caused by current account deficits. The current account deficit arises mainly because of trade deficits, i.e., imports exceeding exports of goods and services. In fact, the first and the most important cause of BOP disequilibrium is imports exceeding exports significantly over time. Thus, a trade deficit is caused mostly by a high rise in imports and a slow rise or decline in exports. The imports of a country, especially of a rapidly developing country, increase for the following reasons:

- A high rate of economic growth leading to a rise in demand for industrial inputs.
- A high rate of inflation compared to foreign countries.
- An increase in the overall demand for foreign goods and services.
- A high income elasticity of demand for foreign goods.
- A low price elasticity of demand for imported goods.
- People's preference for foreign goods.

5. Paul R Krugman and M Obstfeld, *International Economics: Theory and Policy* 6th edn., (Pearson Education, 2003), p. 538.

Countries confronted with these kinds of unfavourable conditions face trade deficits, i.e., a deficit BOP disequilibrium.

36.8.2 Kinds of Disequilibrium

The kinds and causes of BOP disequilibrium are its important aspects from the view point of formulating policies required for correcting it. Here we describe briefly the kinds of BOP disequilibrium.

1. Fundamental disequilibrium When the BOP disequilibrium is caused by a high rate of persistent inflation, it is referred to as a *fundamental disequilibrium*. This kind of BOP disequilibrium is 'fundamental' because it is based on the fundamental laws of demand and supply under a free trade system. A high rate of long-term inflation causes a significant change in the relative prices of imports and exports. Inflation makes domestic goods costlier than imports. If the price elasticity of demand for foreign goods is significantly high, the demand for imported goods increases, given the exchange rate. On the other hand, inflation makes domestic goods costlier for foreigners. Therefore, exports continue to decrease, given the price elasticity of demand in foreign countries. With increasing imports and decreasing exports, the trade deficit keeps widening over time and becomes permanent. A long-run, large, and persistent trade deficit causes an 'obdurate' BOP disequilibrium—one that tends to persist.

2. Cyclical disequilibrium A BOP disequilibrium that arises because of global business cycles is called a *cyclical disequilibrium*. A business cycle refers to an intermittent tendency for growth and depression in an economy—it may even be global like the Great Depression of the 1930s following rapid global growth after World War I. The Global Depression (second only to the Great Depression) of 2008–09 originating in the United States, is another evidence of the repetition of business cycles. During a business cycle, there is a phase of rapid growth followed by a sharp decline in economic activities. An important consequence of business cycles is the foreign trade cycle, since the impact of global business cycles is different on different countries. For example, during the recent global depression of 2008–09, the US and European countries suffered a great deal, while China and India were least affected. Therefore, imports and exports of different countries are affected in different ways. While some countries enjoy a trade surplus, others suffer from trade deficits. As a result, different countries face different kinds of BOP disequilibrium. While some countries have a deficit BOP disequilibrium, others have a surplus BOP disequilibrium. This kind of BOP disequilibrium is called *cyclical disequilibrium*. An important aspect of a cyclical BOP disequilibrium is that it is generally self-correcting.

3. Structural disequilibrium Another major cause of BOP disequilibrium is the structural change in the domestic or foreign economy. Structural changes in the economy may be caused by such factors as depletion of natural resources (coal, iron, oil and other minerals), change in technology, change in the industrial structure of an economy, change in consumer preferences and choices, etc. Such change, if not accompanied by changing demand patterns, reduces the competitive strength of exporting countries in the international market either owing to a high cost of production or a decrease in foreign demand. For example, exhaustion of good quality coal seams in Great Britain in the middle of 20th century converted the country from a net exporter to a net importer of coal.6 Look at another example. The introduction and extensive

6. P.T. Ellsworth, *International Economy,* (London: Collier Macmillan, 1972), p. 235.

use of nylon in the US affected Japanese silk exports heavily. These kinds of changes affect the trade pattern. While some nations gain, others lose and, therefore, while some nations see BOP deficits, others have a BOP surplus. This is called *structural BOP disequilibrium.*

4. Temporal disequilibrium Many countries are often confronted with BOP disequilibrium for a short period. This is known as *temporal BOP disequilibrium.* Some major factors which cause temporal BOP disequilibrium are the following:

- Seasonal crop failures owing to heavy rainfall and droughts, especially in countries producing mainly primary goods. India is one such example.
- Sudden economic depression that causes a sharp decline in the GDP of the country forcing a rapid rise in imports and sharp decline in exports due to a sudden decline in the domestic production.
- Ambitious development programmes requiring heavy imports of industrial inputs, technological know-how, machinery, and equipment.

• Change in consumer choices and preferences owing to the better quality of foreign goods and/or the demonstration effect of advanced countries on the consumption pattern of developing economies, leading to high imports and trade deficits.

36.9 IMPLICATIONS OF BOP DISEQUILIBRIUM

As noted earlier, the BOP disequilibrium, whether with a BOP deficit or surplus, has implications for the economy. The nature and seriousness of the problems arising out of a BOP disequilibrium may be different for different countries at different times. The BOP disequilibrium also creates international problems. In this section, we discuss the major implications of BOP disequilibrium. A ***deficit BOP disequilibrium*** has the following implications for the economy.

(i) Rise in international indebtedness. A large and persistent BOP disequilibrium caused by trade deficits leads to a rapid rise in international borrowings and international indebtedness. For example, look at the increasing trade deficits of India during the last decade and increasing external debts. As shown in Table 36.4, the trade deficit of India has continuously increased over the past decade—from US$10.7 billion in 2002–03 to US$189.7 billion in 2011–12.

Owing to the increasing trade deficit, the external debt of the country has increased over this period, as shown in Table 36.5. The overall external debt of India increased almost continuously from US$101.3 billion in 2001 to US$404.9 billion in 2013. This increase was not owing mainly to an increase in trade deficits but was an important factor in causing an increase in external debt. A consolatory factor is that India has comfortable foreign exchange reserves.

(ii) Adverse impact on economic growth. Another and a very serious implication for a country facing a large and persistent BOP deficit is that its GDP growth rate is affected adversely. This is because a large and persistent deficit acts to restrain the import of industrial inputs and advanced technology required for rapid economic growth. As a matter of fact, this has been the case for most underdeveloped countries struggling to accelerate the pace of their economic growth. Owing to their backward technology, the production and export of underdeveloped countries are largely limited to labour-intensive and primary goods. Prices of such goods are comparatively low. Therefore, their export earnings are also lower in spite of a substantial increase in their exports.

Table 36.4 *India's Current Account and External Debt: 2000–01 to 2009–10*

($ in bn)

Year	*Exports*	*Imports*	*Trade balance*
2002–03	53.8	64.5	–10.7
2003–04	66.3	80.0	–13.7
2004–05	85.2	118.9	–33.7
2005–06	105.2	157.1	–51.9
2006–07	128.9	190.7	–61.8
2007–08	166.2		

		256.6	–91.4
2008–09	189.0	308.5	–119.5
2009–10	182.2	300.6	–118.4
2010–11	256.2	383.5	–580.5
2011–12 (PR)	309.8	499.5	–189.7

Source*: Economic Survey* 2006–07 (Table 6.2, p. 108), 2008–09 (Table 6.2, p. 127) and Economic survey–2013–14, Statistical Appendix, pp. 63–64.

Table 36.5 *External Debt of India: 2000–10*

Year	*External debt (US$ in billion)*
2001	101.3
2002	98.8
2003	104.9
2004	112.7
2005	134.0
2006	139.1
2007	172.4
2008	224.4
2009	224.5
2010	262.3
2011	317.9
2012	360.8
2013 (PR)	404.9

Source: *Economic Survey, 2013–14, Statistical Appendix, p. 601.*

(iii) Rise in unemployment. Large and persistent BOP deficits lead to outflows of income from a country. This kind of situation limits the savings and investments of a country and lowers the growth rate of the economy. Lower growth accompanied with rapid increase in the labour force leads to growth of unemployment. Besides, the low growth of exports limits the expansion of export-oriented industries, limiting the prospects for more employment.

(iv) Financial crisis. An overall implication of a large and persistent BOP deficit over a long period of time is that it leads to financial crisis. As noted above, a persistent trade deficit leads to large external debt. In such a case, not only the borrowing power of the country gets reduced, but also the international organizations like the World Bank and

7. Estimate made by Tang Shuangning, China Everbright Group.

International Monetary Fund, and foreign countries (even those that are financially strong) become reluctant to offer financial help. This was the experience of India in 1991 and of the Asian tigers in the 1980s.

Let us now look at a the effects of ***BOP disequilibrium*** of surplus nature. The most important implication of BOP disequilibrium surplus nature is a rise in the international economic power and strength of the country. For example, China has had a surplus BOP disequilibrium in the 21st century. It had a trade surplus of about $29 billion in July 2010—one of the highest in the recent past. China's exports had risen by about 38 per cent and imports by about 28 per cent. The trade surplus would have led to a high appreciation of the Chinese currency (renminbi, or RMB), also known as yuan, but it did not, because China controlled the exchange rate. This was one of the reasons why China's exports had risen at a higher rate than its imports. This affected exports of the US, China's main trade partner, and led to a depreciation of the US dollar. The US, experiencing a critical economic condition, put pressure on China to allow its currency to fluctuate in tandem with market conditions. This led to depreciation of the Chinese currency, the yuan, against the US dollar, though marginally and over a long period of time. The yuan appreciated to 6.5 yuan per dollar after 17 years. This did affect China's exports, but only marginally. Furthermore, China's trade surplus pushed its foreign exchange reserves beyond US$3 trillion in 2010 which crossed the reasonable limit of US$1.3 trillion.[7] According to Zhou Xiaochuan, governor of China's central bank, China's foreign exchange reserves exceeded the reasonable requirements of the country. These excessive reserves have led to an appreciation of the yuan and affected China's exports adversely.

Another effect of BOP disequilibrium of surplus nature is *inflation*. BOP surplus generally leads to overspending by the government of developing countries. This leads to inflation. A high rate of inflation is a matter of concern.

In conclusion it may be said that although a surplus or deficit BOP disequilibrium has its own economic implications, the problems caused by a BOP deficit affect countries more severely.

36.10 POLICY MEASURES FOR BOP ADJUSTMENT

36.10.1 Why Policy Measures

The implications of the adverse balance of payments have already been discussed in detail in the previous section. As noted there, short-term and small deficits in the balance of payments are quite likely in a wide range of international transactions. Since such deficits do not harm seriously the economy or the economic status of the country, they do not call for immediate corrective actions. More importantly, intermittent short-term changes in domestic economic policies with a view to removing the short-term deficit in the balance of payments may do more harm than good to the economy by causing malallocation of resources in the process of re-allocation of resources. This may cause short-term fluctuations in the economy. The short-term deficits of smaller magnitude are, therefore, not a matter of serious concern for the policy-makers. But, a chronic deficit of a large magnitude has a wide range of economic and political implications. A chronic deficit indicates either a country turning an eternal borrower or depletion of its foreign exchange and gold reserves. Such a country loses its

international liquidity and credibility. Such a situation often leads to a compromise with economic policies and often forces drastic changes in macroeconomic policies of the country. For instance, India faced a similar situation in July 1990 when her forex reserves had fallen dangerously low – it was sufficient to meet import bill of 15 days' imports whereas a forex reserve sufficient to pay for 90 days' imports is required. India had to borrow heavily from abroad and make drastic changes in her economic policies.

A country facing chronic and large deficits in its balance of payments is forced to adopt corrective measures; to bring necessary changes in its internal economic policies for wiping out the deficit, or at least to bring it to a manageable size. It is a widely accepted view that the conditions for the automatic corrective mechanism based on international price-mechanism do not offer a quick and desirable solution. Therefore, the government has no option but to intervene with the market conditions of demand and supply using suitable measures available to them.

The measures that are generally adopted to correct the BOP disequilibrium can be classified as follows: (*i*) **policy measures,** and (*ii*) **direct control measures**. In general practice, policy measures are adopted first for BOP adjustment. But, when policy measures are ineffective, then direct control measures are adopted.

The working and effectiveness of policy measures and direct control measures are discussed here briefly:

36.10.2 Policy Measures to Correct Adverse BOP

Under the free trade system, BOP deficit arises either because aggregate domestic demand for goods and services8 is greater than the domestic supply (prices of domestic and foreign goods being comparable) or because domestic prices are significantly higher than the foreign prices. Thus, the deficit may be removed either by increasing domestic production at an internationally comparable cost of production or by reducing excess demand9 or by using the two methods simultaneously. It may be very difficult to increase the output in the short-run, especially when a country has a near full employment, or when there are other production constraints. Therefore, the only way to reduce a deficit is to reduce the demand for foreign goods. Demand for foreign goods depends on mainly two factors – income and prices. Therefore, generally two kinds of policy measures are adopted: (*i*) **income control measures** and (*ii*) **price control measures.**

(*i*) *Income Control Policies*

Demand may be reduced through a change in its main determinants, viz., income10 and price. Here we discuss how reduction in income can lead to reduction in demand and how it helps in reducing the deficit in the balance of payments. The two policy tools that are applied to change disposable income are monetary and fiscal policies.11 *Monetary policy* operates on the demand for and supply of money while *fiscal policy* operates on the disposable income of the people. The working and efficacy of these policies as instruments of solving the balance of payments problem are described below.

8. Excluding the movement of autonomous capital flows.

9. *See* also, Sodersten, *International Economics* (Macmillan, 1970), pp. 272–77.

10. The term ‘income’ here means ‘disposable income’.

11. For the nature and working of monetary and fiscal policies, see Chapter 27.

Monetary Policy The instruments of monetary policy include discount (or bank) rate policy, open market operations, statutory reserve ratio, and selective credit controls. Of these, the first two instruments are generally adopted in the context of balance of payment policy. This, however, should not mean that other measures are not relevant.

To solve the problem of BOP deficit, a ‘tight money policy’ or dear money policy is adopted. Under ‘dear money’ policy, central monetary authorities raise the interest rate. Consequently, under normal conditions, the demand for bank credit for investment consumer durables decreases. With a fall in investment and through its multiplier effect, the income of the people decreases. If marginal propensity to consume is greater than zero, demand for goods and services decreases. The decrease in demand implies also a simultaneous decrease in imports, imports being income-elastic. This is how ‘a tight money policy’ is supposed to correct deficit in the balance of payments.

The efficacy of ‘tight money policy’ is, however, doubtful under the following conditions: (*i*) when rates of returns are much higher than the increased bank rate, and (*ii*) when investors have already effected their investment in anticipation of an increase in the rate of interest. The tight money policy is then combined with *open market operations*, i.e., sale of government bonds and securities. These two instruments together reduce the demand for capital, on the one hand, and demand for consumer goods, on the other. If all goes well, the BOP deficit tends to decrease.

Fiscal Policy Fiscal policy as a tool of changing household incomes includes (*i*) taxation and (*ii*) public expenditure. Taxation reduces household disposable income. Direct taxes directly transfer household incomes to public coffers and thus reduce the overall demand for consumer goods, both domestic and imported. Direct taxes reduce personal savings directly in a good measure while indirect taxes do it in a relatively small measure. Taxation can be used to curtail investment by taxing capital incomes at progressive rates. As regards the indirect taxes, e.g., sales tax and excise duty, their effect on imports is uncertain.

The government can reduce income and demand by adopting a policy of *surplus budgeting*, i.e., the government keeps its expenditure less than its revenue. When the government adopts a *surplus budget policy*, it increases tax rates and reduces public expenditure. Increase in taxation and decrease in public expenditure both reduce the disposable income of the households at their multiplier effect. For surplus budget policy to be effective, it is necessary that the surplus is so large that the total cumulative effect of taxation on disposable income and the effect of reduced public expenditure is considerably large. Whether the reduction in income achieves a certain target of reducing the BOP deficit depends on the foreign trade multiplier.12

(ii) Price Control Measures: Exchange Depreciation and Devalution

Reducing demand for imports through price measures requires changing the relative prices of imports and exports. Relative prices of imports and exports can be changed through (*a*) exchange depreciation and (*b*) devaluation.13

Exchange depreciation refers to a fall in the market value of home currency in terms of foreign currency and ***devaluation*** refers to government's deliberate devaluation of home

12. For a detailed discussion, interested readers may see Ellsworth *op. cit.,* Ch. 15, Kindleber, *op. cit.* , Ch. 20 and Sodersten, *op. cit.* , Ch. 15.

13. Other measures to change relative price are imposition of import duty and subsidization of export.

currency in terms of gold and reserve currency. However, in terms of purchasing power parity theory both turn out to be the same and their impact on foreign demand is also the same. Therefore, we shall consider them as one in their role as measures for correcting adverse balance of payments.

Devaluation and exchange depreciation change the relative prices of imports and exports—import prices increase and export prices decrease, though not necessarily in the proportion of devaluation. As a result of change in the relative prices of exports and imports, the demand for imports decreases in the country which devalues its currency, and foreign demand for its produce increases provided foreign demand for imports is price elastic. Thus, if devaluation or exchange depreciation is effective, imports decrease and exports increase. The country's payment for imports would decrease and export earnings would increase. This ultimately decreases the deficits in the balance of payments in due course of time.

However, whether expected results of devaluation or exchange depreciation are achieved or not depends on the following conditions.

First, the most important condition in this regard is the Marshall-Lerner condition. The Marshall-Lerner condition states that devaluation will improve the balance of payments only if the sum of elasticities of home demand for imports and that of foreign demand for its exports is greater than unity.14 If the sum of elasticities is less than unity, the balance of payments can be improved through revaluation instead of devaluation.

Secondly, devaluation can be successful only if the affected countries do not devalue their currency in retaliation.

Thirdly, devaluation must not change the cost-price structure in favour of imports.

Finally, the government must ensure that inflation, which may be the result of devaluation, is kept under control, so that the effect of devaluation is not counter-balanced by the effect of inflation.

36.10.3 Direct Measure: Exchange Control

We have so far discussed the policy measures, i.e., the indirect measures, for adjusting the adverse balance of payments—the measures which work through the market mechanism. We now turn to discuss one of the most important direct control measures, i.e. exchange control. The exchange control refers to a set of restrictions imposed on the international transactions and payments, by the government or the exchange control authority. Exchange control may be partial, confined to only a few kinds of international transactions, or total, depending on the requirement of the country.

The main feature of a full-fledged exchange control system may be as follows. The government acquires, through legislative measures, powers to control the foreign exchange transactions. It monopolises the purchase and sale of foreign exchange. The sale and purchase of foreign exchange by the resident individuals is prohibited by law. Even holding foreign exchange without informing the exchange control authority is declared illegal as was the case in India during 1970s and 1980s. All payments to foreigners and receipts from them are routed through the exchange control authority or the authorized agents. Foreign exchange payments are restricted, generally, to the import of essential goods and services such as food items, raw materials, other essential industrial imports

14. For further details, see Kindleberger, *op. cit.*, pp. 328–32, and Sodersten, *op. cit.*, pp. 278–81.

like petroleum products, etc. A system of rationing is adopted in the foreign exchange allocation for essential imports. To ensure the effectiveness of the exchange control system and to prevent any possible evasion, stringent laws (like India's FERA and COFEPOSA) are enacted and a circuitous legal procedure of acquiring import and export licences is brought into force. In the process, the convertibility of the home currency is sacrificed.

Why Exchange Control? With its features mentioned above, the exchange control system as a measure of adjusting adverse balance of payments, differs radically from the indirect control measures. While the indirect control measures work through the market forces—demand and supply conditions —the exchange control works through a control mechanism based on adhoc rules and regulations. In contrast to the self-sustained, automatic functioning of the market system, exchange control requires a cumbersome bureaucratic system of checks and controls. Yet, many countries facing balance of payment deficits opt for exchange control for lack of other options.

In fact, automatic adjustment15 in the balance of payments requires the existence of the following conditions:

(*i*) international competitive strength of the deficit countries is fairly high,

(*ii*) elasticity of demand for imports is fairly high,

(*iii*) international market mechanism is perfectly competitive, and

(*iv*) there is no government intervention with market conditions.

The existence of these conditions has always been doubtful. Owing to differences in resource endowments, technology and the level of industrial growth, countries differ in their economic strength and their industries lack international competitiveness. The protectionist policies adopted by various countries intervene with the international market mechanism. Besides automatic method of balance of payments adjustments requires a strict discipline, economic strength, and political will to endure the destabilizing shocks which the automatic method is expected to bring to a country in the process of adjustment. Since these conditions rarely exist, the efficacy of the international price mechanism to bring about automatic balance of payments adjustment is often doubted.

For these reasons, exchange control remains the last resort for the countries under severe strain of BOP deficits. The exchange control is said to possess a superior effectiveness in providing solutions to BOP deficit. Besides, it insulates the economy against the impact of economic fluctuations in foreign countries. Another positive advantage of exchange control lies in its effectiveness in dealing with the problem of capital movements. The government's monopoly over foreign exchange can effectively stop or reduce the capital movements by simply refusing to release foreign exchange for capital transfer. Many countries, i.e., Germany, Denmark and Argentina adopted exchange control during the 1930s because of this advantage.16

Although the exchange control is positively a superior method of dealing with disequilibrium in the balance of payments, mainly in the less developed countries, it does not provide a permanent solution to the basic causes of deficit problem.

15. As one under the Gold Standard System.

16. Ellsworth P.T., *op. cit.*, p. 370.

SUMMARY

• The balance of payments (BOP) of a country is a systematic accounting of all economic transactions between the residents of the country and the residents of the rest of the world over a period of time, usually a year.

• The purpose of BOP is similar to that of a balance sheet of business firms. All nations having economic transactions with the rest of the world prepare their periodic BOP accounts just to take stock of their receipts and payments, and their assets and liabilities resulting from their international economic transactions.

• For the purpose of BOP accounting, economic transactions between a country and the rest of the world during a specific time period are grouped under current transactions and capital transactions.

• The balance of payments is based on double entry book keeping, in which both sides of a transaction—give and take or payments and receipts—are recorded in the book. Under the double entry accounting system, the balance of payments is always in balance.

• In reality, however, there does arise an imbalance in the balance of payments because (*i*) on the current account, the exports of a country are rarely equal to its imports — there is usually a surplus or a deficit, (*ii*) on the capital account, foreign investment in a country is not equal to its investment abroad, and (*iii*) short-run capital inflows and outflows are not always equal.

• The methods of assessing balance of payments include the assessment of (*i*) balance of trade, (*ii*) current account balance, and (*iii*) balance of payments, i.e., the overall balance.

• In India, current account transactions are of two kinds, viz., (*i*) merchandise trade, i.e., exports and (*ii*) imports, and invisibles. The net of merchandise trade (exports less imports) gives the balance of trade.

• Capital account balance is assessed on the basis of (*i*) net foreign investment, and (*ii*) net foreign loans.

• The sum of the capital and current accounts gives the actual balance of payments—surplus of deficit.

• The double entry book-keeping system shows that the BOP is always in equilibrium. But this does not mean that the BOP of a country is always in equilibrium. There is a specific system of assessing whether BOP is in equilibrium or in disequilibrium.

• The BOP disequilibrium is assessed on the basis of only ***autonomous transactions*** (including both trade and capital transactions). Autonomous transactions are those that are carried out to meet the basic needs of the country and with a business motive. If autonomous receipts and payments are not in balance, the BOP of payment is in disequilibrium. The BOP disequilibrium may be of a surplus or deficit kind.

• The BOP disequilibrium is caused by factors that determine foreign trade and capital flows. If factors are favourable, the BOP disequilibrium is favourable, and if factors are unfavourable, the BOP disequilibrium is unfavourable.

• The BOP disequilibrium is classified as (*i*) fundamental disequilibrium, (*ii*) cyclical disequilibrium, and (*iii*) structural disequilibrium. Whatever the kind of BOP disequilibrium, it has serious implications for the economy.

• If BOP disequilibrium is of the deficit kind, it has serious and long-term adverse implications for the economy. The BOP disequilibrium of a surplus kind, within a reasonable limit, does have some positive effects and favourable implications for the economy. But beyond this reasonable limit, this too has an adverse impact on the economy. The most common effect is inflation.

REVIEW QUESTIONS

1. What is meant by the term balance of payments? What is the purpose behind preparing balance-of-payments accounts?
2. What accounting system is used in balance-of-payments accounting? Explain the double-entry book-keeping system with appropriate examples.
3. Why is the balance of payments always in balance? Does it mean that a country always has its foreign receipts equal to its foreign payments? If not, why not?
4. Distinguish between the balance of trade and the balance of payments? How does the balance of trade affect the balance of payments?
5. What is the different between the current account and the capital account? What are the different items in the capital and current accounts? How is the current account balance and current account balance derived?
6. What is the method of assessing the balance of payments of a country? Explain the system by applying the accounting system used in India.
7. What is meant by a balance of payments equilibrium and disequilibrium? How is equilibrium of the balance of payments assessed?
8. What is meant by autonomous and accommodating transactions? Which of these kinds of international transactions are used to assess the equilibrium or disequilibrium of the balance of

payments?

9. What are the kinds of balance-of-payments disequilibrium? What are the factors that cause fundamental disequilibrium?

10. Distinguish between fundamental and cyclical disequilibrium of the balance of payments. Which of the two balance-of-payments disequilibrium is self-correcting?

11. What are the implications of a deficit balance-of-payments disequilibrium for the economy as a whole? Explain with examples.

12. What policy measures are adopted to correct the BOP disequilibrium of deficit kind? What factors determine the effectiveness of policy measures in correcting adverse balance of payments.

13. What is meant by devaluation of domestic currency? What are the conditions for effectiveness of devaluation in reversing the adverse balance of payments?

14. What is the purpose of exchange control? Under what conditions exchange control is applied to adjust BOP disequilibrium?

15. Distinguish between exchange depreciation and devaluation of a currency. Which of these methods is more affective in improving adverse balance of payments?

FURTHER READING

Carbaugh, Robert J., *International Economics*, (Australia: Thompson South-Western, 2008).

ChaCholiades, M, *International Economics,* (New York: McGraw-Hill, 1990), Chs. 14 and 15.

CleMent, M.o., r.l. Pfister, and K.J. rothwell, *Issues in International Economics,* (London: Constable and Company Ltd., 1967), Ch. 5.

dwivedi, d.n., *International Economics: Theory and Policy,* (New Delhi: Vikas Publishing House, 2013), ch. 16 and 17.

ellsworth, P.t., *International Economy,* (New York: Macmillan Company, 1969), Chs. 16 and 20.

heller, h.r., *International Trade: Theory and Emperical Evidence,* (New Jersey: Prentice Hall Inc., 1973).

KeMP, D.S. 'Balance of Payments Concepts—What Do They Really Mean? in *Readings in International Finance*. (Chicago: Federal Reserve Bank of Chicago, 1987).

Kindleberger, C.P., *International Economics,* (lllinois Richard D. Irwin Inc., 1973), Chs. 18, 19.

Kreinin MordeChat, e., *International Economics: A Policy Approach,* (New York: Harcourt Brace Jovanovich Inc. 1975), Chs. 2, 4 and 7.

leighton riChard J., *Economics of International Trade,* (New York: McGraw-Hill Book Company, 1970), Ch. 7.

Meade, J.e., *The Theory of International Economics Policy,* Vol. 1: *The Balance of Payments,* (London: Oxford University Press, 1972), Part 1.

salvatore, d., *International Economics,* 3rd Edn., (Macmillan Collier Macmillan, 1990), Chs. 15 and 16.

salvatore, sChiavo-CaMPo, *International Economics*, (Cambridge: Winthrop Publishing Inc., 1978), Ch. 9.

sodersten, bo., *International Economics,* (New York: Macmillan Company, 1970), Chs. 13 and 14.

CHAPTER 37 International Monetary System

CHAPTER OBJECTIVES

International transactions of goods and services, manpower and foreign investment generate international flow of foreign currencies. International monetary flow takes place under international monetary system. The present international monetary system has evolved over a long period of time. The objective of this chapter is to describe briefly the evolution of the international monetary system and its present status. The content of this chapter covers the following aspects of the international monetary system.

- Gold standard – the original system
- Breakdown of gold standard
- Growth of Currency Blocks
- Establishment of International Monetary Fund (IMF)
- The international monetary system of the IMF
- The breakdown of IMF system
- Creation of Special Drawing Rights (SDRs)
- The present international monetary system

37.1 INTRODUCTION

In the preceding chapters, we have discussed the theories of international trade, determination of exchange rate, and accounting of balance-of-payments. In this chapter, we discuss the system of international flow of foreign money, i.e., the international monetary system.

The ***international monetary system*** refers to *the organizational set-up that determines the rules and regulation, medium and mode, instruments and facilities for the settlement of international payments.* The international monetary system which prevails today has evolved over a period of more than 150 years. In the process of the evolution of the international monetary system, several monetary systems had come into existence which collapsed due to their inherent shortcomings and efforts were made to modify or to develop a new system. Historically, the international monetary system had started with the *gold standard*. It was followed by some other systems. The final form of international monetary system has been created by the establishment of *International Monetary Fund* (IMF). This

chapter presents a brief discussion of the evolution of the international monetary systems, their weaknesses, reforms and establishment of the IMF, its functions and problems.

37.2 THE GOLD STANDARD: THE EARLIEST INTERNATIONAL MONETARY SYSTEM

The earliest international monetary system was in the form of the *gold standard*. The gold standard mechanism provided a system which was nearest to an international monetary system. The system of gold standard was in existence over a period of four decades before the onset of World War I when most major countries were on gold standard. Their domestic currency system consisted of either gold coins or paper currency fully repayable in gold. Gold could be freely exported and imported in unlimited quantities between the countries on gold standard. This was one of the standard *rules of the 'gold game'*.

The exchange rates between currencies were fixed on the basis of their gold parity. The gold-standard countries maintained convertibility of their currencies into gold at fixed par values. Under this system, gold was the world's common unit of value, an international means of payment, a store of value and a means of international liquidity.

The gold standard system automatically restored the equilibrium through the flow of gold from countries having deficit in their balance of trade to those having surplus. The gold-losing countries experienced reduction in domestic money supply, money incomes and fall in prices. The gold-receiving countries, on the other hand, experienced increase in domestic money supply, rise in money incomes and rise in prices. These conditions made goods and services flow from the deficit countries to the surplus ones. This process continued until trade deficit of the deficit countries and trade surplus of the surplus countries were wiped out and equilibrium was restored.

The gold standard mechanism worked quite efficiently until the break of the First World War in 1914. The system worked because (*i*) the world economy was expanding rapidly; (*ii*) rising trend in world prices provided favourable conditions; (*iii*) there was quick adjustability of wages to price conditions and productivity; (*iv*) international capital movements created sufficient reserves which served as a stabilizing force, and (*v*) gold standard provided until 1914, a smooth payment mechanism.

37.3 BREAKDOWN OF GOLD STANDARD

The Reason for Break down of the Gold Standard

The gold standard, which worked efficiently till 1914, came under pressure with the break of the First World War I which rudely shattered the world economic order. In the face of inflation during the War, virtually all gold payments were suspended. Consequently, the convertibility of currencies in gold broke down. It was expected that freely fluctuating exchange rate would restore a competitive price and cost system which would automatically restore the stability in the world monetary system. But freely fluctuating exchange rates failed to restore the pre-War situation; instead, it stimulated speculation in hot currencies.[1] This led to overvaluation of pound sterling, undervaluation of French and Belgium francs and utter collapse of the German mark. As a result, other countries began to hold

1. Currencies used for short-term borrowing, e.g., pound sterling.

more and more reserves in pound sterling in addition to gold. Due to the War, free-trade and flexible exchange rate system gave way to trade restrictions and exchange controls. As a result, the monetary system of the world went into a complete disarray leading to collapse of the gold standard.

Restoration of the Gold Standard

While the War I was still going on, attempts were made by the major countries to restore the gold standard. Although the process of restoring the gold standard was initiated by the United Kingdom, the United States took lead in abolishing restrictions on gold exports and subsequently returned to the gold standard at pre-War gold parity. Germany and Austria 'scrapped their worthless currencies' and replaced them with new ones, though many other countries continued with their depreciated currencies. However, following the United Kingdom's return to the gold standard on 28 April 1925, the Commonwealth countries followed suit the same year. Between 1925 and 1928, more than forty countries returned to the gold standard. 'By the end of 1928, an international monetary system based on gold had been created,' and currency convertibility was restored.

Re-downfall of Gold Standard

The revived gold standard could not survive for long for the following reasons.

One, the pre-War conditions which had helped the functioning of the gold standard disappeared over time. Due to the after-effects of the War, the economic policies of the countries returning to the gold standard became inward-looking rather than being in tune with the international economic order.

Two, the post-War gold standard had lost its pre-War spontaneity. It was then a managed gold standard and was maintained at a considerable economic cost and strain. The major strain had fallen on Britain which was trying to sustain the gold standard by providing credit facility to the banks of many European countries to help them maintain their reserves and liquidity.

Three, unfavourable phenomenon was the misalignment of pound sterling and French franc, the former being overvalued. The pound was overvalued against the franc and dollar. This led to a gold rush from Britain to France and other European countries as there was free export and import of gold. Consequently, gold reserves in Britain declined sharply. Therefore, Britain suspended the convertibility of foreign balances into gold on 21 September 1931 and went off the gold standard. Following Britain, many Commonwealth countries gave up the gold standard. The International monetary system was once again disrupted.

37.4 GROWTH OF CURRENCY BLOCS

The main feature of the world monetary system after the second collapse of the gold standard was the emergence of several currency blocs on the monetary scene. Although currency blocs had started taking shape earlier on account of historical and political factors, they came into sharp focus during the 1930s through some sort of formal or informal arrangements between the members of currency blocs. The main currency blocs were (*i*) Sterling Area, (*ii*) French Bloc, (*iii*) Council for Mutual Economic Aid (COMEA or COMECON) and (*iv*) the Dollar Area formed round the US in the inter-war period.

The ***Sterling Area*** was a loose association of countries all over the globe—some of them being earlier a part of the British empire and some joined the bloc for its usefulness. Since Britain had held commercial power and pound sterling was dominating the financial scene, many British Empire countries (except Canada) found it beneficial to link their currencies closely to sterling, more so because part of their reserves were held in sterling and most of their trade was with Britain and British Dominions. Since the pound was more stable even during the days of chaos in the international monetary system, many non-Empire countries, e.g., Egypt, Argentina and Japan joined the sterling bloc.

The ***French Bloc*** or Franc Area consisted of former French African colonies. This area was relatively more closely knit. The Bank of France held the reserves of the member nations which had their currencies linked to the franc. Three other African countries—Morocco, Algeria and Tunisia—formed the outer franc area. Their currencies were, however, linked to gold, not to the franc.

Council of Mutual Economic Assistance (COMEA) also known as *COMECON* was the association of East European Countries with the then Soviet Union playing the major role. This association was said to have relatively more political content than economic. Trade between member nations was based on bilateral agreements.

Another import currency area, the ***Dollar Area*** was not a specific area. Canada, Japan and some Latin American countries belonged to *dollar area*. The US dollar emerged as the most powerful currency of the world because the USA had achieved the status of the biggest manufacturing power, the largest creditor and the second largest trader of the world. The US economy was more stable than any other country, which lent stability to the dollar. Thus, during the 1930s, three main currencies, viz., pound sterling, franc, and dollar, had appeared on the monetary scene, which served as reserve and link currencies for their respective areas. The multiple payment system added more confusion to the chaotic monetary system of the 1930s and made it more complicated. Further complications were added to the system by a widespread imposition of high tariffs, widely used quantitative controls, intensification of exchange controls, and competitive devaluation. All these factors led to the complete disintegration of the monetary system. This chaotic system continued until the establishment of the International Monetary Fund which marks a new phase of reconstruction of monetary order in the world economy.

37.5 ESTABLISHMENT OF INTERNATIONAL MONETARY FUND (IMF)

As mentioned above, from the 1930s, international monetary system was in complete disarray. Most currencies had become unstable and inconvertible. At least four currency blocs had appeared on the monetary scene. A system of multiple exchange rates had come into existence. Exchange rate fluctuations were frequent. Import restrictions and exchange controls were widespread. Competitive devaluations had become the order of the day. This anomalous situation had seriously injured the world economic powers like the USA, UK and France. These countries, therefore, initiated efforts in the direction of reconstructing the world monetary system afresh. Initially, efforts were made by the United S tates, Britain, France and Canada during the period of the World War II. The final shape of the new world monetary system emerged out of the American 'White Plan' and British 'Keynesian Plan'. The delegates of the USA and UK held a conference at Bretten Woods, Hampshire, in July 1944 and hammered

out the Articles of Agreement for International Monetary Fund and the International Bank for Reconstruction and Development (IBRD). The IMF came into existence in 1945 and started functioning in 1947. The number of members has since increased from 7 in 1947 to 187 in 2011.

Purpose of the IMF According to the Article I of the Bretten Woods Agreements, the main purposes of the IMF were the following ones.

1. To promote international monetary cooperation through a permanent institution which provides the machinery for consultation and collaboration on international monetary system;
2. To facilitate expansion and a balanced growth of international trade and to contribute thereby to the promotion and maintenance of high level of employment and real income and to the development of the productive resources of all members as primary objectives of economic policy;
3. To promote exchange stability, to maintain orderly exchange arrangement among members, and to avoid competitive exchange depreciation;
4. To assist member-nations in the establishment of a multilateral system of payments in respect of current transactions and in the elimination of foreign exchange restrictions which hamper the growth of world trade;
5. To give confidence to members by making the Fund's resources available to them to correct maladjustments in their balance of payments and to prevent them from resorting to measures destructive to national or international prosperity, and
6. In accordance with the above objectives, to shorten the duration and to reduce the degree of disequilibrium in the international balance of payments of members.

Sources of Fund's Resources

The resources of the IMF consist of contributions made by the member-countries in the form of capital subscription. The capital subscription of each member-country is based on the *quota* allocated to it. Quota for a member-country is fixed mainly on the basis of (*i*) size of its national income, (*ii*) gold and foreign exchange reserves and (*iii*) its share in the world trade. The members' quotas are enhanced periodically to boost IMF's resources. Another source of the loanable resources of the IMF are the loans from the member nations, but a major part of it comes only from the industrial nations and Saudi Arabia. The initial formula for fixing quota on the basis of these factors was described in the following words. "Two per cent of (member-country's) national income in 1940, plus five per cent of its gold and US dollar balances on 1 July 1943, plus 10 per cent of its maximum variation in annual exports in 1934–38, plus 10 per cent of its average annual imports in 1934–38; the total being increased in the same ratio as that which the country's average annual exports in 1934–38 bore to its national income". The actual quota allocated was approximated to 90 per cent of the sum calculated on the basis of the above formula. The remaining 10 per cent was taken into account in respect of countries whose economic strength was not fully reflected in the formula. Later, however, the above formula was used only as guide, and not as basis of calculations.

Financial Assistance by the IMF

In pursuance of its goals, the IMF provides financial assistance to the member nations facing BOP deficits. Member nations can borrow from the IMF to finance their short-run BOP deficits. Deficit nations borrow from the IMF by purchasing the currencies of other member nations (mainly dollars and some other major currencies) or SDRs with their own currencies.

The IMF financial assistance is available only for a short period. After the expiry of the period, member nations that purchase foreign currencies from the IMF are required to buy back their own currency. This is the mode of repayment.

Conditionalities of IMF Loans

The IMF loans are attached with some ***conditionalities***. The most important conditionality is that the deficit nations must agree to implement economic and financial policies suggested by the Fund. For example, when India approached the IMF for loans to overcome its unprecedented foreign exchange crisis of 1990–91, the IMF granted loans only after India agreed to implement sweeping economic reforms, including reforming her industrial policy, privatization of *PSUs*, reforming taxation policy drastically, cutting down the fiscal deficits, rationalization of foreign trade policy and cutting down the import duties.

The quota allotted to the member-countries signifies (*i*) the contribution of a member-nation to the Fund's resources, (*ii*) the relative position of a member-nation vis-à-vis other members, (*iii*) the relative voting power of member-countries, (*iv*) the borrowing rights and limits of the member-countries, i.e., 125 per cent of quota over five years at the rate of 25 per cent annum, (*v*) the gold and foreign exchange reserves of IMF—each member is required to deposit 25 per cent of its quota in gold and remaining 75 per cent in domestic currency and (*vi*) the borrowing limits under credit arrangements, like 'gold tranche' and 'credit tranche'. Besides, the size of quota allotted to a member is also a factor in determining the extent to which the members might be obliged to repurchase their currency from the Fund with gold, SDRs and convertible currencies, if Fund's holding of their currency exceeds 75 per cent of its quota.

Rise in Quota Allocations

As noted above, the number of member countries has increased from the initial number of 7 to 188 in 2012. As a result of new members joining the Fund's membership and also because of revision2 of quota allocations, the total quota has been increasing. The initial total quotas allocated to the various member-countries amounted to $ 8.8 billion in 1947 which has risen to $ 288.0 billion in 2012. In the initial allocations of quota, India (on par with Burma) had ranked fifth with her total quota, of 400 million dollars—the first four being the USA, the UK, China and France. Due to increase in membership and increase in quota of other nations, India now ranks eighth in order of quota allocation.

37.6 IMF AND THE WORLD MONETARY SYSTEM

In pursuance of its objectives, the IMF revived the gold exchange standard. The monetary system which emerged out of the Bretten Woods Agreements combined the characteristics of both the gold standard and the fluctuating exchange rate system of the 1930s. In the new

2. Articles of Agreement of Fund have provided that quota allocations must be reviewed and, if necessary, must be revised after every five years.

system, while gold remained the universally acceptable medium of payment, dollar and sterling emerged as reserve currencies supplementing gold as internationally acceptable assets. Exchange rates of currencies of the member-countries are fixed in gold and also linked to one of the reserve currencies—US dollar or pound sterling. With the fixation of exchanges rates, convertibility of currencies is restored. The IMF also restored the multilateral payments system with the following conditions.

(*i*) No country is supposed to impose new exchange controls without the approval of the IMF, except for stopping undesirable capital movements.

(*ii*) To ensure the flexibility of exchange rates within the prescribed upper and lower limits, IMF has laid the condition that members would not permit the spot exchange rates to vary more than one per cent beyond upper and lower limits.

(*iii*) To deal with short-run payment difficulties, a member would be allowed to borrow only 25 per cent of its *quota* in a year though the total borrowing could go up to 125 per cent over five years.

(*iv*) To overcome a *fundamental disequilibrium,* the IMF, if convinced, may permit devaluation of currency up to 10 per cent—devaluations of a greater degree being subject to negotiations.

With the establishment of the IMF, the multilateral payments system was fully restored within a period of five years. The member-countries gradually cast off their suspicion about the success of the IMF, abandoned their rigid exchange controls and moved toward a simpler system of multiple exchange rate. In 1954, the sterling area liberalized its earlier restrictive transfers towards the dollar area. By 1958, most major currencies had become convertible; restrictions on payments had disappeared; and a free payment system was almost universally established. Briefly speaking, the gold exchange standard was brought back into practice.

37.7 COLLAPSE OF THE IMF SYSTEM AND REFORMS

The international monetary system based on the Bretten Woods Agreements did not work smoothly for long. The problems threatening the working of the new system started brewing in the early 1960s. The problem started with the accumulation of key currency reserves by some non-key currency countries, which undermined the stability of the world monetary system. The key-currency accumulation required steady balance-of-payments deficits in the key currency countries of which the US was most important. With the accumulation of foreign exchange reserves beyond a certain limit, the liabilities of the key-currency countries became larger relative to their gold stocks available for conversion of the accumulated key-currencies. The scarcity of gold stocks was also on account of factors outside the system. While world trade was increasing at the rate of about 6 per cent per annum, the world supply of gold increased only at a rate of 1.6 per cent per annum. Moreover, smaller stocks of gold with the key-currency countries (in relation to the quantity required to maintain the convertibility of reserve currency), on the one hand, and fixed exchange rate system, on the other, created conditions for instability of the system. The fixity of the exchange rates under the conditions of gold scarcity encouraged speculative capital movements to the USA, which aggravated US balance of payments deficit. The pressure on dollar intensified during the late 1960s due to the increasing rate of inflation in the States. The mounting pressure on dollar forced the US government to devalue dollar by 8.57% in 1971. This marked the collapse of the IMF system.

The only reform in the international monetary system made by the IMF to cope with the situation was the introduction of Special Drawing Rights (SDRs) as a new source of international liquidity. The introduction of the SDRs, method of SDR, translation and its limitation are discussed below briefly.

37.7.1 Creation of Special Drawing Rights (SDRs)

The Special Drawing Rights (SDRs) are a kind of *reserve asset* created through the amendment of the IMF system to meet the growing need for international liquidity. It is nicknamed ‘Paper Gold’ since it performs some of the functions of gold in the international payment system. It is neither a paper currency, nor a coin, nor a credit note, nor a treasury bill. SDRs are simply the *entries* in the SDR-Account of the countries participating in the SDR-Scheme, though the SDRs figure in the published reserves of the participating nations.

The Background of SDRs Creation

During the 1960s, the world community experienced a serious inadequacy of *international liquidity*, i.e., shortage of gold and foreign exchange reserves to remove deficits in the balance of payments, because world trade was growing at a much faster rate than the world reserves. For instance, during 1950–71, world trade grew at an annual average rate of 8 per cent whereas world reserves increased at a rate of 2.1 per cent during 1951–60 and 3.2 per cent during 1960–71.[3] This had necessitated the search for a new kind of reserve asset to supplement gold stock and to create additional international liquidity. The efforts in this direction were initiated by the Group of Ten—the top ten members of the Fund. A series of meetings was held to negotiate on the *form* and *role* of the newly created reserve asset. In this regard, conflicting proposals were made. The major conflict over the issue was on whether the new form of international liquidity should be a 'reserve asset' or a 'credit facility' allocated among the members. Ultimately, the conflict was resolved by bringing a reconciliation between the features of both 'credit facility' and 'reserve asset' and the new reserve asset was named Special Drawing Rights (SDRs). The Articles of Agreement of the IMF were amended for the first time in 1969 to create additional liquidity in the form of SDRs.

Method of Transacting SDRs It is important to note that SDRs function as a means of payment, store of value and unit of account. The SDRs serve as a '*means of payment'* only between the participating nations for the purpose of legitimate purchase of foreign exchange, normally the currency of a third country and for making up the deficit in the balance of payments. It cannot be used as routine payment for imports by the government or by individuals.

The SDRs function as a *'store of value'* in the sense that so long as it is not used up, it adds to the international liquidity of a nation.

Since 1978, the SDRs have also been functioning as *'units of account'* after US dollar became unstable and floating became its permanent feature.

An important aspect of the SDRs transaction is that the participants are not required to transfer their currency or any other asset against the SDRs received in allocation. The SDR allocations are simply credited to the participant's SDR - Account and are then available for use under the rules.

3. IMF *Annual Report*, 1971, pp. 10 and 51.

Regarding the transfer of SDRs by one participant to another, there are three principle ways: (*i*) to receive foreign exchange from a participant designated by the Fund; (*ii*) through mutual agreements between the participants to redeem the balance of their own currency held by the other nation and (*iii*) for transaction with the Fund's General Account.4

Transaction with designated country In transactions involving designation, a country willing to buy foreign exchange against SDRs requests the Fund to designate a country from which foreign exchange can be purchased. The Fund then designates one or more participants who can transfer foreign exchange against SDRs to the user country. The Fund generally designates a country with a strong balance of payments in a position to provide convertible currencies.

Transaction by agreement The transfer of SDRs from one participant to another may also take place through a mutual agreement between them. This kind of transaction generally happens when a country wants to redeem its currency held by another participant country. Major part of SDRs transfers are of this nature.

Transaction with general account The SDRs may also be transacted with the General Account of the Fund. The Fund has the ability to acquire and use SDRs in operations and transactions conducted through the General Account. This feature of SDRs 'adds to its characteristics as a reserve asset'. Besides, participants can use SDRs "to pay charges to the General Account for the use of Fund resource"...and also "to repurchase their own currency from the General Account."

37.7.2 Limitation of SDRs

The Articles of Agreements lay down certain restrictions on the use of SDRs which are following:

1. A participant may transfer its SDRs to another participant in order to acquire an equivalent amount of convertible currency and not for other purposes. This facility is not available to the US.
2. A participant will be expected, in general, to use its SDRs only for two legitimate purposes: (*i*) to meet its balance of payment needs, and (*ii*) to cope with the changes in its total reserve position. If a country does not does not follow the rules, the Fund may advise other participants to transfer their SDRs to the misbehaving participant and may thus force the errant participant to surrender the foreign exchange improperly acquired.
3. A participant can normally use only 70 per cent of its average net cumulative

allocation during the first basic period of five years, i.e., the participants are expected to maintain a balance of 30 per cent in their respective SDR-Account.

4. As to the limit of receiving the SDRs, no participating country is obliged to hold more than 300 per cent of its own cumulative allocation of SDRs.

37.7.3 Further Reforms

In view of the pressing need for reforming the International Monetary System and growing demand from the less developed member-nations for adequate representation

4. For further details, See *'Finance and Development'* (IMF), June 1974, p. 6.

in the international monetary negotiations, a Committee on Reform of the International Monetary System of the IMF—known as the 'Committee of Twenty'5 was appointed.

The main issues on which the Committee deliberated were: (*a*) the appropriate degree of flexibility in exchange rates needed for BOP adjustments; (*b*) the need for greater symmetry in the system of adjustment between both deficit and surplus countries; (*c*) the return to convertibility into reserve assets for the dollar and pound sterling, and whether convertibility should be automatic or voluntary and (*d*) the nature of future international reserve assets.

The Committee agreed on (*i*) greater exchange rate flexibility than provided in the Bretten Woods Agreements; (*ii*) the SDRs Reserve as principal reserve asset, and gold and reserve currencies to have a reduced role to play; (*iii*) the valuation of SDRs being changed from a specific amount of gold to a basket of reserve currencies and (*iv*) the rate of interest being raised. But the Committee could not reach an agreement on certain other fundamental issues, such as the criteria for judging the need for BOP adjustments, the form of convertibility, future role of gold, and linking of SDRs with development assistance, etc.

While the negotiations were still going on, three **major changes** took place in the international economy. ***First,*** due to massive outflows of dollar in 1973, the dollar was devalued for a second time in relation to gold which amounted to a 10 per cent devaluation in respect of other major currencies. ***Second,*** because of the increase in oil prices in the late 1973 and early 1974, the world economy faced a very high rate of inflation. The uncertainty associated with these developments and the disagreements within the Committee on certain basic issues prevented it from reaching any final agreement.

Subsequently, however, an Interim Committee of the IMF went into the issues of reforming the international monetary system. On the basis of agreement reached by the Interim Committee in August 1975, the role of gold in the international monetary system was reduced, and official price of gold was to be abolished. The Interim Committee met in Jamaica in January 1976 and provided for legitimizing floating exchange rates and also for incorporating the earlier reform measures.

37.8 CURRENT ROLE OF THE IMF

The IMF is the most important international monetary institution. As noted above, the main functions of the IMF are (*i*) to create conditions for an international monetary system; (*ii*) to ensure the stability of the exchange rate; and (*iii*) to provide financial assistance to member-countries to maintain international liquidity. Before 2000, the IMF made several policy changes to raise funds and to provide financial help.

(*i*) The quotas of member-countries have been increased several times to increase the fund's resources. The total quota was increased from $266.9 billion in 2002 to $237.95 billion in 2011.

(*ii*) The borrowing facilities have been renewed and expanded several times under its lending schemes, viz., the General Agreements to Borrow (GAB) and a New Arrangement to Borrow (NAB).

(*iii*) Some other credit facilities introduced by the IMF include (*a*) the Extended Fund Facility (EFF), (*b*) the Supplement Reserve Facility (SRF), to provide short-run

5. The Committee of Twenty included delegates from Argentina, Australia, Belgium, Brazil, Canada, Ethiopia, France, Germany, India, Indonesia, Iraq, Italy, Japan, Morocco, the Netherlands, Sweden, the United Kingdom, Venezuela and Zaire. The delegates represented a group of countries in addition to their own.

financial facility during the Asian crisis, (*c*) the Contingent Credit Line (CCL), (*d*) Structural Adjustment Facility (SAF), (*e*) Poverty Reduction and Growth Facility (PRGF), (*f*) Enhanced Structural Adjustment Facilities (ESAF), and (*g*) Emergency Assistance.

As noted above, the IMF has not been very effective, ever since its inception, due to (*a*) changing economic conditions in the global economy, (*b*) persistent BOP deficits of the US, and (*c*) gold scarcity. Nevertheless, the fund has been helpful in resolving some serious problems of international financial liquidity. For instance, during the Latin American debt crisis of 1980s and the Asian financial crisis of 1990s, both of which had arisen on account of massive current account deficits, the IMF provided financial assistance to member-countries of these regions—with stringent conditions, of course—to overcome their financial crises and to revive their economies.

Over the past decade, the role of the IMF appears to have not only diminished but also it has proved ineffective for the following reasons:

(*i*) Frequent and violent fluctuations in the exchange rate mainly due to widespread misalignment of the exchange rates between currencies

(*ii*) IMF's failure to promote greater coordination of macroeconomic policies of the major industrial nations, and

(*iii*) Lack of ways and means to prevent international financial crises mainly due to the organization's lack of funds.

It may be added here that to make up its funds shortfall, the IMF planned to sell 400 tons of gold to boost its coffers as its budget fell short by $140 million in 2008. In the meantime the crisis aggravated and the US financial crisis cascaded to a global financial crisis in 2008. Although the US economy is said to reviving slowly, the European financial crisis continued to aggravate. The IMF did not prove to be very helpful. It is argued that IMF could not provide financial help to the countries facing financial crisis because it has spread widely and deeply into private financial institutions including commercial banks, investment banks, insurance companies, and private equity funds, which fall outside the purview of the IMF.

The effective functioning of the IMF is being affected seriously by some other kinds of international economic problems arising out of changing economic conditions in the member countries. Some of these problems are:

(*i*) growth of trade protectionism in industrial nations,

(*ii*) depression in the US economy and job insecurity,

(*iii*) slow growth and high unemployment in European countries,

(*iv*) economic stagnation in Japan, the second-largest economy of the world till 2011, and

(*v*) high rates of poverty in developing economies.

Despite all these limitations, the IMF continues to remain the most important international organization to maintain a systematic international monetary system and to provide financial help to countries facing financial deficits.

SUMMARY

• The earliest international monetary system was the gold standard, which existed between the 1870s and 1914. An important feature of the gold standard is that the balance-of-payments (BOP) adjustment mechanism works automatically.

• The inflationary pressures due to the First World War led to the breakdown of the gold standard. However, by the end of 1928, an important monetary system based on the gold standard had been created and currency convertibility was restored. However, the restored gold standard was in the form of the gold exchange standard. But the revived gold standard could not survive long. It collapsed in 1931.

• Following the breakdown of the gold standard, several currency blocks emerged – Sterling area, French bloc, Council for Mutual Economic Aid (COMECON) and dollar area.

• The international monetary system was in complete disarray during the period from 1931 to 1945—the period of the Great Depression and World War II. To address this situation, two international financial institutions were established, viz., the IMF and the International Bank for Reconstruction and Development (IBRD), called the World Bank, for providing long-run finance for development. The IMF started functioning in 1947.

• The main objectives of the IMF are (*i*) achieving a balanced expansion of world trade,

(*ii*) ensuring stability of the exchange rate, (*iii*) preventing members from indulging in competitive devaluation, and (*iv*) assisting members in correcting their balance-of-payments problems.

• The international monetary system created by the IMF worked fairly effectively in spite of some serious problems cropping up at times, until the system collapsed in 1971. The ultimate factor that led to the collapse of the Bretton Woods System was the growing BOP deficits in the US.

• After the breakdown of the Bretton Woods system, the international monetary system went into disarray, and most industrial nations were confronted with serious international economic problems related to their foreign trade and capital flows.

• Industrial nations made several attempts to restore the international monetary system. In March 1973, most industrial nations agreed to adopt a managed floating exchange rate system , with the dollar remaining inconvertible. When this exchange rate system did not work efficiently, the industrial nations decided, in 1979, to adopt a different managed floating exchange rate system.

• Later, the different groups of member countries adopted different kinds of managed floating exchange rate systems. Half the member countries followed the pegged exchange rate system with their currencies pegged either to the dollar or to the euro or to a basket of currencies. Most of the remaining member countries followed either managed joint or independent floating.

• In 1967, the IMF introduced Special Drawing Rights (SDRs) through amendments of the IMF system, made operational in 1969. The SDRs is a kind of reserve asset created by the IMF to meet the growing needs of international liquidity. It was nicknamed 'paper gold' as it performed some of the functions of gold in the international monetary system created by the IMF.

• Today, the IMF is the world's most important international monetary institution. Howver, its effective functioning is being affected seriously by some other kinds of international economic problems arising out of changing economic conditions in the member countries.

REVIEW QUESTIONS

1. What is meant by the international monetary system? What have been the different stages of the international monetary system?
2. What were the main features of the gold standard? How was the exchange rate determined under the gold standard? How was the balance-of-payments disequilibrium adjusted automatically under the gold standard?
3. What were the 'rules of the game' under the gold standard? Why did the gold standard collapse and how was it restored?
4. What were the factors that necessitated the formation of the Bretton Woods System? What were the objectives with which the Bretton Woods System was created? Was this system successful in achieving its objectives?
5. What were the basic functions of the International Monetary Fund? What were the sources of the financial resources of the IMF? How were quotas allocated to member-countries by the IMF?
6. What kinds of financial assistance are provided by the IMF to member-countries to help them eliminate their balance-of-payments deficits? What were the conditions for providing financial help?
7. What were the purpose and conditions of borrowing from the IMF under General Agreements to Borrow,
8. What were the factors that led to the collapse of the Bretton Woods System? What reforms were adopted by the leading industrial nations to revive the Bretton Woods system? Why did the revived system not work smoothly?
9. What developments had taken place after the collapse of the Bretton Woods System? What were the agreements made under the Smithsonian Agreement?
10. What is meant by gold exchange standard, flexible exchange rate system, and managed floating exchange rate? How was the exchange rate determined under these exchange rate systems?
11. Why was SDR created and how SDRs allocated between the member countries? How are SDRs transacted between the nations? How has the creation of SDRs contributed to the international monetary system?

12. What are the current lines of operation of the IMF? Has the IMF been successful in resolving the problems of the international monetary system?

FURTHER READING

Brian, Tew, *The Evolution of International Monetary System* (Hitchinsons and Co. (Publisher) Ltd., 1977).

Bloomfield, A. I. 1959. *Monetary Policy under the International Gold Standard:* 1880–1914 (New York: Federal Reserve Bank).

Clarke, w.m. and G., Pulay *, The Worlds Money: How It Works* (George Allen and Unwin, London, 1970).

Clive, william r., *International Monetary System and Developing Countries* (Brooking Institution, 1976), Ch.1.

de vries, M. G.. 1986. *The IMF in a Changing World*. Washington, DC: IMF.

de vries, M. G.. 'The International Monetary Fund and the International Monetary System' in M. Fratianni, D Salvatore, and J Von Hagen (eds). 1997. *Handbook of Macroeconomic Policy in Open Economy I* (West Port, C T: Greenwood Press).

dwivedi, D.N., *International Economics: Theory and Policy* (Vikas Publishing House, New Delhi, 2013), ch. 27.

ellsworTh P.T., *International Economy* (Macmillan Company, New York, 1964), Ch. 27.

GarBer, P. N., and svensson, L E O. 1995. 'The Operation and Collapse of the Fixed Exchange Rate Regimes.' *The Handbook of International Economics:* Vol. III. (Amsterdam: North-Holland).

kreinin, mordeChaT, e., *International Economics: A Policy Approach* (Harcourt Brace Jovanovich Inc., New York 1975), Ch. 3.

mCkinnon, R. I. 1996. *The Rule of the Game* (Cambridge, Massachusetts: MIT Press).

salvaTore, D. 2004. *International Economics* (New Delhi: Wiley India).

salvaTore, D. August 2000. 'The Present International Monetary System: Problems, Complications and Reforms.' *Open Economy Review*.

solomon, roBerT, *International Monetary System: An Insider's View—1945–1976* (Harper and Row Publishers, New York, 1977).

Triffin, R. 1961. *Gold and Dollar Crisis* (New Heaven: Yale University Press).

Teriffin, roBerT, *Our International Monetary System: Yesterday, Today and Tomorrow* (Random House, New York, 1968).

CHAPTER

38 Government's Role in the Economy

CHAPTER OBJECTIVES

The objective of this chapter is to discuss the reasons for the government intervention with the market economy, the government role in economic management and its implications for business management. The subject matter of this chapter contains the following aspects.

- Drawbacks of the free market system
- Reasons for the government intervention with the economy
- Economic functions of government and
- Government market intervention and private business

38.1 INTRODUCTION

In order to understand the rationale for government intervention with the market system, let us have a glance at the historical background of the emergence of government role in the economy.

Prior to Great Depression (1929–1934), most economies worked on the principles of *free market system*. During the period of free market system, the role of the government was limited to a few areas, viz., national defence, law and order, and the provision of some essential infrastructural facilities like roads and railways and some social services like health and education. The government was not supposed to play any economic role as, according to the protagonists of free market system, "Business is no business of the government". After mid-1920s, however, the free market system collapsed because of some drawbacks of the free market system. And the world economy had fallen into the Great Depression in 1929. There was unprecedented decline in production, employment, per capita income and price level in industrially advanced countries. After the Great Depression, it was J. M. Keynes, the greatest economist of the 20th century, who emphasized in 1936 the need for government role in reviving the economy by adopting policy measures to prevent the recurrence of economic debacle. It is how the importance of economic role of the government was recognized. After the World War II, most developed countries were economically shattered. The only option open to the economies was that the government must play important economic roles to revive the economy. This is how the government role in economic management came into existence. In modern times, governments in all countries play a significant economic role,

though the nature and level of the government role vary from country to country and from time to time depending on economic system of the country.

Even had the Great Depression not taken place, the need for economic role of the government and its intervention, control and regulation of the market economy would have arisen because of growing needs and aspirations of the people of all countries and the *limitations of the free market system* in meeting the growing needs of the people. Since economic role of the government has now become a reality of modern economic system mainly because of drawbacks of the market system, we begin our detailed discussion on economic role of the government by pointing out some major drawbacks of the free market system.

38.2 DRAWBACKS OF FREE MARKET SYSTEM

The drawbacks of the free market system refer to its deficiencies and failures in fulfilling the ambitions and aspirations of the people of a country. The prime ambition and aspiration of people of a country is *maximization of economic welfare*. The basic factors that maximize economic welfare are following.

• *High economic growth rate* – a high rate of growth in income and a high level of richness and prosperity to satisfy the ever growing needs of the people;

• *Full employment* – all those who are willing to work at the prevailing wage rate are employed;

• *Optimum utilization of resources* – national resources are so allocated between the industries that their marginal productivity is the same in all industries;

• *Equitable distribution of national income* – national income is distributed between the people according to their productivity; and

• *Optimal distribution of consumer items* – consumer goods and services produced at a point of time get so distributed between the people that marginal utility derived from a commodity by each consumer is the same.

It has been the world experience that these aspirations or objectives of the people are not fulfilled under the free market system because of certain serious drawbacks, defects, limitations and shortcomings of the free market system. Let us now look at the major drawbacks of the free market system.

1. ***Inoptimal Distribution of Goods and Services.*** According to Slither1, welfare maximization requires optimum allocation of goods and services. Optimum allocation of goods and services requires two necessary conditions: (*i*) goods must go to the consumers who will derive the greatest pleasure from them, and (*ii*) labour should so allocated in the production of goods that cost of production is minimum. As regards the optimum allocation of consumer goods and services, the optimality rule requires that consumer goods and services are so distributed between the consumers that marginal utility derived by each consumer from the consumption of each good is the same. But in a free enterprise economy, goods and services do not get optimally distributed among the consumers. For examples, clothes hanging in the wardrobe of rich person would definitely yield a higher utility or pleasure to the domestic servant.

2. ***Inoptimal Allocation of resources.*** According to Slither rule, welfare maximization requires optimum allocation of resources, i.e., labour is so allocated in production of goods

1. Sumner H. Slither, *Modern Economic Society* (New York: Henry Holt and Company, 1928), Ch. 8, reprinted in *Readings in Economics* (ed.) by P. A. Samuelson (New York: McGraw Hill Company, 6th Edn., 1970).

that marginal productivity of labour is maximised. But, under free market conditions, labour is so allocated that their marginal productivity is not maximized in all activities nor do the workers get wages according to their marginal productivity.

3. ***Perfect Competition Does Not Exist.*** The ideology of free enterprise economic system assumes the existence of perfect competition for its efficient working. The necessary conditions for the efficient working of the market system are: free competition, increasing cost in all markets, the exclusion principle of consumption, absence of public goods, perfect knowledge, and perfect factor mobility. But the existence of such a perfect market system in the world economy is a very rare phenomenon. Besides, the mere existence of perfect competition is not enough to ensure the efficient working of the system. Scitovsky[2] has pointed out that perfect competition would not ensure perfect efficiency if there are differences between social and private costs, but the existence of such differences cannot be denied.

4. *Individuals Are Not Necessarily the Best Judge.* The whole philosophy of free market system is based on the assumption that individuals are the best judge of their interest and, therefore, their choices and decisions are the best. But many choices and decisions made by individuals, particularly in regard to consumer goods, are generally influenced by 'impulses, habits, prejudices, ignorance or clever sales talk and too little by reflection and investigation of facts or evaluation and comparison of alternative opportunities.'[3] If this were not so, a household would not spend more on liquor and smoking and less on milk, education and health care, and a couple would not produce children whom they cannot bring up properly.

5. *Profit Is the Sole Motivating Force.* As mentioned above, the main motivating force of private enterprise is profit. The private entrepreneurs would, therefore, not like to invest their capital in industries which have a lower profitability even if such industries are of essential nature and of strategic importance for the national economy. Besides, in monopolistic and oligopolistic market settings—which are the order of the day—efforts to maximize profit result in underutilization of resources which means less than potential output and employment.

6. *Public Utilities Get Low Priority.* The services, known as 'public utilities' such as medical care, education, water, electricity, sanitation, etc., are equally important for all the individuals—rich and poor. Certain other facilities in the field of transport and communication (including roadways, railways, airways, telephones, posts and telegraph, etc.) are necessary for the overall growth of the economy. These facilities are jointly called 'socio-economic infrastructure.' The profit miximizing private firms do not invest in these sectors in adequate measure, for at least three reasons: (*i*) these sectors require huge initial investment; (*ii*) the returns in these sectors are relatively low, and (*iii*) most public utility services are in the nature of collective consumption to which the *principle of exclusion* in pricing cannot be applied. Even otherwise, 'public utilities' and other essential services cannot be left to the private sector pricing system because, in that case, only the rich would be able to afford these services, and hence there will be inequitable distribution of essential services.

7. *Growth of Monopolies.* The free enterprise system works through perfect and free competition. Perfect competition requires equality between the competitors. But any two firms are hardly ever equal in efficiency. The competition, therefore, generally becomes imperfect. It has been the universal experience that imperfectly competitive markets

2. Tibor Scitovsky, *Welfare and Competition*, (Unwin University Books, 1968), p. 444.

3. H. Slither, Sumner, *op. cit.*, pp. 35–36.

create conditions for the emergence of monopolistic and oligopolistic competition. This leads ultimately to the growth of private monopolies. Private monopolies result in low production, low employment and high prices, except in the case of natural monopolies. This is one of the greatest drawbacks of the free enterprise system.

8. *Externalities.* The free-market mechanism does not function efficiently where the ***exclusion principle*** is not applicable, i.e., those who do not pay for a good should be excluded from its consumption and those who 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, environment pollution by the factories, automobiles plying in the cities, use of loudspeakers in marriage ceremonies, etc., harm people by causing atmospheric 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, 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 economic growth, it does not ensure a stable, sustained and balanced growth. Therefore, it becomes necessary for the government to intervene in the market mechanism, reduce market distortions, provide conditions for fair competition and help the economy in achieving its goals—efficiency, stability, growth and economic justice.

38.3 GOVERNMENT'S ROLE IN DIFFERENT ECONOMIC SYSTEMS

We have noted that the interference of the government with the market mechanism is indispensable because of the failures of the free market system. Now, the question arises as to what should be the appropriate role of the government in economic management of

the country or what should be the form, nature and extent of the government's interference with the market mechanism? These questions have been debated for long without yielding a precise answer. The issue continues to remain controversial. The reason is that all is not well with government intervention with the market system. Those who tried to fix the limits of the economic role of the government in a free enterprise system have been ridiculed by others. For instance, Colin Clark proposed a law: "The role of government must be held at a ceiling of 25 per cent of the national income". Samuelson commented on the Clark's law, "This is not two-halves truth or even 25 per cent truth".[4] Samuelson's own 'final' law in this regard is, "There are no rules concerning the proper role of government that can be established by *a priori* reasoning".[5] In fact, the issue of appropriate economic role of the government is an ideological question and a matter of collective social choice.

4. Samuelson, P.A., *The Economic Role of Private Activity* in *Readings in Economics* (ed.) by Samuelson, *op. cit.*, p. 91.

5. *Ibid*.

However, the economic roles of the government can be broadly categorized on the basis of the three economic systems presently prevailing in the world, viz., capitalist or free enterprise economies, socialist economies and the mixed economies.

38.3.1 Capitalist Economy

A *capitalist economy* is one which is organized on the principle of the free enterprise system or what is also called a *laissez faire* system. Adam Smith, the founder of economics, had recommended four main functions for the government, viz., (*i*) national defence, (*ii*) maintaining law and order and internal security, (*iii*) judiciary and (*iv*) money supply. The economic activities, he suggested, should be left to the market system. With the passage of time and changing economic order, however, the economists have suggested a wider area for the government activities. In the capitalist system, the primary roles of the government are essentially (*i*) to restore and develop the free market mechanism wherever it is possible to ensure workable competitive conditions; (*ii*) to remove all unnecessary restrictions on the operation of free, competitive market and (*iii*) to provide a background through necessary government interventions and controls in which free competition can work effectively.[6] Besides, government intervention and its economic activities should deliver what the free market mechanism cannot. Meade[7] has recommended the government should perform the following eight kinds of functions.

(*i*) The government should control inflation and deflation mainly through indirect measures, like fiscal and monetary regulations;

(*ii*) The government should control and regulate monopolistic powers of large corporate undertakings with a view to avoiding inefficiency, unemployment and wastage of resources;

(*iii*) The government should have the ownership and state monopoly of essential goods and services, e.g., railway transport and generation and distribution of electricity and such other services on the ground of efficiency and economies of scale;

(*iv*) It should assume the responsibility of promoting equality of opportunity by providing equal access to educational opportunities and curtailing the restrictive trade practices and activities of trade unions, etc;

(v) It is the responsibility of the government to provide administration of justice and maintenance of law and order, and ensuring freedom of activities;

(*vi*) It should help entrepreneurs in planning against uncertainties of the future by some measures of government indicative planning;

(*vii*) The government should make central planning for large structural changes in the economy; and

(*viii*) It must tackle the problems of environmental controls, of the use of exhaustible resources, and of population growth.

It may be inferred from these propositions that the government's role in a capitalist society is supposed to be limited to (*a*) restoration and promotion of necessary conditions

6. Meade, J.E., *The Intelligent Radical's Guide to Economic Policy*, (George Allen & Unwin Ltd., London, 1795), pp. 13–1.

7. *Ibid*. , pp. 14–16.

for efficient working of the free market mechanism and (*b*) to enter those areas of production and distribution in which private entrepreneurship is lacking or is inefficient. Any planning by the government is indicative and should supplement the private plans for safeguard against future uncertainties.

38.3.2 Socialist Economy

In contrast to the capitalist system, the role of the government in a socialist economy is all pervasive. While in a capitalist economy, the government is supposed to play a corrective and complementary role in the economic sphere, in a socialist economy, it exercises comprehensive control on almost all economic activities. In the socialist system, not only is there a complete disregard for private ownership of property beyond the permissible limit, but also free enterprise and market mechanism are abolished by law. The private ownership of factors of production is replaced by state ownership. All economic activities are centrally planned, controlled and regulated by the state. All decisions regarding production, allocation of resources, employment, pricing, etc., are centralized in the hands of the government or its Central Planning Authority. The individual freedom of choice and decision-making in regard to economic activities is drastically curtailed. Individuals are provided freedom of choice, but within the policy framework of the socialist economy. Prior to the disintegration of the Soviet Union in 1989, the Soviet economy was the most prominent example of the socialist economic system. The other countries with socialist economic systems were China, Poland, Czechoslovakia and Yugoslavia.

The social aim of the socialist economic system is the same as in the free enterprise system, viz., efficiency, growth, social justice and maximization of social welfare. But, their methods of achieving these goals are totally different. The motivating force in a capitalist economy is private profit, whereas in a socialist economy, it is maximization of social welfare. Socialist way of management of the economy eliminates many evils of the capitalist system, e.g., exploitation of labour by capitalists, forces generating economic fluctuations, unemployment and social and economic inequality. The socialist economic system in its classical form is, however, disappearing from the economic scene mainly because, (*i*) it does not permit gains from private initiatives, (*ii*) it leads to under-utilization of resources, and (*iii*) it limits growth of the economy, as experienced by the erstwhile USSR and communist China.

38.3.3 Mixed Economy

A *mixed economy* is one which combines the features of both capitalist and socialist (centrally planned) economic systems. In this system, the major part of the economy, called *private sector,* is allowed to function on the principle of free enterprise system or free market mechanism within a broad political and economic policy framework of the government. The other part of the economy, the *public sector,* is constituted of industries and utilities promoted, owned and managed by the government, largely on the principles of a socialist economy. The public sector is created by reserving certain industries, trade, services and activities for government ownership, management and operation. The government prevents by law the entry of private capital into the industries reserved for the public sector. Another way of creating or expanding the public sector is the *nationalization* of private sector industries. The nationalization of private commercial banks and insurance companies in 1969 are prominent examples of the public sector extension in India. The

promotion, control and management of the public sector industries is the sole responsibility of the government.

Apart from controlling and managing the public sector industries, the government controls and regulates also the private sector through its industrial, monetary and fiscal policies. If necessary, direct controls are also imposed.

It is noteworthy here that the so-called free enterprise economies are, in fact, mixed economies, since in all such economies there co-exist private and public sectors. The mixed economies of free enterprise system can, however, be distinguished from the mixed economies of 'socialist pattern' on the basis of the rationale of public sector in the two systems. The public sector in a free-enterprise system is a matter of pure economic necessity and is complementary to the free market mechanism. It functions with the objectives of aiding, supplementing and strengthening the free enterprise system. On the other hand, creation of the public sector in a mixed economy like India, is a matter of ideological and social choice. Its creation and functioning is aimed at creating a 'socialist pattern of society'. It is another thing that India has failed to achieve any of these social goals.

In the *mixed economy*, the role and responsibilities of the government are much wider than in the free enterprise system, and much less than in the socialist society. The government in this system, undertakes to perform all the functions that the state performs in a free enterprise economy. In addition, it assumes the responsibility of making and implementing the plans for economic development of the country. The government also has to perform the task of coordinating private sector activities with the public sector, and controlling and regulating the former to bring it in tune with public sector policies. Since most countries today are of the nature of mixed economy, we discusses have the role of government in a mixed economy.

38.4 GOVERNMENT'S ROLE IN A MIXED ECONOMY

The government's role in a mixed economy with a large private sector may be broadly grouped under the following two categories.

(*i*) regulatory roles, and

(*ii*) promotional or development roles.

Regulatory roles include all direct and indirect policy measures which the government employs from time to time to control and regulate private business to prevent the growth of socially undesirable business activities, to prevent concentration of economic power and to direct private business activities towards the goals of growth and prosperity, employment and social justice.

The ***promotional roles***, on the other hand, include all the activities that are undertaken and all the policies that are adopted to build the development infrastructure (i.e., the economic and social overhead capital) necessary for industrial growth; to enhance the resource potential of both men and material, to enlarge the production capacity of economy and to create all other facilities deemed to be necessary for the overall growth of the economy. In a mixed economy like ours, the development activities are carried out by the government through a comprehensive programme of development.

In this section, we answer the question whether regulatory policies and development plans of the government help or hamper the private business. Since regulatory policies are a part of plan strategy, it is difficult to separate the impact of the two kinds of government roles on the private business. We have nevertheless discussed them separately.

38.4.1 Regulatory Measures and Private Business

The regulatory measures of the government include, by and large, the following policies:

(*i*) industrial and licensing policies,

(*ii*) taxation policy,

(*iii*) monetary and credit policy,

(*iv*) income and wage policy,

(*v*) price policy,

(*vi*) policy regarding technology and employment,

(*vii*) labour policy,

(*viii*) control and regulation of monopolies,

(*ix*) import and export policy,

(*x*) foreign exchange policy, and

(*xi*) industrial safety and environment policy.

The probable adverse impact of the regulatory policies of the government are pointed out here with reference to some major regulatory policies.

Adverse Impact on Private Business

From the angle of a free enterprise system, all the regulatory policy measures of the government are tantamount to interference with the market mechanism. Government regulatory policies restrain the market forces from their free play and hamper the automatic market functioning by introducing distortions in the price structure. The price structure becomes inoptimal which in turn prevents optimum allocation of resources. Cost-structure becomes inoptimal. To the extent such market distortions affect the private business decisions, business in private sector is adversely affected. For instance, the minimum wage policy of the government keeps wages artificially higher than the competitive market rate. Private entrepreneurs employing labour in accordance with the marginal productivity theory would employ less number of workers than they would in the absence of Minimum Wages Act. To the same extent their production and profit are reduced.

Similarly, when the government adopts a ***dear money policy*** and raises the bank rate, it discourages borrowings by the private entrepreneurs from the commercial banks and other financial institutions. If profitability is not relatively high, private investment will be restrained by the rise in the interest rate, other things remaining the same. Due to an increase in interest rate, businessmen may maintain lower levels of inventories, employ less number of workers, buy a lower quantity of raw materials, postpone the replacement of worn-out capital and defer new investment plans. Such private business decisions indicate adverse impact of the government's dear credit policy on the private business.

Consider some of the deterrent effects of ***business taxation policy*** of the government. If government taxes business income at a highly progressive tax rate, it reduces the after tax profit and, thereby, reduces the stimulus for investment and reduces the desire and capacity to save and invest. However, the impact of progressive business taxation on private business depends on the businessmen's capacity to shift the tax burden backward and forward and also on their ability to evade and avoid the tax. However, taxation affects private business through a complicated process. The adverse effects of taxation will be discussed elaborately in the next chapter.

Positive Social Impact. It may appear from the preceding discussion that the government's regulatory policy measures are always a deterrent to private business. This is, however, not a correct view of the impact of government control and other policy measures. Government policies should be examined in the overall social perspective and should be judged in the light of their overall social costs and benefits. The task of analyzing social cost and benefit of government policies falls outside the scope of this book. We will nevertheless mention some *a priori* justification for government control and regulation. Consider, for example, the fixation of minimum wage by law. It is the government's responsibility to protect and promote the interest of all sections of the society. Therefore, the government is justified in implementing such policy measures as are necessary to protect the interest of the majority of the people. The Minimum Wage Act is a similar measure—it protects labour from exploitation in a labour-surplus economy. What is equally important and often lost sight of is the fact that it helps private business in two ways. *One*, a higher wage rate keeps the aggregate purchasing power higher and, thereby, the aggregate demand. *Two*, it reduces the chances of avoidable confrontation between employers and employees. Similarly, 'tight money' and 'cheap credit' policies help in stabilizing the economy which is in the interest of both business community and other sections of the society. Furthermore, although business income taxation may appear to be a deterrent to growth of private business, the government expenditure counter balances its adverse effect to a great extent by enhancing the aggregate demand and purchases from the private sector. Likewise, the MRTP Act prevents the growth of monopolies and, thereby, provides safeguards to the smaller and less competitive firms—this maintains the competitive condition and promotes efficiency. Also, this Act prevents concentration of economic power which is one of the greatest economic evils of the free enterprise system. This enlarges the prospects for the expansion of private business activities.

It is, therefore, erroneous to have the view that government interference with the market mechanism always affects private business adversely. Government intervention helps private business in many ways, particularly under the conditions of widespread market imperfection. But, ill-conceived policy measures and even the most appropriate policies implemented badly serve neither the society nor the private entrepreneurs.

38.4.2 Government's Promotional Role and Private Business

The most important promotional or developmental role of the government is to create economic and social overhead capital which are the prerequisites of rapid economic growth. The creation of economic overhead capital includes development infrastructure including:

(*i*) creation of transport and communication facilities (e.g., railways, roadways, airports and sea ports, telephone and telex services, etc.);

(*ii*) creation of means of irrigation (e.g., dams, canals, tubewells, etc);

(*iii*) generation and distribution of electricity, and exploration and development of other sources of energy (e.g., coal, petroleum, natural gas, atomic energy and solar energy);

(*iv*) promotion and expansion of industries which are of strategic importance but lag behind for lack of private investment; and

(*v*) development and adoption of new and more efficient technology.

Creation of social overhead capital includes investment in education, health, housing, community development programmes, technical training etc., which enhance the productivity and growth potentials.

The infrastructure created by the government helps private business in at least two ways.

First, the creation of economic and social overheads helps to accelerate the pace of production and overall economic growth. A growing economy provides an environment in which entrepreneurs can look forward to expand their business with confidence and optimism. Economic growth enlarges the market size to the advantage of private business through a sustained increase in aggregate demand.

Second, an adequate supply of economic and social overhead capital creates external economies, reduces private cost of production and reduces capital-output ratio. Besides, economic and social overhead facilities created by the government help growth of private business by facilitating acquisition of inputs (labour and raw materials), by helping in the expansion of market for their product, by facilitating quick movements of goods and services, by making technical and skilled labour easily available, and so on. The lack of such facilities in the less developed countries restricts growth of the economy and also that of private business activities. Thus, by providing the necessary infrastructure, the government plays a significant role in promoting private business.

SUMMARY

- Government intervention with free market system and participation in economic activities started in a great measure after the World War II on the thoughts and principles suggested by J. M. Keynes.
- The rationale and need for government control and regulation of economic activities of the private sector had arisen due to drawbacks of the free market system in fulfilling the desire and aspirations of the people.
- The desire and aspirations of the people include (*i*) high economic growth rate, (*ii*) optimal utilization of resources, (*iii*) equitable distribution of national income, and (*iv*) optimal distribution of consumer goods and services.
- The drawbacks of free market system lie in its failure to ensure (*i*) a high rate of economic growth, (*ii*) optimal utilization of resources, (*iii*) equitable distribution of national income, and (*iv*) optimal distribution of consumer goods and services.
- Government role in controlling and regulating the economy and interference with the economy have become now a uniform factor in all kinds of economies – capitalist economies, socialist economies and mixed economies. But government role varies

depending on the nature of the economy. The government role is limited in capitalist system, it is almost total in socialist economies, and mixed role in mixed economies.
- Most economies in modern world today are largely mixed economies, though government role varies from country to country. The government role in mixed economies consists of two kinds of roles: (*i*) control and regulation of economic activities, and (*ii*) promotional role.
- Control and regulation of the economies are carried out by different kinds of policy

measures and promotional role is performed by direct investment in and participation in building socio-economic infrastructure and in industries of national defence, etc.

REVIEW QUESTIONS

1. Why is government interference indispensable in a free enterprise system? Does government interference ensure the fulfillment of social objectives?
2. What is 'mixed economy'? How does the government's role in a mixed economy differ from its role in a free enterprise economy?
3. Do the government's economic activities necessarily restrict the growth of private business activities?
4. How will you argue for minimizing the government's interference in the free market mechanism?
5. Explain how monetary and fiscal policies can help and hamper private business.
6. Do private enterprises have any social responsibility to fulfil? How can they bring a reconciliation between their profit maximization objective and their social responsibility?

FURTHER READING

Friedman, m., *Capitalism and Freedom* (University of Chicago Press, Chicago, 1962).

Jacoby, neil H., *Corporate Power and Social Responsibility: A Blueprint for the Future* (Macmillan, New York, 1973).

learner, a.P., *The Economics of Control* (Macmillan Company, New York, 1944).

meade, J.e., *The Intelligent Radicals Guide to Economic Policy* (George Allen and Unwin, London, 1975).

melvin ansHen (ed.), *Managing the Socially Responsible Corporation* (Macmillan, New York, 1970).

samuelson, P.a., *The Economic Role of Private Activity*, in his *Readings in Economics* (McGraw-Hill Co., New York, 1970), 6th Edn.

slitcHer, sumner, H., *Modern Economic Society* (Henry Hold and Company, Inc., New York, 1928), Ch. 8.

sPengler, J.J., *The Social Structure*, *The State*, *and Economic Growth—Brazil, India, Japan,* edited by Welbert E.M., and Spengler J.J., (Duke, 1995).

stiener, g.a., *Government Role in Economic Life* (McGraw-Hill Inc., New York, 1953).

CHAPTER

39 Monetary Policy

CHAPTER OBJECTIVES

The objective of this chapter is to discuss monetary policy, the most important policy instrument used by the government to control and regulate the monetary sector of the economy. The main aspects of monetary policy discussed here are following.

• Meaning and objectives of monetary policy
• The scope of monetary policy
• Monetary policy measures and
• Effects of the monetary policy measures on the economy

39.1 INTRODUCTION

In the preceding chapter, we have discussed the reasons and need for government control and regulation of the economy and economic role of the government. The government controls and regulates the economy by formulating and applying certain policy measures. There is a wide range of economic policies used by the government. However, given the purpose and scope of this book, we will confine our discussion to three major economic policies, viz., (*i*) monetary policy, (*ii*) fiscal policy, and (*iii*) industrial policy. In this chapter, we present a brief discussion of monetary policy. The main aspects of monetary policy discussed here are mentioned above. The various monetary policy measures discussed here include conceptual and theoretical aspects in the context of the Indian economy.

39.2 MEANING AND SCOPE OF MONETARY POLICY

Meaning of Monetary Policy Different monetary economists have defined monetary policy in different ways. For example, Harry Johnson defines monetary policy as a "policy employing central bank's control of the supply of money as an instrument of achieving the objectives of general economic policy"[1]. Shaw defines monetary policy as "any conscious action undertaken by the monetary authorities to change the quantity, availability or cost... of money"[2]. Note that Johnson emphasizes only the control of money supply

1. Johnson, Harry G., "Monetary Theory and Policy", *Am. Eco. Rev.* , Vol. LII, No. 3, June 1962, p. 335. Reprinted in his *Essays in Monetary Economics* (ed.), (George Allen and Unwin, London, 1969), p. 15.

2. Shaw, G.K., *An Introduction to the Theory of Macroeconomic Policy*, *op. cit.*, p. 65.

as the objective of monetary policy, Shaw emphasizes both supply of and demand for money. *Monetary policy is essentially a programme of action undertaken by the monetary authorities, generally the central bank, to control and regulate the demand for and supply of money with the public and the flow of credit with a view to achieving predetermined macroeconomic objectives*. The objectives of monetary policy, in general, are following.

(*i*) Economic growth,

(*ii*) Employment promotion,

(*iii*) Economic stability, and

(*iv*) Stability of the price level.

Priority of these objectives is determined on the basis of the need of the economy.

Scope of Monetary Policy The scope of monetary policy covers the area of economic transactions and the macroeconomic variables that monetary authorities can influence and alter through the monetary policy. In broad sense of the term, the scope of monetary policy encompasses the area of *demand for money* and *supply of money*. From monetary instruments point of view, the scope of monetary policy measures includes (*i*) bank rate, (*ii*) cash reserve ratio, (*iii*) open market operations by the central bank, and (*iv*) also the direct control measures deemed fit by the central bank.

In a fully *monetized economy*, the scope of monetary policy encompasses the entire gamut of economic activities. For, in such an economy, all economic transactions are carried out with money as a medium of exchange. In that case, monetary policy works by changing the general price level. It is, therefore, capable of affecting all economic activities—production, consumption, savings, investment and foreign trade. The monetary policy can influence all major macroeconomic variables— *GDP*, savings and investment, employment, the general price level and the foreign exchange.

The operational scope of the monetary policy depends also on the level of capital market development. While the change in the supply of money affects the level of economic activities through the price level, the other instruments of monetary control (bank rate and cash reserve ratio) work through the capital market. Where the capital market is fairly developed, monetary policy affects the level of economic activities through the changes in the capital market. It works faster and more effectively. Incidentally, a developed capital market is one which has the following features: (*i*) there is a large number of financially strong commercial banks, financial institutions, credit organizations and short-term bill market, (*ii*) a major part of financial transactions are routed through the capital markets, (*iii*) the working of the various capital sub-markets is inter-linked and inter-dependent, and (*iv*) the commodity sector is highly sensitive to the changes in the capital market. It is important to note that the changes in the bank rate and cash reserve ratio work through the commercial banks. Therefore, for the monetary policy to have a widespread impact on the economy, it is necessary that the capital sub-markets have strong financial links with the commercial banks.

39.3 INSTRUMENTS OF MONETARY POLICY

The instruments of monetary policy refer to the monetary measures that the central bank can use at its discretion with a view to controlling and regulating the money supply and the demand for money. The instruments are also called ‘weapons of monetary control’.

Samuelson and Nordhaus call these factors as “The Nuts and Bolts of Monetary Policy”. The measures of monetary policy are generally classified under two categories:

(*i*) General credit control measures, and

(*ii*) Selective credit controls.

39.3.1 General Credit Control Measures

The *general credit measures* or the *traditional measures* of monetary control are the measures that are used by the central bank to control the demand for and supply of money. The general credit control measures are following.

(*i*) Open Market Operations,

(*ii*) Bank rate, and

(*iii*) Cash Reserve Ratio (CRR).

In addition to these measures, a new measure called ‘Repo rate’ has been introduced to control the money supply and demand. Repo rate was introduced by RBI in October 2005. Let us now briefly discuss the meaning and working of these measures.

(i) Open Market Operations (OMO)

The 'open market operation' refers to the sale and purchases of government securities and treasury bills3 by the central bank of the country. When the central bank decides to increase the supply of money with the public, it purchases the government securities, i.e., bills and bonds, from the public and when it decides to reduce money in circulation, it sells the government bonds and securities. The open market operation is the most powerful and widely used tool of monetary control.

Let us look at the operational method of open market operations. The central bank carries out its open market operations through the commercial banks—it does not deal directly with the public. The buyers of the government bonds include commercial banks, financial corporations, big business corporations and individuals with high savings. These customers of government bonds hold their accounts with the banks. When they buy government bonds, money is transferred from their account to RBI account. Therefore, when the central bank carries out its open market operations, it affects bank deposits and reserves and thereby, their capacity to create credit. For instance, suppose the central bank decides to reduce money supply with the public and the availability of credit with the objective of preventing inflation. To this end, the central bank will offer the government bonds and treasury bills for sale through the commercial banks. When people buy the government bonds and securities generally through cheques drawn on the commercial banks in favour of the central bank, the money is transferred form the buyers' account to the central bank account. This reduces the total deposits with the commercial banks and also their cash reserves. As a result, credit creation capacity of

3. In India, Treasury Bills are short-term promissory notes issued by the Government of India through the Reserve Bank of India (RBI). There are two kinds of Treasury Bills—91-Day and 182-Day Bills. The 91-Day Treasury Bills are issued by the RBI on behalf of the Government at a fixed discount rate of 4.6 per cent. The RBI provides rediscounting facility within 14 days of issue at an 'additional rediscounting fees'.

The 182-Day Treasury Bills, introduced in 1986, are sold by way of auction to residents of India (excluding State Governments and Provident Funds), for a minimum value of `100,000. The auction bid is invited every fortnight and the 'discount rate' is decided on the basis of the auction rate. As regards the auction procedure, the auction bids invited are scrutinized by a Committee headed by the Deputy Governor of the RBI. The Committee decides on the cut-off price or the minimum official price. Bids quoting a price equal to or higher than the cut-off price are accepted and other bids are rejected.

commercial banks decreases. Consequently, the flow of credit from the commercial bank to the society decreases.

As regards the *demand for credit*, when the central bank sells government bonds, their prices go down and, therefore, the rate of interest goes up. This causes an upward push in the interest rate structure. The rise in the rate of interest reduces the demand for credit. Thus, not only the supply of credit but also the demand for credit is affected by the open market operations.

On the contrary, when the central bank decides to increase money supply, it buys back government bonds and securities. Then the money flows out from the central bank account to the people's accounts with the commercial banks. As a result, deposits with the commercial banks and their cash reserves increase. This enhances their capacity to create credit. Other things given, the flow of money from the banks to the public increases.

Effectiveness of Open Market Operations

The effectiveness of open market operation as a weapon of monetary control depends on the following factors.

(*i*) When commercial banks possess excess liquidity, the open market operation does not work effectively.

(*ii*) In a buoyant market situation, the effective control of demand for credit through the open market operation is doubtful. And, during the period of depression, open market operations are not effective for lack of demand for credit.

(*iii*) If banking system of the country is not fairly developed and security and capital markets are not interdependent, open market operations have a limited effectiveness. This is the case of underdeveloped countries.

(*iv*) The popularity of government bonds and securities with the public also matters a lot. The government debt instruments are generally not popular due to low rate of return. The central bank then has to use coercive measures and force the commercial banks to buy the government bonds, as is the case in India.

(ii) Bank Rate Policy

Bank rate, known also as *discount Rate*, is the rate at which the central bank rediscounts the bills of exchange presented by the commercial banks. The RBI Act 1935 defines 'bank rate' as the *"standard rate at which (the bank) is prepared to buy or rediscount bills of exchange or other commercial papers eligible for purchase under this Act"* .

The RBI rediscounts only approved bills and the 'first class bills of exchange'. Why do commercial banks get their bills of exchange rediscounted? What happens, in fact, when commercial banks are faced with a shortage of cash reserves, they approach the central bank to get their bills of exchange rediscounted. It is a common method of borrowing by the commercial banks from the central bank. The central bank rediscounts the bills presented by the commercial banks because it is a part of its functions—it is the *lender of the last resort*. For rediscounting the bills of exchange, the central bank charges a rate. This rate is traditionally called *bank rate*. A more appropriate name in usage now is the ***discount rate***. However, *for all practical purposes, bank rate is the rate which the central bank charges on the loans and advances to the commercial banks.*

The central bank can change this rate—increase or decrease—depending on whether it wants to expand or reduce the flow of credit from the commercial bank. When it wants to increase the credit creation capacity of the commercial banks, it reduces the discount rate and *vice versa*. This action by the central bank is called the bank rate policy or more appropriately, the *discount rate* policy. The *bank rate policy* was first adopted by the Bank of England in 1839. The RBI has been using this measure since its inception in 1935.

The working of the discount rate policy is simple. When the central bank changes its discount rate, commercial banks change their own discount rate. That is, when the central bank raises its discount rate, commercial banks raise their discount rates too. Generally, the central bank rate is 1 percentage point higher than the discount rate charged by the commercial banks.

Let us now look at how change in bank rate changes the flow of credit.

One, when RBI raises the discount rate (virtually the interest rate), it reduces the net worth of the government bonds (the treasury bills and promissory notes) against which commercial banks borrow funds from the central bank. This reduces the commercial banks' capacity to borrow from the central bank. As a result, commercial banks find it difficult to maintain a high cash reserve. This reduces the credit creation capacity of the commercial banks. So the flow of credit to the market is reduced.

Two, when the central bank raises its discount rate, commercial banks raise their discount rate. This raises the *cost of credit* which discourages the business sector to get their bills of exchange discounted. Besides, a rise in the bank rate pushes the interest rate structure up. The rise in the interest rate reduces the demand for funds too. Such a policy is, therefore, called a '***dear money policy***'. A reverse process is used when central bank adopts a ***cheap money policy***.

Three, In general, bankers' lending rate is quickly adjusted to deposit rates. Therefore, a rise in the bank rate causes a rise in the deposit rate. This turns borrowers into depositors. Therefore, savings flow to the banks in the form of deposits. This is called *the deposit mobilization effect.*

Limitations of Discount Rate Policy

The discount rate policy has lost its effectiveness as a weapon of monetary control over time for the following reasons.

1. The variation in the discount rate works effectively only when commercial banks approach the central bank for borrowing. In modern times, the commercial banks have built their financial resources. They are not dependent on the central bank for financial support. Therefore, their discount rate is not affected significantly when central bank raises the bank rate.

2. With the growth of credit institutions and financial intermediaries, the capital market has widened extensively. The share of banking credit has declined. Therefore, variations made by the central bank in the discount rate, especially when it raises the rate, have only a limited impact on the credit market—it is limited to only bank credit.

3. From the credit demand angle, variations in the discount rate become effective only where demand for credit is interest-elastic. The structure of the credit market in the less developed countries is such that the interest rates are sticky. Hence change in the discount rate has not been found to be very effective.

(iii) The Cash Reserve Ratio or Statutory Reserve Ratio

The cash reserve ratio (*CRR*) is the percentage of total deposits which commercial banks are required by the central bank to maintain in the form of cash reserve with the central bank4. The objective of the cash reserve ratio is to prevent shortage of cash in meeting the demand for cash by the depositors. The cash reserve ratio depends, normally, on the banks' experience regarding the cash demand by the depositors. But, "If there were no government rules, banks would probably keep only a very small fraction of their deposits in the form of reserves"5. Since cash reserve is *non-interest bearing*, commercial banks often keep their cash reserves below the safe limits. This situation might lead to financial crisis in the banking sector. So, in order to prevent this eventuality, the central bank imposes a *CRR* on the banks. This has become a handy tool for the central bank to control money supply. The central bank enjoys the legal powers to change the cash reserve ratio of the banks at its own discretion. The cash reserve ratio is a *legal requirement*. Therefore, it is also called *statutory reserve ratio (SRR)*.

By changing the *CRR*, the central bank can change the money supply overnight. When economic conditions demand a contractionary monetary policy, the central bank raises the *CRR*, and when economic conditions demand monetary expansion, the central bank cuts down the *CRR*. The effect of change in the *CRR* on the supply of money and credit can now be briefly explained. Suppose commercial banks possess a total deposit of `100 million and *CRR* is 20 per cent. It means ( *a*) that the banks can loan `80 million and (*b*) that the credit or deposit multiplier6 equals five. It also means that the banks can create, through the process of credit multiplier, a total credit of `500 million or an additional credit of `80 million × 5 = `400 million.

Now let the central bank decide to reduce the money supply with the public and raise the *CRR* to 25 per cent. Then the credit multiplier will go down to 4. With this provision, the banks can provide loan only to the extent of `75 million (= `100 million – `25 million). Thus, the total credit created by the banks goes down tò100 × 4 = `400 million and additional credit goes down tò75 × 4 = `300 million. A fall of `100 million in the bank credit is supposed to have considerable impact on the money market. The effect will be reversed when the central bank cuts back the *CRR* to 20 per cent.

Effectiveness of CRR as an Instrument of Monetary Control This method alone is effective where other measures fail. It proves more handy where open market operation and bank rate policy prove less effective. However, its effectiveness in terms of impact on the capital market depends on the share of the banking credit in the credit market.

4. In India, the scheduled commercial banks were required until 1956 to maintain 5% of the demand liabilities and 2% of the time liabilities in the form of cash reserves. The RBI Amendment Act, 1956 empowered the RBI to vary minimum cash-deposit ratio between 5% and 20% for demand deposits and between 2% and 5% for time deposits. In 1962, however, this distinction between the demand and time deposits was removed and a flat rate of 3% was fixed for all deposits with the provision that this could be raised to 15%.

5. Samuelson, P.A. and W.D. Nordhaus, *Economics*, 1995, *op. cit*, p. 511.

6. The formula for deposit multiplier (*Dm*) is given below:

1

1

1

D =

=

=

=

m

5

CRR 20 100 0.20

The total credit creation (*TCC*) can be worked out as follows:

1

1

TCC =

(Deposite) =

CRR

20 100 (100 million) = 500 million

It is relatively more effective in the advanced countries with advanced banking system accounting for a major share in the capital market.

The Statutory Liquidity Requirement (SLR):

A Special Provision in India

In India, the RBI has imposed another kind of reserve requirement in addition to *CRR*, called *Statutory Liquidity Requirement* (SLR). The SLR is the proportion of the total deposits which commercial banks are required to maintain with them in the form of liquid assets (cash reserve, gold and government bonds) in addition to cash reserve ratio. This measure was undertaken to prevent the commercial banks from liquidating their liquid assets (government security holdings) when *CRR* is raised. Prior to the imposition of SLR, commercial banks used to convert their liquid assets into cash to replenish the fall in their loanable funds due to a rise in the *CRR*. The RBI has been almost regularly raising the SLR since the early 1970s. It had raised the SLR from 25% to 30 % in November 1972, to 32% in 1973, to 33% in 1974, to 34% in 1978 and then to 38.5%. Since April 3, 1992, however, the RBI reduced the SLR on the additional demand and time deposits. It cut down the SLR on additional deposits to 30% on April 3,1992. At present the scheduled commercial banks are required to maintain SLR at 25% whereas the effective rate worked out to 28% of the total Net Demand and Time Liabilities (*NDTL*) in March 1996. The RBI keeps changing the SLR in order to achieve objectives of its monetary policy.

39.3.2 Selective Credit Controls

The *quantitative methods* of monetary control, when they are effective, affect the entire credit market in the same direction. In other words, quantitative monetary controls hava their impact on all the sectors of the economy in a uniform manner. This may not be always desirable or intended by the policy-makers. The monetary authorities are often faced with the problems of (*a*) rationing the credit, (*b*) diverting the flow of credit from the non-priority sectors to the priority sectors, and (*c*) curbing speculating tendency based on the availability of bank credit. These objectives of credit control are not well served by the quantitative measures of credit control. The monetary authorities then resort to *selective credit controls*. Selective credit controls include the following credit control measures.

(i) Credit rationing. When there is shortage of institutional credit available for the business sector, the large and financially strong sectors or industries tend to capture the lion’s share in the total institutional credit. As a result, the priority sectors and weaker but essential industries are starved of necessary funds, mainly because bank credit goes to the non-priority sectors. In order to curb this tendency, the central bank resorts to credit rationing measures. Generally, three measures are adopted: (*a*) imposition of upper limits on the credit available to large industries and firms, (*b*) charging a higher or progressive interest rate on bank loans beyond a certain limit, with a view to making bank credit available to relatively weaker sectors, and (*c*) providing credit to weaker sectors at lower internal rates.

(ii) Change in lending margins. In general, banks advance money against a mortgage of property – land, building, jewellery, shares, stock of goods, etc. The banks provide loans only up to a certain percentage of the value of the mortgaged

property. The gap between the value of the mortgaged property and amount advanced is called ‘lending margin’. For example, if the value of stock is `10 million and the amount advanced is only `6 million, the lending margin is 40 per cent. The central bank is empowered to increase or decrease the lending margin with a view to decreasing and increasing the bank credit. This method was used for the first time by the RBI in 1949 with the objective of controlling speculative activity in the stock market. Since 1956, the RBI has made an extensive use of this method with a view to preventing speculation in scarce agricultural products, namely, foodgrains, cotton, oilseeds, vegetable oils, sugar, *khandsari* and *gur*, and cotton textiles and yarns. However, it is important to note here that selective credit controls were eliminated in October 1996 (*Economic Survey*—1997–98, p. 43.)

(*iii*) *Moral suasion.* The moral suasion is a method of persuading and convincing the commercial banks to advance credit in accordance with the directive of the central bank in the economic interest of the country. This method is adopted in addition to quantitative and other selective methods, particularly when effectiveness of these methods is doubtful. Under this method, the central bank writes letters to and holds meetings with banks on money and credit matters with the objective of persuading banks to act according to the instructions and advise of the central bank in the interest of the economy as a whole.

(*iv*) *Direct controls.* Where all other methods prove ineffective, the monetary authorities resort to direct central measures with clear directive to the banks to carry out their lending activity in a specified manner. There are (however) rare instances of direct control measures.

39.3.3 Repo and Reverse Repo Rates

In addition to traditional monetary control measures, RBI uses repo rate (repurchase operation rate) and reverse repo rate under its Liquidity Adjustment Facility (LAF) programme. ***Repo rate*** is the rate at which RBI lends short-term money to the commercial banks when they borrow from the RBI. ***Reverse repo rate*** is the rate that RBI offers to the banks willing to keep their money with it. Depending on the need of the country, the RBI keeps changing these rates. On 28 January 2014, the RBI had fixed repo rate at 8 per cent.

39.4 LIMITATIONS OF MONETARY POLICY

The effectiveness of monetary policy, or any policy for that matter, depends on a number of factors.

1. *The time-lag*[7]. The first and the most important factor that delays the effective working of monetary policy is the *time-lag*. Time lag refers to the time taken by policy-makers for chalking out the policy action and its implementation. The time-lag is divided in two parts: (*i*) 'inside lag' or preparatory lag, and (*ii*) 'outside lag' or response lag. The 'inside lag' refers to the time lost in (*a*) identifying the nature of the problem, (*b*) identifying the sources of the problem, (*c*) assessing the magnitude of the problem, (*d*) choice of

7. For a comprehensive analysis of time lag in monetary policy, see Hamberger, Michael J., "The Lag in the Effect of Monetary Policy: A Survey of Recent Literature", *Federal Reserve Bank of New York Monthly Review*, December 1971.

appropriate policy action, and (*e*) implementation of policy actions. The 'outside lag' or the response lag refers to the time taken by the households and the firms to react in response to the policy action taken by the monetary authorities. If inside and outside lags are long, not only the nature and the magnitude of the problem may change rendering the policy ineffective, but also it may worsen the situation.

It has been the experience of many countries including developed ones that 'time-lag' has been unduly long making monetary policy less effective than expected. The time-lag of monetary policy, particularly its response lag, has been found to be generally longer than the time lag of fiscal policy. However, the issue of time-lag of monetary policy is controversial. Friedman and Schwartz find an average time-lag of 18 months between peaks (or troughs) of money supply and peaks (or troughs) of business cycle. Their findings have, however, been questioned by the findings of other economists[8]. Nevertheless, 'the evidence from several sources suggests that the lag associated with monetary policy is long and possibly variable'[9] and 'the consensus seems to be that the lag is about 12 to 16 months long'[10].

2. *Problem in forecasting.* The formulation of an appropriate monetary policy requires a reliable assessment of the nature and magnitude of the problem—recession or inflation. More important is to forecast the effects of monetary actions. Despite advancement in forecasting techniques, reliable forecasting of macroeconomic variables remains an enigma. Look at the views of Stephen McNees[11] on problems in forecasting.

"How can forecasters go wrong? They may not predict disturbances (the Gulf War, for example); they may misread the current state of the economy and hence base their forecasts on a wrong picture of the present situation; and they may misjudge the timings and the vigour of the government's monetary and fiscal responses to booms or recessions. The fact is that forecasting has not reached

perfection, particularly at major turning points in the economy. ...”

With this status of forecasting and prediction of the outcome of a policy action, formulation of an appropriate monetary policy has remained an extremely difficult task. An inappropriate policy based on guesswork is bound to be unsatisfactorily effective. The empirical evidence proves the point.

3. *Non-banking financial intermediaries.* The structural change in the financial market has also reduced the scope of effectiveness of monetary policy. Non-banking financial institutions have expanded rapidly over time. The proliferation of non-banking financial intermediaries including industrial finance corporations, industrial development banks, mutual saving funds, insurance companies, chits and funds, etc., has reduced the share of the commercial banks in the total credit. Although financial intermediaries cannot create credit through the process of credit multiplier, their huge share in the financial operations reduces the effectiveness of monetary policy.

8. For details, see Edgmand, Michael R., *Macroeconomics: Theory and Policy* (Prentice-Hall of India, 2nd edn., 1985), Ch. 18.

9. Edgmand, Michael R. *op. cit.*, p. 373.

10. Glahe, Fred R., *Macroeconomics: Theory and Policy* (Harcourt Brace Jovanovich, Inc., New York, 1973), p. 287.

11. McNees, Stephen, “How Large Are Economic Forecast Errors?” *New England Economic Review*, July-August 1992, part reproduced in Rudigar Dornbusch and Stanley Fischer, *Macroeconomics*, 6th Edn., 1994, (McGraw-Hill, Inc.), p. 456.

4. *Underdevelopment of money and capital markets.* The effectiveness of monetary policy in less developed countries is considerably low because of the underdeveloped character of their money and capital markets. Their money and capital markets are fragmented while effective working of monetary policy requires that money market and the sub-markets of the capital market work interdependently. For this reason, the effects of change in money supply and particularly the changes in the interest rate remain confined to the banking sector.

39.5 MONETARY POLICY OF INDIA

Having discussed theoretical and some empirical aspects of monetary policy, we proceed now to discuss briefly India’s monetary policy. The main aspects of India’s monetary policy discussed here are (*i*) objective of India’s monetary policy, (*ii*) policy measures adopted to achieve the objective of monetary policy, and (*iii*) effectiveness of policy measures.

39.5.1 Objective of India’s Monetary Policy: Price Stabilization

In general, all macroeconomic policies of a country have four main objectives: (*i*) economic growth, (*ii*) employment promotion, (*iii*) price stability, and (*iv*) social justice – equitable distribution of income. However, the two main objectives of monetary policy of most countries are (*i*) economic growth, and (*ii*) controlling inflation and ensuring price stability. Of these objectives, *growth and price stability* have been, in general, the main objectives of India’s monetary policy. However, Chakravarty Committee[12] emphasized that *price stability* is “the dominant objective of monetary policy”. According to Rangarajan, a former Governor of the RBI, “It is price stability which provides the appropriate environment under which growth can occur and social justice can be ensured”[13]. The price stability as the dominant objective of India’s monetary policy was recognised because growth process in the country, although slow, was building inflationary pressure and the country needed a monetary policy to maintain a reasonable balance between price stability and economic growth. Let us have look at the inflationary trend in India and the need for price stabilization.

Inflationary Trend in India. Price stabilization has been the main objective of India’s monetary policy because the country has been under inflationary pressure since the Second Plan Period. The inflation rate was around 6 per cent on y-t-y basis, though it had shot up to 12 per cent during the period from 1971 to 1976. The main cause of inflationary pressure was deficit financing—a fiscal measure—adopted by the Government of India (GOI) to finance the development programmes. As a result, money supply was increasing at a faster rate (16–17 per cent p.a) against low economic growth (3-4 per cent p.a) causing inflation 5–7 per cent. The Chakravarty Committee emphasized the importance of controlling growth of money supply, and suggested that the growth of money supply should be regulated with the objective of maintaining price stability in compliance with the Plan objective of output growth. In order to control inflation to an acceptable level, the RBI adopted measures to control monetary expansion, at the same time making bank credit available for private investment at a reasonable interest rate.

12. The Working Committee to Review the Working of the Monetary System (1985) set up by the Reserve Bank of India in 1982 under the Chairmanship of Prof. Sukhmay Chakravarty with the membership of C. Rangarajan, the then Governor of the RBI.

13. C. Rangarajan, "Issues in Monetary Management", and "Monetary Policy Revisited", in his *Indian Economy: essays on Money and Finance* (New Delhi: UBS Publishers' Distributors Ltd., 1998).

Let us look at the inflationary pressure that India has faced from time to time. India had inflation rate of about 6 per cent during the period from 1956–57 to 1970–71. The average rate of inflation during the period 1971–76 had shot up to 12 per cent—the highest rate of inflation India had witnessed till then. The inflation rate, however, declined thereafter and averaged 7.5 per cent till 1990–91. But the quinquennial average inflation rate shot up again to 10.6 per cent. This was the second phase of high inflation in India. Since 1996–97, however, the rate of inflation declined and remained stable around 5 per cent till 2007. This rate of inflation was close to the Chakravarty Committee recommendation. But, on 5th July 2008, the inflation rate shot up again to about 12 per cent (11.91% to be precise) which is considered to be the highest in the past 13 years. The inflation record shows the inflation rate could hardly be ever maintained to the level of the rate recommended by the Chakraverty Committee. It is understandable, as C. Rangarajan (the then member of the Chakraverty Committee and the Governor of RBI), had remarked, "absolute price stability is not feasible in a large and complex economy undergoing structural transformation. Imbalances to some extent are inevitable."14

The question that arises here is: What rate of inflation would comply with Plan objective of growth rate? The Chakravarty Committee (1985) had examined the issue at both theoretical and empirically levels and had recommended an annual inflation rate of 4 per cent which was, in its opinion, socially tolerable and conducive to growth. However, since India has had only "moderate inflation"—crossing double digit rarely—'inflation targeting' was not considered to be appropriate policy measure for India. In 2013 and 2014, however, controlling inflation has become the objective of India's monetary policy because inflation rate had gone up to 8–10 per cent.

39.5.2 Monetary Measures

The RBI has been using various monetary measures from time to time including some non-traditional measures for price stabilization and other monetary policy objectives. We give here a brief description of the measures adopted by the RBI, and also their effectiveness.

1. *Bank Rate* The bank rate has been one of the important instruments used by the RBI to control inflation, whenever required. As mentioned above, the bank rate remained unchanged at 3 per cent during 1935–1950. Since 1951, however, bank rate has been frequently changed – mostly increased. The RBI was using bank rate infrequently as a weapon of monetary control till mid-1990s with the purpose of mitigating mounting inflationary pressure in the country. After mid-1990s, however, inflation rate declined with rise in the growth rate of the economy, due mainly to economic reforms. As a result, the RBI started reducing bank rate from the year 1997 which continued till May 2008. However, the RBI started enhancing the bank rate and raised it to 7.5 per cent in July 2008 due to rate of inflation crossing double digit. However, due to fall in the inflation rate in late 2008, the bank rate was cut down to 6 per cent in January 2009. This rate is likely to be maintained in fiscal year 2008–09.

As regards the effectiveness of bank rate as an instrument of monetary control, India's experience, and also that of other countries, shows that the bank rate has not proved to be an effective method of controlling money supply. The reason is that banks do not depend on the RBI greatly for their financial requirements. Besides, even if commercial banks

14. Rangarajan, C., "The Analytical Framework of the Chakraverty Committee Report on the Monetary System", *Reserve Bank of India Bulletin*, September 1987, p. 702.

borrow from the RBI, their total borrowing accounts for a small proportion of the total credit created by the commercial banks, especially when there are other sources of credit.

2. Cash Reserve Ratio (CRR) The *CRR* is another traditional monetary tool that RBI has been using to control inflation in the country, and also to restrain credit flow to the business sector. Recall that *CRR* refers to the percentage of net demand and time liabilities (NDTL) which commercial banks are required to maintain in the form of ' *cash reserves*'. The NDTL are essentially the net demand and time deposits. The *cash reserves* are practically divided under two heads: (*i*) 'required reserves (*RR*)', and (*ii*) 'excess reserve'. The ***required reserve*** is the cash reserve that commercial banks are statutorily required to maintain with the RBI. Incidentally, this is **a non- traditional method**. The **'required reserve'** is calculated fortnightly (on the second Friday of the month) on the basis of average daily deposits. The *excess reserve* is the cash reserve which banks maintain as 'cash in hand' with the purpose of meeting the currency demand by the depositors. The excess reserves are determined generally by the bank's own experience regarding the 'currency drain'.

As regards the use of the *CRR* method as monetary control, till 1973, the RBI used this method only once in 1960. However, since 1973, the RBI has been using *CRR* quite frequently as a major instrument of controlling the excess supply of money. The RBI raised the statutory *CRR* from 3 per cent fixed in 1935 to 5 per cent in 1960 and raised it further frequently. As a result, the bank rate had gone up to 15 per cent in July 1989. This rate was maintained till 1994. But, since 1995, the *CRR* has been regularly reduced by the RBI until January 2006, as shown below. However, due to inflationary pressure in the economy, the RBI began to raise the *CRR* and raised it 8.75% in July 2008. With inflation rate declining, the RBI cut down the *CRR* to 5 per cent in June 2009.

3. Statutory Liquidity Ratio (SLR) In addition to *CRR*, the RBI was empowered to impose 'statutory cash reserve ratio' (*SLR*) between 3–15 per cent of banks' demand and time deposits to control and regulate the credit creation by the banks for the private sector and the availability of finance to the government. Under the *SLR* scheme, the commercial banks are required by statute to maintain a certain percentage of their total daily demand and time deposits in the form of liquid assets. Liquid assets, as specified by the RBI, include (*i*) excess reserves, (*ii*) unencumbered government securities, e.g., bonds of IDBI, NABARD, Development banks, cooperative debentures, debentures of port trusts, etc., and (*iii*) current account balance with other banks. The method of determining the SLR can be specified as follows.

$$SLR = \frac{ER + GS + CB}{DD + TD}$$

where *ER* = excess reserves, *GS* = Government (unencumbered) securities, *CB* = current account balance with other banks, *DD* = demand deposits, and *TD* = time deposits.

The basic purpose of using *SLR* was to prevent the commercial banks from converting their assets into liquidity when *CRR* was raised to control money supply. When *CRR* was raised, what commercial banks used to do was to convert their liquid assets into cash to replenish the fall in their funds due to the rise in the *CRR* and maintained their credit creation ability. This made monetary policy ineffective.

The *SLR* was first imposed in 1949, and was fixed at 20 per cent, and remained unchanged till August 1964. In September 1964, the *SLR* was raised to 25 per cent and was maintained at the same level till September 1970. Since then, the *SLR* has been raised

quite frequently. The *SLR* was raised in September 1990 to 38.5 per cent – very close to the prescribed upper limit of 40 per cent. The *SLR* has been used in fact as a monetary-fiscal tool. The deficit financing method – a fiscal measure – led to rapid increase in money supply which continued to build inflationary pressure in the economy. The RBI now used the *SLR* for controlling the short-term money supply. Since 1992, however, the *SLR* has been gradually reduced. It was reduced to 25 per cent in April 1992, mainly because the rate of inflation had declined to around 5 per cent in the early 1990s. The *SLR* continues to be maintained at 25 per cent.

4. Open Market Operations (OMO) In developed countries like the USA and the UK, open market operation is considered to be a very powerful and efficient tool of monetary management. But in India, the open market operation has not been used until 1980 as a successful instrument of monetary management for two reasons: (*i*) security market was not well developed, and (*ii*) government bonds were not popular. It is for these reasons that open market operation was not used until the mid 1980s to control money supply, nor was this tool effective when used. However, some important changes were made on the recommendations of the Chakravarty Committee (1985). The interest rate on Government securities was raised during the late 1980s and scheduled commercial banks were granted freedom to determine their own prime lending rates. These two factors made open market operation a fairly effective tool to control short-term credit. After the economic reforms of 1991–92, *OMO* was assigned a greater role in monetary management.

5. *The Repo Rate: A New Monetary Tool* Till the late 1980s, the RBI had been using the traditional methods of monetary control. However, on the recommendations of the Chakravarty Committee (1985), some important changes were made in the monetary policy. Some major changes were introduced in the monetary policy only after the foreign exchange crisis of 1990 and subsequent economic reforms. Some major monetary reforms and some new tools of monetary management were introduced including the *repo rate*. We describe here briefly a new monetary tool that is often used by the RBI, i.e., ***Repurchase Operation Rate – the repo rate.***

In April 1997, the RBI introduced a new system, called *Repurchase operation rate* (abbreviated as *repo rate),* to manage the short-run liquidity of the banking system. As mentioned above, under the *SLR* system, the commercial banks are required to invest a certain percentage of their demand and time deposits in government securities. This system blocks the bank money with the RBI, often causing liquidity problem. The repo system provides a solution to this problem of liquidity. Under the repo system, the RBI buys securities back from the banks and, thereby provides funds to the banks. It is a form of lending money to the banks for a short period 1–14 days. The rate of interest at which the RBI lends money to the bank is the *repo rate*. In contrast, there is *reverse repo rate*. The reverse repo rate is the rate at which the banks can buy the securities or deposit money with the RBI.

The operational rule of the repo rate is quite simple. When the central bank aims at increasing liquidity or money supply, it buys back the securities at a low repo rate. This increase the funds with commercial banks which can be used to create credit. On the other hand, when the objective is to control the money supply, the RBI uses the reverse repo rate and increases the repo rate. In June 1998, the repo rate was fixed at 5 per cent. However, due to anticipated increase in liquidity via Resurgent India bonds and East Asian crisis, the repo rate was raised to 8 per cent in August 1998. But it was later reduced gradually to 4.5 per cent in 2004, to

5 per cent on April 28, 2005, and to 6.25 per cent on October 26, 2006. However, due to mounting inflationary pressure in the economy, repo rate was increased to 7.25 per cent in 2006–07. Along with the changes made in the repo rate, the reverse repo rate was also simultaneously raised. In 2008, the Indian economy was facing a 13-year high rate of inflation which was touching 12 per cent. With the objective of controlling inflation, the RBI kept increasing the repo rate. On July 29, 2008, the RBI raised the repo rate from 8.5 per cent. On January 28, 2014, it was reduced to 8 per cent.

39.5.3 Evaluation of India's Monetary Policy

Monetary policy, or any policy for that matter, has to be evaluated by examining whether its objectives have been achieved over time. As mentioned above, on the recommendation of the Chakravarty Committee, the RBI had adopted 'price stability', i.e., controlling inflation, as the 'dominant objective of the monetary policy', while at the same time, maintaining an adequate liquidity in the economy. The Chakravarty Committee had recommendation price stability at 4 per cent rate of inflation. Even other economists have suggested, on empirical basis, that a 3–5 per cent annual inflation is desirable for a developing economy. Examined against the price stability objective, India's monetary policy appears to be only partially successful.

In India, inflationary pressure started building up during the 1960s, due to the Chinese war in 1962, the Pakistan war in 1965, and near-famine conditions in 1965–66. As a result, inflationary pressure started mounting from 1962–63, and inflation rate shot up to 13.9% in 1966–67. The decennial average rate of the 1960s was worked out at 6.4 per cent. The things were much worse in the 1970s. The inflation rate during the 1970s was much higher – the highest rate during the period was 25.2 per cent in 1974–75. This has been the highest rate of inflation in India so far. The decennial average inflation rate was 9 per cent, due mainly, to the failure of the kharif crop and rise in oil prices.

During the 1980s, things improved marginally. The decennial rate of inflation declined from 9 per cent in the 1970s to 8 per cent in the 1980s, with the highest inflation rate of 18.2 per cent in 1980–81. However, there was an upsurge of inflationary pressure during the first five years of 1990s. The average rate of inflation during the period from 1990–91 to 1995–96 was worked out at 10.6 per cent. Thereafter, however, the inflation rate declined considerably. The inflation rate varied between 3.4 per cent in 2002–03 and 6.4 per cent in 2004–05. The annual average rate of inflation during the period from 1995–96 to 2006–07 works out to be 5 per cent. This was quite close to the economically and socially desirable rate of inflation.

If one compares the high rate of inflation (varying between 6% and 10%), one would conclude that during the entire period from 1960–61 to 1995–96, i.e., during a period of 35 years, the monetary policy was not very successful in achieving its objectives. Although inflation rate was quite within the desirable limit 4–5% during the period from 1995–96 to 2006–07, it can hardly be attributed to the monetary policy. The lower rate of inflation was mainly due to high growth rate— 7–9 per cent. The only point that goes in favour of the monetary policy is the fact that things might be much worse in the absence of monetary controls adopted by the RBI. What is alarming is the fact that, in spite of all monetary measures, inflation rate shot up to about 12 per cent—to be precise, 11.98 per cent—in June-July 2008. However, had RBI not adopted a monetary policy with prime objective of price stabilization, inflation rate could have been much higher.

SUMMARY

- Monetary policy refers to the actions to be taken by the monetary authority to control and regulate demand for and supply of money with the purpose of achieving some predetermined objectives.
- As is the case with the objectives of any macroeconomic policy, objectives of monetary policy are (*i*) economic growth, (*ii*) employment promotion, (*iii*) economic stability, and (*iv*) stabilization of price level.
- The scope of monetary policy refers to the economic variables that can be influenced by change in monetary policy instruments. The scope of monetary policy covers all the macroeconomic variables that are affected by the demand for money and supply of money.
- Monetary policy instrument include (*i*) bank rate, (*ii*) cash reserve ratio, (*iii*) open market operation by the central bank, and (iv) direct controls and moral suasion. In India, monetary policy instruments include also (*a*) statutory liquidity ratio (SLR), and (*b*) repo rate.
- The effectiveness of monetary policy depends on (*i*) time lag – time required by the monetary instruments to affect the targeted economic variable, (*ii*) how accurately is forecast the effective working of monetary policy, (*iii*) how large, diversified and developed are the non-banking financial institutions, and (*iv*) how high is the level of development of money and capital markets.
- The objectives of India’s monetary policy are (*i*) price stability, and (*ii*) economic growth. However, according to the Chakravarty Committee, the dominant objective of India’s monetary policy is stabilization of inflation at 4 per cent.
- For price stabilization, the Reserve Bank of India has used over time almost all the instruments of monetary policy including (*i*) bank rate, (*ii*) cash reserve ratio, (*iii*) statutory liquidity ratio, (*iv*) open market operation, and (*v*) repo rate.
- As regards the success of monetary policy of India in controlling inflation, the inflation rates that prevailed in India since mid-1980s prove that monetary policy has not been successful. In fact, inflation rate in India had crossed double digits in many years.
- However, by examining the monetary policy against the build-up of high inflationary pressure in the country over the Plan period, economists conclude that had RBI not adopted anti-inflationary monetary policy, inflation rate in India would have been much higher.

REVIEW QUESTIONS

1. What is meant by monetary policy? How does it differ from fiscal policy?
2. Describe the instruments of monetary policy. How do they work and what are their limitations?
3. Distinguish between total credit control and selective credit control measures of monetary control. Under what conditions are the qualitative controls preferred to quantitative controls?
4. What is open market operation? How does it work to affect the money supply? Why is this measure considered to be more effective than other measures of monetary control?
5. Explain the transmission mechanism of monetary policy. How does a change in money supply changes the levels of income and prices? What is portfolio adjustment? How does Keynesian approach in this regard differ from the monetarist approach?

6. What are the factors that determine the effectiveness of monetary policy? How does empirical evidence corroborate with theoretical propositions?

7. What is meant by 'time lag'? How does it affect the effectiveness of monetary policy?
8. What are the limitations of the monetary policy in developed and less developed economies? Why is monetary policy less effective in the developing economies?
9. What are the objectives of India's monetary policy? Have the policy measures used by the RBI been effective in achieving the policy targets?
10. What monetary measures have been used by the RBI to control in the country? Which of the measures proved to be more effective.
11. Price stabilization has been the dominant objective of India's monetary policy. Has the RBI succeeded in stabilizing the price level at a desirable level.
12. Inflation rate in India touched 12 per cent in July 2008. What was the dilemma faced by the RBI and the government in adopting strong measures for controlling inflation. Explain in detail.

FURTHER READING

Chakraverty , Sukhmoy, *Report of the Working Committee of the Monetary System*, 1985.

DwiveDi, D.N., *Macroeconomics: Theory and Policy* (Tata McGraw-Hill, 2010), Ch. 30.

FrieDmaN, M., "The Role of Monetary Policy", *Am Eco. Rev.,* March 1968.

FrieDmaN, M., "Monetary Policy: Theory and Practice", *Journal of Money, Credit and Banking,* February 1982.

moDigliaNi, Franco, "The Monetarist Controversy: Should We Forsake Stabilization Policy". *Am. Eco. Rev.,* March 1977.

raNgarajaN, C., "The Analytical Framework of the Chakravarty Committee Report on the Monetary System", *Reserve Bank of India Bulletin,* September 1987, p. 702.

raNgarajaN, C., "Issues in Monetary Management" in *Indian Economy: Essays on Money and Finance* (Oxford, Delhi, 1998).

Shapiro, E., *Macroeconomic Analysis* (Galgotia, New Delhi, 1994), Ch 26.

tobiN, James, "Monetary Policy: Rules, Targets and Shocks", *Journal of Money, Credit and Banking,* November 1983.

CHAPTER

40 Fiscal Policy

CHAPTER OBJECTIVES

The objective of this chapter is to discuss some important aspects of another macroeconomic policy, the fiscal policy, in general and India's fiscal policy in particular along with its impact on private business. The discussion here contains the following aspects of fiscal policy.

- Meaning and purpose of fiscal policy
- Instruments of fiscal policy
- Scope of fiscal policy
- Some basics of India's fiscal policy
- Taxation policy of India and
- Impact of fiscal policy on private business

40.1 INTRODUCTION

Having discussed monetary policy in the preceding chapter, we proceed now to discuss fiscal policy. While monetary policy deals with aggregate demand for and supply of money, fiscal policy deals with money income transferred from public purse to government coffer and from the government coffers to public purse. In fact, money is transferred from the public purse to government coffer by means of taxation and borrowings and money that flows back to public purse takes the form of government expenditure. Government's action plan related to government revenue (taxation) and expenditure constitutes the fiscal policy. In this chapter, we will discuss the various aspects of fiscal policy as noted above.

40.2 MEANING AND OBJECTIVES OF FISCAL POLICY

40.2.1 Meaning of Fiscal Policy

Look at the origin of the world 'fiscal'. The word 'fisc' means 'state treasury'. Thus, the term 'fiscal policy' is related to 'public treasury', i.e., how government raises its revenue and how it spends the treasury funds. Thus, fiscal policy refers to the action plan of the government to raise its funds and to spend the funds to achieve certain predetermined objectives. Fiscal policy has, however, been defined variously by the economists. For example, Arthur Smithies has defined fiscal policy as "a policy under which government uses its expenditure and

revenue programs to produce desirable effects and avoid undesirable effects on the national income, production, and employment.''1 Samuelson and Nordhaus offer a more comprehensive definition of fiscal policy. By fiscal policy they "mean the process of shaping taxation and public expenditure to help dampen the swings of the business cycle and contribute to the maintenance of a growing, high-employment economy, free from high or volatile inflation."2 In their opinion, the role of fiscal policy is confined largely to stabilization of employment and the price level. It seems, they have defined fiscal policy keeping in view the problems of the developed countries. Fiscal policy can be defined in more general terms as follows.

Fiscal policy is the government programme of making discretionary changes in the pattern and level of its expenditure, taxation and borrowings in order to achieve certain economic goals such as economic growth, employment promotion, income equality, and stabilization of the economy on a growth path.

A narrow concept of fiscal policy is ***budgetary policy.*** While *budgetary policy* refers to current revenue and expenditure of the financial year, fiscal policy refers to budgetary operations including both current and capital receipts and expenditure. The essence of fiscal policy lies, in fact, in the budgetary operations of the government. The two sides of the government budget are *receipts* and *expenditure*. The ***total receipts*** of the government are constituted of tax and non-tax revenue and borrowings (including deficit financing). These items in the budget represent the budgetary resources of the government. The ***government expenditure*** refers to the total expenditure made by the government in the fiscal year. The total government expenditure consists of payments for goods and services, wages and salaries, interest and loan repayments, subsidies, pensions and grants-in-aid, and so on. The government can, by using its statutory powers, change the magnitude and composition of inflows and outflows and thereby the magnitudes of macroeconomic variables—aggregate consumption expenditure and private savings and investment. The magnitude and composition of inflows and outflows can be altered by making changes in taxation and government spending. The policy under which these changes are made is called *fiscal policy*.

40.2.2 Objectives of Fiscal Policy

Fiscal policy as defined above, is the government's programme of taxation, expenditure and other financial operations to achieve certain national goals. The objectives of fiscal policy, like those of other economic policies of the government, are derived from the 'aspirations and goals' of the society. Since the economic, political and social conditions of the nations vary, the aspirations of the people and, therefore, the objectives of their fiscal policy may be different. However, the most common objectives of fiscal policy of different countries are:

(*i*) economic growth;

(*ii*) promotion of employment;

(*iii*) economic stability and

(*iv*) economic justice or equity.

1. Arthur Smithies, "Fiscal Budgeting and Fiscal Policy." in *A Survey of Contemporary Economics,* Vol. I (The Blakiston Co., Philadelphia, 1949), p. 174.

2. Samuelson, P. A. and Nordhaus, W. D., Economics (McGraw-Hill, NY, 15th Edn., 1995), p. 626.

The emphasis and the order of priority of these objectives may vary from country to country and from time to time. For instance, while stability and equality get higher priority in the developed nations, the objectives like growth, employment and equality get higher priority in the less developed countries. Whatever the objectives and the order of their priorities, the two basic instruments that are used to achieve the social goals are *taxation* and *public expenditure*. In this chapter, we will concentrate on India's taxation policy and its possible impact on the private business of the country.

40.3 FISCAL INSTRUMENTS AND TARGET VARIABLES

Fiscal policy is implemented through fiscal instruments also called 'fiscal handles', 'fiscal tools' and 'fiscal levers'. The changes made in fiscal tools work through their linkage to the target variables. The policy instruments and target variables are briefly described below.

40.3.1 Fiscal Instruments

Fiscal instruments refer to the budgetary measures which the government uses and manipulates to achieve some predetermined objectives. The major fiscal instruments include the following measures.

(*i*) Budgetary policy—deficit or surplus budgeting,

(*ii*) Government expenditure,

(*iii*) Taxation, and

(*iv*) Public borrowings.

The features and the working of these fiscal instruments are briefly described here.

(i) Budgetary Policy. Budgetary policy refers to government's plan to keep its budget in balance, in surplus or in deficit. This kind of budgeting is in itself a fiscal instrument. When the government keeps its total expenditure equal to its revenue, as a matter of policy, it means it has adopted a ***balanced-budget policy.*** When the government decides to spend more than its expected revenue, as a matter of policy, it is pursuing a ***deficit-budget policy.*** And, when the government adopts a policy of keeping its expenditure substantially below its current revenue, it is following a ***surplus-budget policy.*** Balanced, deficit and surplus budgets affect the economy in different ways, to different extents, and in different directions.

(ii) Government Expenditure. The government expenditure includes total public spending on purchase of goods and services, payment of wages and salaries to public servants, public investment, infrastructure development, transfer payments (e.g., pensions, subsidies, unemployment allowance, grants and aid, payments of interest, and amortization of loans). Given the expendable resources of the government, the size and the composition of government expenditure is a matter of government discretion based on its objectives.

The government expenditure is an injection into the economy; it adds to the aggregate demand. The overall effect of government expenditure on the economy depends on how it is financed and it is spent and what is its multiplier effect.

(iii) Taxation. A tax is a ***non quid pro quo*** payment by the people to the government, i.e., tax is a payment by the people to the government against which there is no promise for any direct return to the tax payers. By this definition, taxation means ***non quid pro quo*** transfer of private income to public coffers by means of taxes. Taxes are classified as ***direct taxes*** and ***indirect taxes.***

Direct taxes include taxes on personal incomes, corporate incomes, wealth and property. Personal income tax and corporate income tax are the two most important direct taxes imposed by the central government in India. In 2010–11, personal income tax contributed 17.5 per cent and corporate income tax 37.7 per cent of the gross tax revenue. These two direct taxes together contributed 55.3 per cent of the gross tax revenue of the central government. More importantly, corporate income tax has of late emerged as the most important single source of government revenue.

Indirect taxes includes taxes on production and sale of the goods and services. Indirect taxes are also called ***commodity taxes***. In India, the two most important central indirect taxes are excise duty (or VAT) and customs. In 2010–11, central excise yielded 17.4 per cent and customs 17.1 per cent of the gross tax revenue3. These two taxes together contributed 34.5 per cent of gross revenue in 2010–11.

(iv) Public Borrowings. Public borrowings include both internal and external borrowings. The governments make borrowings, generally, with a view to financing their budget deficits. ***Internal borrowings*** are of two types: (*i*) borrowings from the public by means of government bonds and treasury bills, and (*ii*) borrowing from the central bank. The two types of borrowings have different kinds of effects on the economy. Borrowings from the public to finance budget deficit is, in effect, simply a transfer of purchasing power from the public to the government, whereas borrowings from the central bank for financing budget deficits, i.e., ***monetized deficit financing***, is straightaway an injection into the economy. ***External borrowings*** include borrowings from (*a*) foreign governments, (*b*) international organizations like World Bank and IMF, and (*c*) market borrowings. It has the same effect on the economy as deficit financing. In India, the total borrowing accounted for about 8 per cent of the total central government expenditure in 2012–13.

40.3.2 Target Variables

In the Keynesian framework of analysis, the ultimate target variable of fiscal policy is the intended change in the ***aggregate demand.*** The change in aggregate demand is sought through the change in its various components and level, and in the price structure. The target variables of fiscal policy, i.e., the variables which are sought to be changed through fiscal instruments are following:

(*i*) Private disposable incomes,

(*ii*) Private consumption expenditure,

(*iii*) Private savings and investment,

(*iv*) Exports and imports, and

(*v*) Level and structure of prices.

40.4 INDIA'S FISCAL POLICY

India's fiscal policy was formulated with the purpose of finding solution to economic problems confronted by the country at the time of Independence. The most dominant problem faced by the country was economic backwardness and wide-spread poverty. The level of backwardness was so low that the rate of capital formation was estimated to be only 5% of national income. According to the First Five Year Plan (1951–56), the rate of

3. *Economic Survey*, 2012–13, Economic Division, Ministry of Finance, Government of India, p. 61, Table 3.4.

capital formation had to be increased from "5 per cent of the national income to, say, about 20 per cent" (p.17). In view of this kind of development targets, the most challenging task for the Government was to mobilise adequate financial resources to finance the development programmes chalked out in Five Year Plans. Raising resources was a challenging task because none of the known sources of developmen t finance – taxation, domestic borrowing, external borrowing, and foreign aid and external assistance4 – had the potentials of yielding adequate financial resources to meet the development needs of the country. Therefore, India had to rely on her domestic resources. It was under these conditions that India had formulated her fiscal policy. Let us first have a look at the objectives of India's fiscal policy and proceed to discuss taxation policy, the main source of government revenue.

40.4.1 Objectives of India's Fiscal Policy

In general, the objectives of India's fiscal policy are the same four objectives as mentioned above, viz., (*i*) achieving economic growth, (*ii*) employment promotion, (*iii*) ensuring economic stability, and (*iv*) providing social justice to the people of the county. The objectives provide the overall framework of the fiscal policy of India. Though these objectives are generally interrelated, the priority of and emphasis on these objectives change from year to year depending on the need of the country. This is indicated in the annual budget of the Government. Let us now look at the allocation of resources as shown in budgetary framework.

Fiscal policy is operated in India by two action plans: (*i*) mobilising financial resources, and (*ii*) allocation of financial resources to *development* and *non- development plans*. Practically, these actions take, respectively, the form of (*a*) public revenue, and (*b*) public expenditure. In India, the main source of revenue is *taxation*.

As mentioned above, the important issue for our purpose here is to see how fiscal policy affects private sector business. As regards the ***effect of fiscal policy on private business,*** taxation is deemed to affect private business adversely and public expenditure is supposed to affect private business positively. From analytical point of view, whether fiscal policy affects private business positively or negatively depends on the *net effect* of public expenditure and taxation. Examining effect of taxation involves a wide perspective. Therefore, we will discuss taxation policy, taxation structure, and effects of taxation on private business in the next section. Here we look the possible effects of public expenditure on private business.

40.5 PUBLIC EXPENDITURE IN INDIA

Public expenditure is the main instrument of achieving the objectives of fiscal policy, the objective of economic growth. In India, public expenditure is classified in government budget under two categories: (*i*) Plan expenditure, and (*ii*) Non-Plan expenditure. Plan expenditure is basically the *development expenditure* and non-plan expenditure is the *administrative expenditure* on maintaining law and order, national defence, and social infrastructure including transport, communication, education and health. However, both kinds of public expenditure have their impact on private sector economic functions. Before we describe how public expenditure affects private economic activities, let us have a look at growth of public expenditure in India.

4. Incidentally, India could manage 'external assistance' to a considerable extent which accounted for 22.4% of the Second Plan outlays and 28.2% of the Third Plan outlays. But external assistance declined drastically in the subsequent plans.

40.5.1 Growth of Public Expenditure

Public expenditure has increased in India intensively as well as extensively, i.e., in terms of both amount and areas of public expenditure since 1950–51. Decade-wise growth in total public expenditure (including expenditure of both central and state governments) is given below in Table 40.1. As can be seen in the table, public expenditure (at current price) increased from `900 crore in 1950–51 tò2,631 crore in 1960–61 and tò7,843 crore in 1970–71. Since 1980–81, public expenditure tended to increase more rapidly. Public Expenditure increased from `37,218 crore in 1980–81 tò163,520 crore in 1990–91 and tò588,233 crore. In 2010–11, public expenditure shot up tò22,40,369 crore.

Table 40.1 *Growth of Public Expenditure in India: 1950–51 to 2012–13*

Year	*Public Expenditure (` in crore)*	*Percentage to GDP*
1950–51	900	9.1
1960–61	2,631	15.3
1970–71	7,843	17.2
1080–81	37,218	25.6
1990–91	1,63,520	28.7
2000–01	5,88,233	28.0
2010–11	22,40,369	29.2

Sources: RBI, *Handbook of Statistics on Indian Economy —2009– 10,* Table 115 and *Economic Survey: 2011– 12,* Governement of India, Table 3.12. (Reproduced in S.K. Mishra and V.K. Puri, *Indian Economy* (Mumbai: Himalaya Publishing House, 2012, p. 669).

More importantly, the ratio of total public expenditure to GDP increased from 9.1 per cent in 1950–51 to 29.2 per cent in 2010–11. Thus, total public expenditure accounts for about 30 per cent of the national income.

40.5.2 Effects of Public Expenditure on Private Business

The high rate of increase in public expenditure accounting for about 30 per cent GDP is bound to have a serious impact on the private business. Both development and non-development expenditure have a positive effect on private business entrepreneurs, but in different way and to a different extent. Measuring effect of public expenditure on private business is a complex and difficult task. We point out here how development expenditure and non-development expenditures are supposed affect the private business activities.

Development expenditure brings about growth in the economy. Economic growth widens the prospects for private business. Therefore, private sector gains from development expenditure. As regards the non-development expenditure, it adds to personal income of those who get employment or contractual deals with government agencies. Increase in income creates demand which widens the business prospects. Besides, given the time-lag between the demand rise and supply, prices tend to increase. Price rise adds to profit because cost does not increase shortly and at the same rate – cost rise has a time lag. Rising profit promotes private investment and growth of private business. To sum up, public expenditure has a positive effect on private business provided it does not lead to hyper inflation.

Now we proceed to discuss in detail the nature, scope and pattern of taxation and their impact on private business.

40.6 TAXATION POLICY OF INDIA

Taxation policy of India was first formulated in 1950–515. The taxation policy underwent several modifications and changes during the period from 1950–51 to 1990–91. In 1991, taxation policy was reformed drastically. Therefore, taxation policy of India should be analysed under two stages, i.e., (*i*) taxation policy from 1953–54 to 1990–91, and (*ii*) taxation policy in post-1991 period.

40.6.1 Formulation of India's Taxation Policy: 1953–54 to 1990–91

The tax policy that existed during 1950–1990 was the result of recommendations of several tax enquiry committees and review panels and deliberations on the policy in the Parliament. The tax policy was formulated, reformulated and reformed several times with view to making it fair, equitable and efficient. The efforts to formulate an appropriate tax policy began in the early 1950s with the appointment of a number of enquiry and reform committees6 one after another.

40.6.2 Basic Functions of Taxation Policy

A tax policy is supposed to perform multiple functions, depending on the needs of the country. India's taxation policy has been changing from time to time in the process of tax reforms. However, India's taxation policy has, in general, been designed to perform two basic functions:

(*i*) Revenue function, and

(*ii*) Regulatory functions.

1. *Revenue Function.* Revenue collection has been the primary objective of India's tax policy. In pursuance of its revenue objective, the Central government and also the state governments, used their taxing powers extensively and intensively. As regards the *extensive use of taxes*, the governments cast their tax net as far and wide as possible. The Centre imposed five new taxes—estate duty7, wealth tax, gift tax, expenditure tax8 and capital gains tax during the 1950s. Progressive and high tax rates were imposed in case of both direct and indirect taxes. Both direct and indirect tax rates were, in general, revised

5. First Five-Year Plan (1951–1956), p. 17.

6. (*i*) *Taxation Enquiry Committee (TEC)* in 1953–54 to suggest suitable tax measures for mobilizing additional tax revenue;

(*ii*) *Nicholas Kaldor Committee,* under the chairmanship of Prof. Nicholas Kaldor, a renowned tax expert of Britain, was set up in 1956 to suggest new tax measures to augment government revenue;

(*iii*) *Direct Taxes Administration Committee* (Tyagi Committee) was appointed in 1958 to suggest measures for (*i*) a scheme of integration of direct taxes, (*ii*) preventing tax evasion and (*iii*) simplifying the procedure of tax compliance;

(*iv*) The *Committee on Rationalization and Simplification of the Tax Structure* (Bhoothalingam Committee) was set up in 1967 to suggest measures to reform the tax system and to prevent tax evasion;

(*v*) *Direct Taxes Enquiry Committee* (Wanchoo Committee) was appointed in 1971 to suggest tax reform measures to prevent tax evasion;

(*vi*) *Indirect Taxation Enquiry Committee* (Jha Committee) was appointed in 1976 to find and examine the sources of anomalies in the indirect tax system and to explore the possibility of implementing Value Added Tax (VAT) system in place of excise duties; and

(*vii*) *Direct Tax Laws Committee* (Choksi Committee) was set up in 1978 to suggest measures to simplify and rationalize tax laws and to improve their implementation.

7. Estate duty was abolished in 1991 because revenue from this tax was insignificant.

8. Expenditure tax was abolished in 1963 due to administrative problems in its administration and insignificant revenue yield.

upwards intermittently. Central excise duty was extended to all imaginable non-agricultural products. Extension of excise duties in India was 'unprecedented'9. Prohibitively high import duty was imposed on almost all items of imports.

The state governments on their part, imposed several new taxes, in addition to sales tax, viz., agricultural income tax or large holding tax, surcharge on cash crops, profession tax, tax on urban property, sales tax on motor spirit (petrol), motor vehicle tax, tax on passengers and goods, entertainment and betting tax, state excise on alcohol and narcotics, though many of these taxes did not prove revenue productive.

The upward revision of excise duty in India was 'unparalleled'10. As a result, excise duty rates increased manifold between 1960–61 and 1971–72. For example, excise duty on synthetic and artificial silk increased by 2232%, on vegetable product (vanaspati) by

1428%, on matches 737%, cigars and cigarettes by 545%, on kerosene by 350%, on cotton cloth (A) by 280% and on tea by 174%.

Another important feature of India's taxation policy prior to 1975–76 was to make income tax rates highly progressive with increased tax rates on all the income slabs, though exemption limit was raised too. Consequently, income tax rates for higher income groups had risen so much that it amounted to almost confiscation of income beyond a certain level. For example, in 1973–74, marginal rate of personal income tax on an income of `500,000 was 89% and on incomes over `10,00,000, the tax rate was 95.7% and with surcharge added, the marginal tax rate for the latter category rose to 97.75%. The corporate income tax rate was raised from 35% in 1957 to 45% in 1966 and to 55% in 1978–79. With surcharge tax added, the corporate tax rate in 1978–79 rose to 61%.

2. *Regulatory Function.* Regulatory functions of taxation include:

(*i*) reducing disparities in income and wealth distribution;

(*ii*) restraining consumer demand with a view to containing inflation;

(*iii*) promoting savings and investment and

(*iv*) shifting investment from non-essential and non-priority sectors to priority sectors.

In pursuance of these goals, progressive tax rates were imposed; excise duties were enhanced more and more on the so-called luxury goods; incentives were provided for savings and tax holidays were granted for new investments.

It is noteworthy that revenue and regulatory functions of taxes work together and it is extremely difficult to separate the effects of the two functions.

40.7 TAXATION POLICY REFORMS (1991)

The taxation policy that existed in India between 1950 and 1990 had a number of defects and anomalies which defeated many objectives of taxation policy. Ironically, the anomalies in India's tax system continued to grow, tax evasion increased at an increasing rate, tax-evaders continued to thrive and inequity in taxation continued to widen in spite of volumes of suggestions made by the expert committees for reforming the tax system and their suggestions implemented partially or wholly. In fact, India's tax system remained beset with many dangerous anomalies. Let us have a brief look at India's taxation policy prior to 1991.

9. Due, John F., *Indirect Taxation in Developing Economies* (Johns Hopkins Press, 1970), p. 27.

10. Mahler, W.R., Jr., *Excise and Sales Taxation in India* (Orient Longmans, 1970), p. 105.

40.7.1 Anomalies in India's Tax Policy

The Tax Reform Committee (TRC) pointed out the *defects and anomalies in India's tax system*. In its opinion, India's tax system had grown over time in such a way that it had become inefficient, inequitable, regressive, unjust, cumbersome and difficult to administer. Some of the major defects and anomalies of India's tax system till 1991, as pointed out by the TRC, are briefly described here.

1. *Very high tax rates.* Central tax rates in India were much higher compared to the tax rates in comparable countries. The high tax rates had harmed the economy, the society, the taxpayers and the administration. High rates of income tax, company tax and excise duties were solely responsible for the rampant tax evasion in the country. According to the TRC, "High tax rates… without indexation and lack of effective enforcement are the main factors contributing to large scale tax evasion. Lack of effective enforcement means very little fear of being detected and punished".

2. *Tax system of cascading nature.* All indirect taxes are of cascading nature. Tax cascading means taxes piling on taxes and tax on tax so that actual tax burden borne by the society is much higher than one calculated on the basis of the tax rates. In the absense of tax cascading, the actual tax burden on the society would be equal to the calculated tax burden. Apart from excess tax burden, tax cascading distorts the price structure and, thereby, the resource allocation.

3. *Administratively complex tax system.* The tax system over time has become so complex that it was extremely difficult to understand and administer. Complexity arises due to multiplicity of tax laws, provisions and sub-provisions, sections and sub-sections with respect to definition of tax base, exemptions and concessions. The complexity of the tax system and tax laws gives tax administrators ample opportunity to interpret the law in their own way with a view to harass the tax payers with the aim of extracting a share in the concessions and making money. This, practice is ubiquitous in all the tax departments.

One is flabbergasted to come across the news, 'CBI raids the houses of tax officials'.

4. ***Anomalies in individual taxes.*** Apart from pointing out the defects in the tax system as a whole, the TRC pointed out the anomalies in the individual taxes. In this regard, we will confine our discussion to three major taxes, viz., personal income tax, import duties and excise duties. The TRC did not find much against the *corporate income tax* except that its effective rate was very high at 51.75 per cent.

Personal income tax has 'serious anomalies and inequities'. It is *anomalous* because it provides 'tax shelters' to the members of parliament, central government ministers, a section of government officials and top executives of private firms by leaving their perks tax-free. There is no rationale for not taxing perks which account for a considerable proportion of their real income. It is *iniquitous* because it discriminates between different categories of tax payers. For example, the salary income of the government servants has a lower tax burden than other categories of salary earners because the former category is provided housing at nominal rent. Personal income tax is regressive because, according to TRC, at 1990–91 prices, the tax payers with incomes between `50,000 and `100,000 had the highest income tax burden and those with incomes abovè500,000 had the lowest tax burden.

Import duties had, according to the *TRC*, the following anomalies: (*i*) the average rate of import duty (125%) was much higher than international standards, (*ii*) tariffs were widely dispersed and complex, and administratively cumbersome, (*iii*) import tariffs had a multiple rate system—basic, auxiliary as well as additional, and (*iv*) tariffs had a complicated system of concessions granted by notification which made the system not only administratively inefficient but also created room for arbitrary use of discretionary powers by the customs officials.

Excise duty, the largest contributor to the central revenue, had the following defects: (*i*) it had unimaginably complex multiple rate system for a commodity classified under different categories, (*ii*) it is of cascading nature as it makes the same commodity taxable at different stages of its production, (*iii*) excise duty fell also on capital goods like machinery, tools, accessories, office equipment, etc., and (*iv*) the excise concession for small and tiny sectors encouraged manipulation of firm's size for the purpose of tax evasion.

40.7.2 Tax Reforms in 1991

It was due to the foreign exchange crisis of 1990 and the *IMF* conditions for providing financial help to tide over the crisis, a *Tax Reform Committee* (Chelliah Committee) was appointed in 1991 to examine the direct and indirect tax structure and to suggest measures to (*i*) improve the elasticity of tax revenue, (*ii*) make the tax system fairer and more broad based, (*iii*) rationalize the direct tax system by removing its anomalies, (*iv*) improve equity and sustain economic incentives, (*v*) identify new areas for taxation, (*vi*) improve compliance of direct taxes and strengthen enforcement, (*vii*) simplify and rationalize customs so as to improve international competitiveness of Indian exports, and (*viii*) simplify and rationalize the structure of excise duties for better tax compliance and to widen the scope for MODVAT (a modified value added tax) scheme.

Although the basic tax structure of the country remains intact, the Government of India made sweeping changes in the taxation policy on the recommendations of the TRC. The core tax reforms and those with far reaching consequences are described here tax-wise.

1. ***Personal income tax.*** The rate structure of the personal income tax has been reduced from about six slab-rates in the 1980s to three slab-rates: 10%, 20% and 30%. The most significant reform is cutting down the tax rate on the top income bracket from about 67% in the late 1980s to 30%. This has solved many problems but has made the tax system regressive.

2. ***Company income tax.*** The company tax rate has been reduced to 40% for domestic and 45% for foreign companies. The number of concessions granted to the companies under Sections 35CCA and 34AC have been withdrawn.

3. ***Excise duties and MODVAT.*** Excise duty rates have been modified across the board. The number of classifications of commodities under tax laws has been substantially reduced. The procedure of excise calculation has been simplified. This makes tax compliance a much easier task. The process to replace the excise duties with MODVAT is underway. Most state governments have now adopted MODVAT.

4. ***Import duties.*** Import duties which appeared to be "a bewildering picture of combinations of 'basic' and 'auxiliary' duties" have been simplified. The duty rates have been slashed across the board so that the weighted-average effective rate comes down from about 85% to 45% and then to 35% in 2006–07 budget. Depending on the needs of the country, import duties are changed almost regularly in annual budgets.

40.8 INCOME TAXATION AND THE PRIVATE BUSINESS

A multiple tax system has widespread ramifications on the economy and different kinds of taxes have different kinds of effects on the private business. Taxes affect the economy in many ways by affecting macro variables like consumption, saving, investment, price structure, price levels and work effort. In India, there is a wide range of direct and indirect taxes. **Direct taxes** include personal and corporate income taxes on current earnings, wealth tax and gift tax on transfer of property. **Indirect taxes** include excise duty, sales tax, custom duties and a number of other taxes imposed by the states. Not only is there multiplicity of taxation, but also double taxation of incomes and commodities.

The total revenue of the country accounts for about 22 per cent of the national income. Such widespread and heavy taxation cannot be supposed to be neutral to private business activities. In fact, it is alleged by the business community, even after the tax reforms of 1991 and 1992, that the existing Indian tax structure is seriously undermining the incentive to save and invest for both individuals and corporations. However, taxation reforms made in India since 1991 have reduced the tax rates to internationally comparable levels and are expected to have much less adverse impact on the private business.

Measuring how tax has affected the private business is an extremely complex affair. In this section, however, we will confine our discussion to how income tax affects the private business in general. The effect of the corporate income tax on private business will be discussed in the next section.

The impact of income taxation on the growth of private business in general, and on private investment in particular, may be examined through its effects on (*i*) people’s work-efforts; (*ii*) saving of the households in general and of private firms in particular, and (*iii*) incentive and ability to invest.

It is **important** to note at the outset that there is little evidence available in case of India to support or refute the proposition regarding the adverse effects of taxation on saving and investment.11 The empirical evidence available for other countries is not strictly conclusive, and even if it is, the same may not be applicable to the Indian economy. We will, therefore, confine our discussion to only theoretical propositions regarding the effects of various taxes on private investment.

40.8.1 Effects of Income Taxation on Work Efforts

The effect of taxation on private enterprise depends, among other things, on how income tax affects people’s desire to work. The additional work effort depends, in fact, on peoples’ choice between leisure and work.12 Leisure gives a kind of satisfaction (or pleasure) while work yields income which yields another kind of satisfaction. Taxation of personal income reduces return from labour and, therefore, it alters peoples’ choice between leisure and work. When a tax is imposed or income tax rate is increased, wage income decreases. As a result, the reward for an additional labour and the price of additional leisure, i.e., opportunity cost of leisure, are both lowered. Under this condition, ‘the worker will tend to substitute leisure

11. The only comprehensive and reliable study, *Taxation and Private Investment*, was made for 1950s by the National Council of Applied Economic Research, New Delhi.

12. For a simple analysis of tax-effect on choice between work and leisure, see, A.R. Prest, *Public Finance in Theory and Practice*, (Weidenfeld and Nicolson, London, 1975), pp. 74–8.

for work.’ Thus, taxation reduces the supply of labour. But, at the same time, increase in tax rate reduces the total income from given hours of labour. It makes the worker poorer but poor workers normally wish to enjoy fewer hours of leisure in order to earn more. The workers would, therefore, like to work more to raise their income. Thus, taxation has both negative and positive effects on labour supply. The net effect of taxation on work effort (or labour supply) depends on the relative strength of the two effects13.

A number of surveys and econometric studies carried out in the United States and England on this aspect of taxation have not yielded any definite measure of the net effect of taxation on work effort14. According to Musgrave and Musgrave, “There is no *a priori* basis on which to judge the direction in which the net effect will go, although it is reasonable to assume that effort will decline."15 They have, however, contradicted themselves (in the next paragraph) by saying. " it should not be readily assumed that an income tax must reduce effort." Sanders has found that "the typical (business) executive [does not] put forth his best efforts, taxes or no, to fulfill the requirements of his job and to progress on the promotional ladder of his company."16 George F. Break17 interviewed 306 lawyers and accountants in England—an ideal group to study as they belonged to the category of tax-payers who can easily adjust their working hours with changes in their incomes. According to his findings, '40 men reported definite adverse effect on incentive' for additional work, 32 men reported to have worked harder due to taxation as some of them wanted to accumulate wealth and some wanted to maintain their standard of living. The remaining 234 men reported minor or no effect on their work effort.

It may be inferred from these empirical evidences that taxation of income has only marginal effect on work effort. Although under the conditions mentioned above, any generalization would be risky, much of tax effect on work efforts depends on (*i*) the level of income; (*ii*) tax-rates—proportional, progressive or regressive; (*iii*) the productivity or marginal efforts and (*iv*) non-monetary benefit, such as free accommodation, education of children, health care, travel benefits, etc.

40.8.2 Effect of Taxation on Savings

Let us now examine the expected effects of taxation on household and corporate savings. Household savings will be examined in respect of personal income-tax, and corporate savings in respect of corporate income tax and the tax-treatment of depreciation and retained earnings.

Effect of Income Taxation on Personal Savings As a matter of economic law, saving is a function of disposable income: the higher the income, the higher the average propensity to save. This law states that when income increases, the marginal propensity to consume decreases and marginal propensity to save increases. This implies that the households in higher income brackets save more and supply a major proportion of savings to the capital market.

So the Question Arises What about the effect of tax on savings of higher income groups? Taxation reduces the disposable income and the capacity to save. Since marginal propensity to

13. Break, George F., 'The Incidence and Economic Effects of Taxation, in *The Economics of Public Finance*, (The Brookings Institution, Washington, D.C., 1974), p. 180. See also R.A. Musgrave, and P. Musgrave, *Public Finance: Theory and Practice,* 2nd Ed., (McGraw-Hill Kogakusha Ltd. Tokyo, 1976), p. 407.

14. Break, George F. *op. cit.*, p. 185.

15. *Op. cit.*, p. 484.

16. Quoted from Otto Ekstein, *op. cit.* p. 81.

17. "Income Taxes and Incentive to Work: An Empirical Study, *"Amercian Economic Review,* Vol. XLVII, No. 5, September 1955, Quoted in Otto Ekstein, *op. cit*.

save is higher in the higher income classes, a highly progressive income tax reduces the overall rate of saving. "Taxation effect on saving may result not only because the tax payer's income is reduced, but also because an income tax reduces the net rate of return on savings."18 For this reason, households may prefer to increase their present consumption rather than to save and invest their income for future larger consumption. If the rate at which the households substitute future consumption for present consumption decreases, the rate of saving also decreases, given the rate of interest. Through a study of 750 active investors, mostly belonging to high income-brackets, Butters19 *et. al.* have found that income taxation had substantially reduced the capacity of the investors to accumulate new investible funds and it had changed the investment pattern.

In case of India, a study of the 'attitudes and reaction of individual investors and of the trends in the stock market' indicates that the willigness of the public to invest in equity shares has not been perceptively affected.20 The incentive to invest, particularly in the higher income groups, was found to be sustained "to some extent by the concessional treatment of capital gains," though their capacity to invest out of current saving was materially curtailed. The study also indicates that 'the combined impact of income and wealth taxes tends to severely curtail the capacity to save of active entrepreneur-investors in the larger ranges of income and wealth.'

The Effect of Indirect Taxation on Saving**:** ***A Digression.*** According to Musgrave and Musgrave,21 the effect of indirect taxes on saving is comparatively less retarding at least for two reasons. ***First,*** the incidence of indirect taxes, unlike the direct taxes, tends to be regressively distributed. The regressiveness of indirect taxes is based on a decreasing marginal propensity to consume as income rises. ***Second,*** 'consumption tax' (or commodity tax) does not reduce the rate of return on savings and, therefore, avoids the substitution effect of the income tax, which is adverse to saving.” In case of India, however, this argument is not tenable because as the studies made by the Ministry of Finance (MOF) for 1958–59 and 1963–64, and by the National Institute of Public Finance and Policy (NIPFP) for 1973–74 have revealed, incidence of indirect taxes in India is fairly progressive. The findings of these studies are summed up in Table 40.2.

Table 40.2 *Incidence of Indirect Taxes: Tax as Percentage of Consumer Expenditure*

Ministry of Finance			*National Institute of Public Finance and Policy*	
Monthly Household Expenditure Group (`)	*1953–54*	*1963–64*	*Monthly per Capital Expenditure Group* (`)	*1973–74*
0–50	2.4	6.5	0–15	2.96
51–100	2.7	7.0	15–28	3.63
101–150	3.1	8.0	28–43	4.89
151–300	3.3	10.1	43–55	6.85
301 and above	5.5	16.6	55–75	7.92
			75–100	11.40
			100 and above	21.96

Sources: (*i*) *Incidence of Indirect Taxation*—1963–64, (Ministry of Finance, Government of India, New Delhi, 1969). (*ii*) Chelliah, R.J. and Lal, R.N., *Indirect Taxation in India*—1973–74, National Institute of Public Finance and Policy, 1978, p. 17.

18. Musgrave and Musgrave, *op. cit.*, p. 491.

19. Quoted from Otto Ekstein, *op. cit.*

20. National Council of Applied Economic Research, *Taxation and Private Investment.* pp. 77–78.

21. *Op. cit.*, p. 491.

The progressiveness of indirect taxes may be relatively less than that of income taxes. It is however difficult to show that indirect taxes are less of a deterrent to saving than income tax since both kinds of taxes are simultaneously payable by the richer section of the society.

40.9 CORPORATE INCOME TAXATION, BUSINESS SAVING AND INVESTMENT

Corporate income taxation is the most important factor that affects directly the growth of private sector business. Corporate income tax affects the firm's own savings and investment by (*i*) decreasing the incentive to save and invest by reducing the rate of return; (*ii*) reducing the internal investible funds, and (*iii*) increasing the risk in the investment22. Apparently, the corporate income tax is expected to be a deterrent to the firm's capacity to save and invest. But the overall effect of this tax on corporate savings and investment depends on a number of other tax factors23 listed below.

(*i*) Statutory allowances and disallowances;

(*ii*) Statutory tax concessions, e.g., development rebate, depreciation allowance, etc;

(*iii*) Corporate income tax rates (proportionate or progressive);

(*iv*) Tax treatment of losses—carry-forward facilities or otherwise;

(*v*) The system of depreciation allowance;

(*vi*) Tax free reserves, e.g., contingency reserves for bad debts, anticipated losses, etc;

(*vii*) Tax-treatment of corporate retentions.

(*viii*) Taxation of personal capital gains mainly in case of corporate common stocks;

(*ix*) Tax treatment of inter-corporate dividends—whether subject to double taxation and

(*x*) Taxation of corporate capital gains.

To these may be added another important factor affecting the private investment, i.e., the *ability of firms to shift the tax-burden*. Let us now briefly discuss the important aspects of corporate income taxation in India in relation to private investment.

Company Tax and Incentive to Invest Corporate income tax reduces the expected rate of return on investment in proportion to tax paid. In India, the corporate income was being taxed before 1991–92 at rates varying between 45 per cent and 65 per cent, depending on (*i*) whether public investment was 'substantially' involved in the company, and (*ii*) whether profit exceeded certain limits. Most large scale companies were subject to taxation at the rate of 55 per cent, the effective rate being 51.75 per cent. Such a high rate of company income taxation reduced the after-tax profits rate by about 52 per cent, i.e., post-tax profit is reduced to less than half of the pre-tax profits. Such a heavy reduction in firm's earning cannot remain without reducing the incentive and also the ability and desire to invest.

Company Tax and Ability to Invest Most industrial corporations rely mainly on their internal resources for their expansion for such reasons as high cost of borrowed capital and intervention in management by new equity holders. The major sources of internal finance are *depreciation reserves* and *retained profits*. The magnitude of these sources of internal funds depends on depreciation allowance permissible under corporate income taxation and tax-treatment of retained earning.

22. Otto Ekstein, *op. cit.*, p. 84.

23. Gandhi, Ved, *Some Aspects of Indian Tax Structure*, (Vora & Co. Publishers Private Ltd., Bombay, 1970), p. 93.

In India, not much evidence is available to show the effect of tax-treatment of depreciation and retained earnings (taxable) on corporate saving and investment. Regarding the retained earning, the remarks by the NCAER24 are worth mentioning, "the fluctuations in retained profit (during fifties) were caused mainly by fluctuations in gross profit because tax provisions as a proportion of profit did not vary widely." It also mentions, "It is difficult to say definitely what would have been the effect of greater variation in tax rates. Most probably its impact on retained profits would have been no less than that of fluctuations in gross profits".

It may be finally mentioned that the Indian government had provided a number of incentives to private entrepreneurs. Given the incentives for investment, it may be said that corporate income taxation in India would have not affected the corporate saving and investment much because the effective tax burden on individuals as well as on corporations is considerably reduced. Some important incentives provided before 1990 for savings and investment are enumerated below.

(*a*) *Tax holidays* for new undertakings including industries, shops and hotels established after 31 March 1976, for a period of 5 years, on the income upto 7.5 per cent of the capital employed.

(*b*) *Special-tax holidays* for new industrial undertakings set up in backward areas after 31 March 1973, to the extent of 20 per cent of the total profit.

(*c*) *Liberal depreciation allowance* on building, furniture, plant and machinery. Rates for plant and machinery varied from 5 per cent to 100 per cent on the basis of seven categories thereof, depending on the useful life.

(*d*) *Investment allowance* in the form of depreciation at the rate of 25 per cent of the cost of new physical capital installed after 31 March 1976. Investment allowance is comparable to development rebate.

(*e*) *Allowance of all revenue expenditure* incurred on scientific research related to business, during three years immediately preceding the commencement of business.

(*f*) *Development allowance* for tea industry in new areas at 50 per cent of the actual cost of planting done after 31 March 1965 and at 30 per cent of actual cost of replantation between 1 April 1965 and 31 March 1970.

(*g*) *Concessional treatment of intercorporate dividends*—60 to 100 per cent of intercorporate dividend received by a domestic company were exempted from tax.

(*h*) Other incentives and concessions included:

(*i*) Five year wealth tax holidays for initial equity;

(*ii*) Wealth tax exemption for share holdings;

(*iii*) Export market development allowance;

(*iv*) Exemption of dividend and interest upto a certain limit;

(*v*) Tax concession for book publishing;

(*vi*) Carry forward and set-off of accumulated losses;

(*vii*) Amortization of prospecting and development expenditure for certain minerals.

24. *op. cit.* (fn. 20).

With such a wide range of incentives, Indian company income taxation might be supposed to have been least deterrent to saving and investment. However, the available data show that saving and investment by the corporate sector have not increased in a great measure. Many of the incentives have now been withdrawn during the fiscal reforms since 1990. The effects of the tax reforms of 1990 remain to be investigated.

SUMMARY

• Fiscal policy refers to the government plan of making discretionary changes in the pattern and level of its expenditure, taxation and borrowing with the purpose of achieving some economic objectives. In the context of annual plan of spending and raising revenue by the government, fiscal policy is called budgetary policy.

• In general, the major objectives of fiscal policy are (*i*) economic growth, (*ii*) employment promotion, (*iii*) economic stability, and (*iv*) economic justice, i.e., economic equity.

• Instruments of fiscal policy include (*i*) public expenditure, (*ii*) taxation, and (*iii*) public borrowing.

• Target variables of fiscal policy (*i*) private income, (*ii*) private consumption, (*iii*) private saving and investment, (*iv*) export and import, and (*v*) price level.

• The objectives of India's fiscal policy are, in general, the same as the general objectives of fiscal policy, including (*i*) increasing growth rate, (*ii*) providing employment to unemployed, (*iii*) removing poverty, and (*iv*) controlling inflation.

• The total government expenditure (including central and the state governments spending) aimed at achieving these objectives have increased at a very high rate. The percentage of the total government expenditure to GDP has increased from 9.1% in 1950–51 to 29.2 per cent in 2010-11.

• The main source of government revenue is taxation. Indian's taxation system involves (*i*) central taxes, and (*ii*) state taxes. Central taxes include direct taxes – personal income tax, corporate income tax, wealth tax – and indirect taxes – excise duty and import duty.

• Different kinds of taxes affect private business in different ways. In case of direct taxes, personal income tax reduces disposable income and, thereby, the demand for consumer goods, which affects business adversely. Corporate income tax reduces the rate of return on investment and incentives for investment.

• As regards the excise duty, it has a double effect on private business. First, it puts tax burden on the firms initially, though they add excise duty to the price of their products. Second, addition of excise duty to product price increases price which reduces consumer demand. All these effect are negative.

• Finally, public expenditure has positive effect and taxation has negative effect on private business. So the overall impact of fiscal policy on private business depends on the net effect public expenditure and taxation. It all depends on a number of factors.

REVIEW QUESTIONS

1. What is meant by fiscal policy? What were the main objectives of India's fiscal policy before 1990 reforms?
2. What are the main objectives of fiscal policy? Briefly describe the role of fiscal policy in the economic growth of an underdeveloped economy.

3. How does taxation affect private business activities? Is taxation always a deterrent to the growth of private business?
4. Describe the effects of taxation on (*a*) work-efforts of the people, (*b*) desire and capacity to save and invest.
5. How does corporate income taxation affect the incentive and capacity of the large corporations to save and invest?
6. Indirect taxes are more of a deterrent to private saving and investment than the direct taxes. Elucidate.
7. How do the indirect taxes affect private business? Explain with examples from India.

FURTHER READING

Chelliah, R.J., *Fiscal Policy in Underdeveloped Countries* (George Allen and Unwin Ltd., London, 1969).

Due, John F. and a.F. FRieDlanDeR, *Government Finance: Economics of Public Sector* (Richard D. Irwin, Inc., 1977), Ch. 9.

DwiveDi, D.n., *Macroeconomics: Theory and Policy* (McGraw-Hill, New Delhi, 3rd Edn. 2010), ch. 21.

DwiveDi, D.n., *Readings in Indian Public Finance* (Wiley Eastern Limited, New Delhi, 2nd Edn., 1994), Chs. 1, 2, 10 and 11.

ekstein otto, *Public Finance* (Prentice Hall, Inc., New Jersey, 1967), Chs. 1 and 6.

helleR, w.w., 'Fiscal Policies for Underdeveloped Countries', in *Readings on Taxation in Developing Countries* (ed.) by Richard M. Bird and Oliver Oldman (The Johns Hopkins Press, Baltimore, 1967).

hevemen, R.h., *The Economics of Public Sector* (John Wiley and Sons Inc., New York, 1970), Chs. 3, 5, and 7.

kalDoR, noCholas, 'The Role of Taxation in Economic Development', in *Fiscal Policy for Economic Growth in Latin America* (The Johns Hopkins Press, Baltimore, 1965).

musgRave, R.a., *Theory of Public Finance* (McGraw-Hill Book Co.), Asian Student Edition, 1959.

musgRave, R.a., and P. musgRave, *Public Finance In Theory and Practice* (McGraw-Hill Kogakusha Ltd., Tokyo, 1989), International Student Edition, Ch. 22.

PRest, a.R., *Public Finance in Theory and Practice* (Weidenfeld and Nicolson, London, 1975), Chs. 4, 16.

CHAPTER 41 Industrial Policy of India

CHAPTER OBJECTIVES

Industrial policy of country determines the nature, scope and prospects of industrial growth of a country. The objective of this chapter is to have a brief discussion on the following aspects of industrial policy of India.

- The reasons for formulation of industrial policy
- Industrial policies adopted during pre-1991 period, i.e., the period from 1948 to the economic reform year 1991
- Features of industrial policy in Post-1991 period
- Meaning and purpose of joint sector and
- The current industrial policy

41.1 INTRODUCTION

In preceding two chapters, we have discussed monetary and fiscal policies of India. Both these policies are intended to achieve some macroeconomic goals of the country. These policies do affect private business positively and negatively by changing the economic environment of the country. However, these policies do not serve the purpose of promoting, organizing and determining the desirable industrial structure and growth of industry.

Therefore, the government has to formulate a separate policy for organising, developing and managing industrial growth of the country. This policy is known as *industrial policy*. Industrial policy refers to the rules and regulatory laws formulated by the government to determine the ownership and pattern of growth of industries in the country. In this chapter, we present a brief description of different kinds of industrial policies adopted by the government since the year of Independence, their objectives and limitations.

Industrial policy of India has undergone a sea change during the post-Independence period – over a period of 65 years. In this chapter, we present a brief review of India's industrial policy in historical perspective just to show the process of formulation of industrial policy. It may be added here that a detailed historical review of India's industrial policy is not over objective here as it does not serve any useful purpose. From historical review point of view, the history of India's industrial policy can be divided two phases: (*i*) Pre-1991 phase of industrial policies, and (*ii*) Post-1991 phase of industrial policy. The year 1991

marks the year in which India was forced to make drastic changes in her economic policies including industrial policy. So we will present our review of India's industrial policy in the two phases. We begin our discussion by looking at 'why industrial policy is required'.

41.2 WHY INDUSTRIAL POLICY?

In a democratic country and mixed economy like India where the government has adopted economic planning as an instrument of guided industrialization for establishing a 'socialist pattern of society,' it becomes indispensable for the government to control and regulate private industries with a view to bringing them in tune with national goals. The regulatory functions of the government1 in a mixed economy like India, include: (*i*) determining the areas for private and public sector industries; (*ii*) laying down the policy guidelines and norms for controlling and regulating the private and public sector industries; (*iii*) assisting, promoting and directing the private entrepreneurship for rapid growth of industries, and (*iv*) protecting the small-scale industries from competing large scale industries for promoting self-employment.

The government carries out these regulatory functions through (*i*) legislative measures, and (*ii*) discretionary changes in its fiscal and monetary policies. The legislative control measures or the direct controls are formulated in the form of 'Industrial and Licensing Policy.'

41.3 INDUSTRIAL POLICIES OF PRE-1991 PERIOD

As noted above, industrial policy of India has undergone numerous changes and reformulations. Here, we point out the basics of industrial policy formulated and used during the period from 1948 to 1990.

Industrial Policy -1948 Industrial policy of 1948 was the first systematic attempt to formulate a comprehensive industrial policy of India. The basic purpose of this policy was to demarcate the areas of public and private sectors and bring out guidelines for the control and regulation of the industrial sector. The industrial policy of 1948 had divided industries into four broad categories, viz., (*i*) *defence and strategic industries* including arms and ammunitions, atomic energy and railway; (*ii*) *basic and key industries* including iron and steel, coal, aircraft manufacturing and ship-building; (*iii*) twenty other *important industries* including heavy chemicals, sugar, cement, paper, cotton and woollen textiles, machine tools, etc. and (*iv*) residual industries, including all other industries not included in the other three categories.

Industries of category (*i*) were put under the exclusive monopoly of the government. 'Basic and Key Industries' of categories (*ii*) were open for private sector participation but were specified as state-controlled industries. Industries of category (*iii*) being mainly in the private sector were put under general state control and regulation. 'Residual industries', i.e., category (*iv*) were left purely to the private sector but were subject to general control and regulation of the state. In addition, cottage and small-scale industries, though placed, by and large, in the private sector were to be developed on cooperative lines, and coordinated and integrated with the large scale industries. With these provisions, the Industrial Policy (1948) laid the foundations of the mixed economy in India.

Industrial Policy — 1956 The Industrial Policy (1948) was, revised and a new industrial policy was announced in 1956. The change in the industrial policy was necessitated by the changing conditions, such as acceptance of the national goal of establishing a 'socialist

1. See also, Redford, E.S., *Administration of National Economic Controls*, New York, 1952.

pattern of society' during the Second Plan and the need for (*i*) rapid industrialization of the economy through the expansion of public sector; (*ii*) rapid growth of capital goods and heavy industries; (*iii*) preventing growth of monopolistic tendency in the economy, and (*iv*) diversification and dispersal of industries towards the backward regions.

The main objective of the Industrial Policy Resolutions (1956) was to recategorise the industries and determine their ownership. Industries were regrouped under *three* broad categories: Schedule A, Schedule B and 'residual industries'.

The Schedule A of industries was formed by merging together the first two categories of the 1948-policy. Schedule A consisted of 17 major industries (e.g., arms and ammunitions, atomic energy, iron and steel, heavy machinery, railways, ship-building, aircraft manufacture, telephone and telegraph, generation and distribution of electricity, etc.). The development and promotion of these industries was made the exclusive responsibility of the state.

The industries of Schedule B consisted of 12 industries, viz., aluminium, ferro-alloys, machine tools, antibiotics, fertilizers, synthetic rubber, minerals, dye-stuffs and plastics, road and sea transport, etc. These industries were to be progressively state-owned, with the provision that new units would be set up by the state and that private enterprise would be allowed to develop and assist the state in the promotion and growth of the industries of this category. The third category of industries consisted of the 'residual industries' of the 1948-policy. The ownership, management and promotion of industries in this category was left to the initiative and entrepreneurship of the private sector. It was, however, open to the state to enter and undertake production in any of these industries in the national interest.

Another important objective of the Industrial Policy Resolutions (1956) to promote cottage and small-scale industries in view of the ever growing need for (*i*) creating job opportunities for the unemployed; (*ii*) preventing further industrial and economic concentration and helping economic decentralization; (*iii*) ensuring an equitable distribution of national income and (*iv*) mobilizing and utilizing resources—local capital and skills available in small measures.

The Industrial Policy (1956) remained effective until 1973, although its rules governing the licensing system were modified several times in view of changing conditions and to meet the emerging requirements.[2] The 1956-policy and licensing system were, however, found by R.K. Hazari[3] to be lacking in many respects. His enquiry revealed that the licensing policy had failed to prevent economic concentration. It had instead encouraged economic concentration. The share of four big business houses in the total approved investment had increased from 22.4 per cent in 1959–60 to 24.6 per cent in 1965–66.

In view of the shortcomings of Industrial Policy (1956), the Government appointed an Industrial Licensing Policy Inquiry Committee under the Chairmanship of Subimal Dutt.

The major findings of the Dutt-Committee: First, the large business houses had secured undue favour in licensing against other applicants. ***Second,*** some big industrial houses, most notably the Birlas, managed to secure unduly large number of licences by making a large number of applications for the same product through the firms controlled by them, while they implemented only a few. ***Third,*** the Committee disclosed that 31 per cent of

2. For details see *Report of the Study Team of the Directorate General of Technical Development,* Govt. of India, 1967 and Report of the Study Team on Economic Administration, Administrative Reform Commission, Govt. of India, 1967, pp. 14–16. For a brief description, see also, P.K. Ghosh, *Government and Industry*, (Orion Publication, New Delhi, 1977), Ch. 2.

3. *Industrial Planning and Licensing Policy*, Final Report, Govt. of India, (Planning Commission, New Delhi, 1967).

the licences issued were never implemented and, on the contrary, many licences had built up production capacity in excess of licensed capacity and had secured regularization of capacity later. ***Fourth,*** the Committee did not find any clear evidence of compliance with the Plan priorities in respect of regional dispersal of industries, import substitution and development of small industries. In many cases there was overlicensing from 50 to 100 per cent. ***Finally,*** the committee had also found that the share of 20 large business houses in the finance provided by the public financial institutions was unduly large.

Industrial Policy—1970 Following some of the recommendations of the Dutt-Committee[4], the Government announced in February 1970, a revised industrial policy and licensing procedure. The main objectives of the Industrial Policy (1970) were: (*i*) widening the scope for expansion of the public sector; (*ii*) restricting the growth of large industrial houses and proliferation of foreign companies, and (*iii*) accelerating the growth of industries in small and medium scale sectors. Different kinds of licencing system was adopted for the industries of different categories.

Dutt Committee had also recommended setting up a ***joint sector***. The 'joint sector' concept implies participation and cooperation of public and private sectors in the promotion and development of priority industries. In this sector, private managerial skill and public funds would work together for the promotion of industries. In effect, however, government participation appeared in the form of (*i*) conversion of financial assistance (loans and debentures) of IDBI and IFCI, etc., given to the private sector into equity capital, and (*ii*) representation of the public financial institutions in the management of companies getting assistance from them.

Industry Policy—1973 Industrial licensing policy (1970) created a good deal of apprehension among the industrialists and also an atmosphere of uncertainty, particularly in the large scale sector. It caused a slackening in the overall investments in the industrial sector. In view of this experience and the requirements of the Fifth Plan which emphasized the need for additional job creation and 'removal of poverty', Government announced its yet another revised industrial policy on February 2, 1973.

The Industrial Policy (1973) was not essentially a new policy but a mere reiteration of Industrial Policy (1956). The 1973-policy created a new sector, the **joint sector** (discussed later) but it remained an ambiguous entity. Furthermore, the new policy liberalized the licensing which affected the interests of small and medium scale units. However, despite liberalization of licensing, the 1973-policy failed to create confidence in the private sector. It was for these reasons, the Government had to revise that the licensing policy and liberalize it further in 1975.

Industrial Policy—1977 When Janata Party came to power in 1977, it reviewed the earlier industrial policies. It concluded that despite its certain merits and desirable elements, the Industrial Policy (1956) along with its subsequent amendments had many shortcomings and created certain distortions in the economy.5 With a view to removing the distortions caused by the earlier industrial policy and to achieve the social objectives, the Janata Government announced its New Industrial Policy in December 1977.

The *main thrust* of the 1977-industrial policy was an efficient *promotion of cottage and small-scale industries* widely dispersed in rural areas and small towns. It emphasized

4. Government did not accept all the recommendations of Dutt-Committee.

5. Distortion included (*a*) widening rural-urban disparity, and (*b*) stagnation of investment.

that what could be produced by cottage and small-scale industries should be reserved for them. Thus, the new industrial policy had shifted the emphasis from large-scale industries to cottage and small-scale industries. The area reserved for the large scale industries were (*a*) *basic industries* essential for the promotion of infrastructure and for development of small-scale and village industries; (*b*) *capital goods industries*; (*c*) *high technology industries* which required large scale production and met the development requirements of small-scale industries and agriculture and (*d*) industries other than those reserved for small-scale sector.

The Industrial Policy (1977) was however criticized on the following grounds.

First, the new policy was regarded as a mere extension of the 1956-Policy with some not-so-serious modifications. ***Second,*** it was also said that the emphasis on the growth of small-scale and cottage industries was merely an eye-wash. For, out of about 2400 items produced by the industries of this sector,6 only 504 items were reserved for the small sector which were not very different from the 180 items reserved by the 1956 policy.

Third, although development of small-scale and cottage industries was the main thrust of the new industrial policy, the relative emphasis on the sector had, in effect, gone down. Only about 2 per cent of the total plan outlay had been earmarked for the development of small-scale and cottage industries compared to 3.8 per cent in the Second and Third Plans.

Industrial Policy—1980 When the Congress Party returned to power in 1980, it liberalized the industrial policy with a view to accelerating industrial growth of the country. The 1980-policy stressed the growth of infrastructure industries like power, transport, communication and finance. The limits of investment in small and ancillary units were substantially enhanced to help their modernization. The 1980-policy emphasized the need for making public sector units efficient and economically viable. Restrictions on the expansion of private sector industries were relaxed and conditions were liberalized. Large private industries were automatically permitted to expand capacity upto 5% per annum subject to 25% expansion in five years. Licencing of private industries was simplified and streamlined with the objective of minimizing bureaucratic delays and preventing harassment of licence seekers. It also provided for permitting the use of advanced technology where necessary to modernize industrial units to help in increasing their international competitiveness.

The industrial policy (1980) remained in operation till July 1991. This policy was, in fact, the first step towards the liberalization of the industrial policy. It had no doubt proved helpful in increasing industrial investment but private sector industries remained shackled by licensing controls and regulations.

41.4 THE NEW POLICY (1991)

Industrial sector of India had made considerable progress; industrial structure was fairly diversified; and India had achieved near self-sufficiency in consumer goods industries. However, growth rate was not enough to absorb increasing manpower supply and alleviate poverty. There was a need for accelerating industrial growth rate that could help achieve these goals. But no major change in industrial policy was made. However, a new industrial policy was announced in July 1991 which was necessitated by the ***foreign exchange crisis*** of 1990. The country had reached close to financial bankruptcy. There was no alternative to borrowing from the IMF and the World Bank (WB) for tiding over the crisis. The IMF

6. According to the Small-scale Industries Census (1973–74).

and the WB agreed to provide financial help on the condition that India would make drastic changes and structural reforms in her economic policies — industrial policy was one of them. The government had no option but to meet the IMF and the WB conditionality.

In compliance with one of the conditionalities, the government announced the New Industrial Policy (NIP) on July 24, 1991, along with other economic reforms. The following were the main features of the NIP.

1. ***De-reservation of Industries of the Public Sector.*** The NIP had dereserved 9 out of 17 industries reserved for the public sector since 1956. The eight industries reserved for the public sector include those where security and strategic considerations were predominant. The industries dereserved include iron and steel, electricity, air transport, ship-building, heavy machinery industries like heavy electrical plants and telecommunication cables and instruments.

2. ***Privatization of the Government Share in Public Sector Units.*** The government resolved to offer 49 per cent of the government shareholding in 31 public sector undertakings (PSUs) to mutual funds, financial institutions, the general public and workers. A beginning had already been made. A part of the equities of the selected PSUs had been placed with mutual funds. The government put on sale the shares of seven PSUs (including SAIL).

3. ***Abolition of Industrial Licensing.*** The industrial licensing regime which was the main source of corruption and bureaucratic delays and also of most of the industrial ills, had been abolished for all industries irrespective of the level of investment, except for the industries involving national security and strategic considerations, social safety, environmental issues and industrial hazards.

4. ***Abolition of Phased Manufacturing Programmes.*** The New Industrial Policy abolished the phased manufacturing programme in force in a number of engineering and electronics industries, intended to force indigenization in manufacturing. The abolition of phased manufacturing programme removed a major irritant that a large number of firms had felt, i.e., discretionary power and government's interference in business decisions.

5. ***Removal of Mandatory Convertibility Clause.*** The NIP removed the convertibility clause from the lending conditions of the financial institutions. Under this clause, the financial institutions financing an industrial project had the option of converting their loans into equity, if they wanted to do so. This clause was mandatory. Although this option was not often exercised, it was considered to be a hanging threat of takeover by the financial institutions. This clause was no longer mandatory.

6. ***Removal of Investment Control on Large Business Houses.*** The industrial units covered under MRTP Act (1969), i.e., those having assets worth `100 crore, were required to obtain a separate licence for additional investment and capacity expansion. Under the NIP resolutions, such firms would not be required to obtain prior approval of the government for investment in delicensed industries.

7. ***Liberalization of Foreign Investment.*** The earlier foreign investment policy was extremely restrictive. In the NIP, automatic permission was available to foreign investors upto 51% equity shares in the high priority industries. This facility was made available to the firms able to finance their capital imports through their foreign equity.

Although some changes have been made and industrial policy continues to be modified from time to time, the main thrust of the policy continues to remain the same.

Evaluation The NIP marked the end of the old policy regime and the beginning of a new era in the government's approach towards the control and management of the industrial sector. The departure from the old industrial policy was of a drastic nature. Some of the changes were as follows:

1. The earlier policy resolutions had put a greater emphasis on such policy goals as strengthening the public sector, building public sector enterprises, preventing economic concentration, reducing economic disparities, expanding gainful employment and self-reliance. In the NIP, the top priority was accorded to industrial efficiency, growth and international competitiveness.
2. The earlier industrial policies had made the public sector the main instrument of industrial growth. In the new policy, the private sector was made the main instrument for future industrial growth of the country.
3. Another striking feature of the NIP was liberalization of foreign investment. Foreign investment was earlier allowed on selective basis with only 40% equity. The new policy invited direct foreign investment to 34 industries with 51% equity. This had been done in view of the need for globalization of the Indian industries and the advantage of transfer of new technology.

The new industrial policy had made Indian industries more efficient and internationally competitive. It created a healthy environment for industrial growth, free from bureaucratic shackles. It encouraged fresh private investment in the new areas opened to the private sector.

41.5 JOINT SECTOR

In addition to the private and public sectors, the joint sector which is of relatively recent origin has appeared in India's mixed economy. In this section, we have briefly described the concept, growth and functioning of the joint sector.

41.5.1 Concept of Joint Sector

The concept of joint sector implies partnership between the state and private entrepreneurs in industrial ventures. The joint sector refers to those industries which involve 'co-ownership and co-management' of industrial units by the government and private entrepreneurs. Although the Dutt-Committee had suggested for the first time elaborate measures to enlarge the joint sector, it did not provide an appropriate definition of the joint sector. J.R.D. Tata in his *Memorandum on Industrial Growth* to the Government tried to provide a 'minimum agreed definition' of the joint sector as "A joint sector enterprise is intended to be a form of partnership between the private sector and the government in which State participation in capital will be not less than 26 per cent, day-to-day management will normally be in the hands of the private sector partner, and control and supervision will be exercised by a board of directors on which Government is adequately represented."

The Tata concept of the joint sector has a bias in favour of the private sector as a partner. It visualizes state-partnership in the form of only a minor capital-supplier. Although 'adequate representation' is not expressed in clear terms, it may mean state representations on the board of directors in proportion to capital contribution. In that case, state representation would become ineffective if the state contributes less than 50 per cent of capital. Thus, Tata visualized the state to be a kind of 'sleeping partner' in the joint sector. A correct approach would be to

visualize and define joint sector keeping in view the objectives which this sector is supposed to fulfil. The basic idea of bringing together the state and private entrepreneurs in industrial ventures is to work towards growth and economic justice, to create conditions against growth of monopolies and economic concentration. The joint sector should, therefore, be treated as a partnership between the forces of growth and social justice. So long as this objective is achieved, the controversy over 'which sector should dominate' is meaningless. Moreover, with a view to removing apprehension and suspicions, which is not unlikely under the conditions prevailing in our economy, equal participation of both sectors in financial contribution, control and management and decision-making would be helpful and should be acceptable.

41.5.2 Rationale of the Joint Sector

In India's mixed economy model, the country has both private and public sectors working in separate fields specified under the Industrial Policy. Why is there then a need for a joint sector and what is the rationale of the joint sector?

The rationale of the joint sector lies in the following facts: (*i*) there is an urgent need for ensuring social and economic justice; (*ii*) curbing monopolistic tendency and economic concentration without adversely affecting growth is our social goal, and (*iii*) both private and public sectors have their strength and weakness and their positive qualities may be combined to further social gains. We have already noted that private sector functioning under a free-enterprise system may create conditions for monopolistic tendency and thereby economic concentration. The growth of monopolies or large industrial houses may be prevented through legal measures. But it may affect investment and growth adversely. This possibility may be prevented if private business is brought under social control through state partnership in private business.

Besides, it is generally acknowledged that private sector possesses better managerial capabilities, initiative, enterprise and dynamism, and hence is more efficient. These are positive qualities of the private sector. But profit-mongering, desire to accumulate wealth through anti-social business ethics and practices, exploitation of labour and consumers, are some of the negative aspects of private sector business. On the other hand, the states possess enormous public funds but, as a business entity, it lacks managerial ability and perspective. More often than not public investment turns out to be wasteful expenditure. The joint sector provides areas where managerial skill of the private sector and public sector funds may be utilized for promoting economic growth and social welfare. The state as partner will prevent profiteering and exploitation of consumers and private sector as a partner will prevent inefficiency.

Furthermore, as the Dutt-Committee had pointed out, large amount of public funds provided by public financial institutions are being used by the private sector for private gains. There is no reason why the government should not become a partner in the industrial undertakings using public funds by converting public loans into equity, and direct the activities of such industrial units towards social gains.

41.5.3 Advantges from the Joint Sector

The enterprises which have been created or brought under joint control and management of private and public sectors during the post-Independence period can be supposed to have formed the 'nucleus' of the existing joint sector. The joint sector is also reported to have worked satisfactorily. The smooth functioning of the joint sector has dispelled some of the earlier

misgivings and apprehensions. Some of the major apprehensions were: (*i*) that dominance of the private sector in the management and control might defeat the objective of social control on private business; (*ii*) that joint sector might result in strengthening private capitalism with state assistance; (*iii*) that unnecessary government interference with management might introduce bureaucracy and inefficiency in business management, and (*iv*) that mismanaged industrial concerns might try to get refuge in the joint sector. But these doubts now seem to have been disapproved.

The smooth functioning of the joint sector is attributed to a healthy, cooperative environment and mutual trust between the partners under which they are functioning.[7] The other factors[8] which are said to have contributed to the successful functioning of the joint sector are: (*i*) the purpose of collaboration, other than social control of business has been to curb the market power of large industrial concerns; (*ii*) the management of the joint sector undertakings confines exclusively in the hands of the private sector; (*iii*) nomination of leading businessmen as the chairmen of the boards, and (*iv*) the appointment of managing and technical directors is the prerogative of the private sector. The future course of joint sector operation will depend on, by and large, to what extent the government tries to implement the basic objective of the joint sector, i.e., social control of private business and curbing the market power of monopoly and oligopoly firms.

41.5.4 The Present Policy

The present policy of the government in respect of the joint sector is based on the recommendations of the Dutt-Committee (1969). The major recommendations of the Committee regarding joint sector are following:

(*a*) The large industrial houses which have received substantial funds from financial institutions should be brought under the joint sector by converting loans into equity.

(*b*) The existing industrial units which have been set up jointly with private and public investments and in which the state is to play an active role in control and management should be treated as joint sector concerns.

(*c*) The large industrial units of Schedules B and C, justifying their size on account of technical and economic considerations should be in the joint sector. Such units should be treated as public sector units by virtue of large share of public investment in these units, and private sector should be allowed to participate in the development of the joint sector as a small partner.

The government accepted the concept of joint sector in its Industrial Licensing Policy of 1970 and reiterated its stand in its Industrial policy of 1973. The present policy of the government in respect of joint sector may be summarized as follows: (*i*) the industries established in the joint sector will be required to fulfil the social and economic objectives; (*ii*) the government would allow public and private sector participation in the areas so far excluded from the list of industries; (*iii*) while the joint sector would play a promotional role, it would remain under effective control and direction of the government and (*iv*) the government would also determine the share of various partners (viz., Centre, State, private entrepreneurs and foreign companies) in the equity capital.

7. Tata Memorandum, *Economic Times,* 29 August 1972, Bombay.

8. Ghosh, P.K. *op. cit.*, p. 84.

SUMMARY

• Industrial policy refers to the rules and regulations formulated by the government to determine the ownership and to control and regulate the growth of industrial sector of the country with some specific purpose.

• The need for industrial policy arises because ownership pattern and nature of industries and growth pattern of industrial sector under free market system may not be in accordance with the economic and social needs of the country.

• India’s industrial policy has a long history since 1948. The formulation and reformulation of industrial policy can be put under two phases: (*i*) pre-1991 industrial policies, and (*ii*) post-1991 industrial policy.

• In the first phase, the first industrial policy of India was formulated in 1948. Industrial policy was later reformulated and reformed in subsequent years due to changing economic needs and economic conditions of the country. A significant reformulation of industrial policy took place in 1956. Industrial policy was reformulated in subsequent years – 1970, 1973, 1977, 1980. The 1980-industral policy was in operation till 1991.

• India’s industrial policy was reformed drastically in 1991 for two reasons. *One*, the industrial growth and structure of the country had changed significantly and it needed some reformulation. *Two,* India had faced in 1990 an unprecedented *foreign exchanges crisis* was forced to borrow funds from the World Bank and the IMF. These provided financial help under the conditions that India would make some major economic reforms in her economic policy, including industrial policy. The current industrial policy is the same as one formulated in 1991.

• Another important change in industrial policy was made over time was the formation of a joint sector with joint ownership and partnership of the private sector and government for the establishment and development a section of essential industries. The objective of setting up of a joint sector was to use financial resources of the government and management skill and technology of the private sector.

REVIEW QUESTIONS

1. What is the significance of industrial policy in a mixed economy? What purpose does it serve?
2. What are the main features of the New Industrial Policy? How is it different from the 1956 Industrial Policy?
3. How do you think the licensing policy of India had affected growth of industries in India?
4. How do you think the New Industrial Policy (1991) is conducive to industrial growth of India? What are its shortcomings?
5. What is the meaning and purpose of the joint sector? How do you think the promotion of joint sector would achieve a greater industrial growth and social justice?

FURTHER READING

Ghosh, P.K., *Government and Industry* (Orion Publication, New Delhi, 1970).

hazari, r.K., *Indian Planning and Licencing Policy*, Final Report, 1961.

Kumar, arun, *New Economic Policies*: *An Assessment*, *EPW*, Vol. XXVIII, No. 50, December 11,

1993, *Indsutrial Policy in India*, (Birla Economic Research Foundation, New Delhi, 1995).
ParanjaPe, h.K., *New Indsutrial Policy: A Capitalist Manifesto, EPW,* Vol. XXVI, No. 43, October 26, 1991.
ranGneKar, D.K., "Industrial Policy," *Economic Times*, Annual Number, 1975.
Report of Industrial Licensing Policy, Inquiry Committee, Main Report, July 1969.

CHAPTER

42 Some Economic Legislations

CHAPTER OBJECTIVES

The objective of this chapter is to discuss some important economic legislations which have been formed to control the emergence of socio-economic evils that arise due to free functioning of the private sector entrepreneurs and distort economic structure of the country. The legislations discussed here include the following three legislations.

• Monopolistic and Restrictive Trade Practices (MRTP) Act
• Law governing conflict between trade unions and corporations and
• Industrial Relations and Industrial Dispute Act

42.1 INTRODUCTION

We have discussed in three preceding chapters, the major policy instruments—fiscal, monetary and industrial policies—which the government uses to control and regulate the corporate sector to bring its activities in harmony with prime economic goals of a sustainable economic growth, fuller employment, stability and economic equity. In addition, the government uses a host of legislative measures to minimize socio-economic problems that arise—intended or unintended—in the process of profit–motivated working of the private enterprises. Some widely known economic and social problems and public-private conflicts that arise out of functioning of the private business are following.

(*i*) Growth of monopolies and monopolistic trade practices leading to underutilization of resources and higher prices;

(*ii*) Exploitation of labour leading to conflict between the management and labour unions, loss of production and inconvenience to the society;

(*iii*) Unfair and unethical trade practices, e.g., cheating consumers in respect of promised quality, quantity and price; unauthorized imitation of popular brands of products; production and supply of prohibited goods and adulterating foodstuffs;

(*iv*) Using money power to influence politics and bureaucracy for personal favours against the society resulting into widespread corruption of politics and administration, and

(*v*) Causing environmental pollution1, including that of rivers and underground water, which threaten the survival of both human and animal life and destruction of environment.

There are **two major sources** of such socio-economic problems. ***One,*** there is inherent conflict between private and social interest. ***Two,*** private enterprises have certain responsibilities (discussed ahead in Chapter 43) towards the society which they ignore altogether. The attitude of indifference of the business community towards its social responsibilities aggravates the conflict between private and social gains. It is the responsibility of the government to control and regulate the private business to minimize the loss to the society through a system of preventive measures and punishment. To this end, the government enacts certain economic legislations.

In this chapter, we will discuss some economic legislations which have been enacted by the Government of India to deal with socio-economic problems arising in free enterprise economics. Our purpose here is not to deal extensively with the legal provisions of economic legislations but to make the students of business management aware of the causes of socio-economic problems and implications of economic legislations in business decisions.

42.2 MONOPOLIES AND RESTRICTIVE TRADE PRACTICES (MRTP) ACT

42.2.1 The Background

Growth of monopolies has been a widespread phenomenon in the industrial countries with capitalistic economic systems. Almost all industrialized, free enterprise nations have promulgated laws to control and regulate prices and production of existing monopolies and prevent further growth of monopolies. The emergence and existence of monopolies in India can be attributed to the introduction of the British industrial culture by the Britishers during their colonial rule in this country. Independent India inherited an economic system and industrial structure characterized by monopolies and economic concentration. Although the Directive Principles of the Indian Constitution laid down provisions for reducing concentration of economic power—wealth and means of production—no action was taken until 1970 and, as a result, monopolies and concentration of economic power continued to grow during the post-Independence period.

The need for controlling monopolies arises mainly because growth of private monopolies (except where it is permissible on the grounds of efficiency) and the concentration of economic power in the hands of a few individuals or firms are detrimental to the economic welfare of the society at large. Monopolies and economic concentration limit the promotion of society's economic welfare in following ways.

(*i*) Profit maximizing monopolies limit production much below their efficient level and, thereby, restrict the supply of goods and services much below the potential level. This limits the availability of goods and services and the prospect of economic welfare.

(*ii*) Private monopolies set prices of their products at a level much higher than their competitive level and, thereby, reduce the consumer surplus and economic welfare.

(*iii*) By limiting the output, monopolies limit the scope of employment and income generation much below its potential level. This is detrimental to economic growth and promotion of economic welfare.

1. According to World Health Organisation (WHO) Report (2014), about half of the urban population is exposed to air pollution. In general, environment pollution is attributed to industrial expansion all over the world.

(*iv*) Monopolies prevent the growth of competition in the industrial sector and, thereby, the efficiency in production that can be achieved in the environment of competition.

(*v*) By using their economic power, monopolies influence the government's decisions on policy matters pertaining to industrial growth, equity and social justice.

42.2.2 Monopoly Inquiry Commission (1965)

In view of the disadvantages of monopolies, the Government of India appointed a Monopoly Inquiry Commission in 1965 under the Chairmanship of Shri K.C. Dasgupta, to perform the following functions.

(*i*) To inquire into the extent and effects of economic power in private hands and the prevalence of monopolistic and restrictive trade practices in economic activities other than agriculture, and

(*ii*) To suggest, in the light of its findings, the legislative and other measures to control monopolies and to protect essential public interest.

Approach and Findings of Monopoly Inquiry Commission (1965)

In its approach to assess the concentration of market power, the Monopoly Inquiry Commission examined (*i*) product-wise concentration, and (*ii*) country-wide concentration (i.e., virtually the industrial house-wise concentration). In assessing the product wise concentration, the Monopoly Commission adopted the following criteria of concentration.

Degree of concentration	*Share of top three firms*
1. High	75% or more
2. Medium	60% to 74.9%
3. Low	50% to 59.9%
4. Nil	Less than 50%

Product-wise concentration. The commission investigated the concentration of 100 products and concluded as follows:

(*i*) *High concentration* in 65 products including infant milkfood, kerosene, oil, petroleum, pump-stove, fluorescent tubes, dry batteries, domestic refrigerators, sewing machines, typewriters, tooth pastes, footwear, cigarettes, motor vehicles, etc.

(*ii*) *Medium concentration* in 10 products including biscuits, electric fans and lamps, radio receivers, cement, bicycles, etc.

(*iii*) *Low concentration* in eight products, including woollen fabrics, worsted knitting yarn, hurricane lantern, paper and pencils, stovewares, pipes, etc.

(*iv*) *Nil concentration* in 17 products including tea, coffee, sugar, vegetables oils, coal, cotton textiles, sanitary wares, etc.

Company- wide concentration. The Commission's findings revealed that the *top five* industrial houses, viz., Birlas (151), Bangurs (81), Surajmal Nagarmal (76), Bird Heilders (64) and Tatas (53)[2], together controlled 425 companies. In all, top 75 business houses controlled 1536 companies with 44.1% joint share in the total assets of all the non-government and non-banking companies in 1963–64.

2. The numbers given in the parentheses indicate the number of companies controlled by business houses.

If one looks at the present status of the monopolies and concentration of economic power, things have not changed materially. In mid-1990s, for instance, five top-most business houses (Tatas, Birlas, Reliance, Thapars and Singhanias) had total assets of `25,000 crore which accounted for 60 per cent of the total assets of `41,610 crore of top 20 business houses and companies[3].

Recommendations of the Monopoly Commission Based on its findings, the Monopoly Commission made the following recommendations:

(*i*) setting up a permanent body like Monopolies and Restrictive Trade Practices Commission;

(*ii*) liberalization of industrial licencing policy to facilitate the entry of small and medium industrialists to monopolized industries;

(*iii*) issuing import licences to all actual users of imported inputs rather than issuing it to a few importers; and

(*iv*) countervailing actions by the public sector to restrain monopolistic practices expect where monopolies contribute to growth and are socially and economically desirable.

42.2.3 Monopolies and Restrictive Trade Practices Act (1969)

Following the recommendations of the Monopoly Commission, the Government of India enacted the Monopolies and Restrictive Trade Practices Act, popularly known as the MRTP Act, in 1969. This Act was made effective from June 1, 1970. This Act is applicable to entire India, except Jammu and Kashmir. The MRTP Act (1969) has three main objectives: (*i*) to control and regulate the concentration of economic power, (*ii*) to control monopolies and restrictive trade practices, and (*iii*) to prohibit restrictive trade practices unless it is in public interest. To these ends, the MRTP Act defined the *monopolistic or dominant firms, monopolistic trade practices*, and *restrictive trade practices* in as great details as necessary for legal action. Here, we present the definitions in brief.

Monopolies or Dominant Firms The MRTP Act (1969) provides that the companies of the following categories be treated as monopolies or dominant firms: (*i*) companies with assets of `20 crore, ( *ii*) inter-connected companies with total assets of `20 crore, (*iii*) companies whose own supply accounts for one-third of the total supply, and (*iv*) interconnected companies working together as dominant companies with total assets of `1 crore or more.

Monopolistic Trade Practices The MRTP Act makes a distinction between *monopolistic trade practices* and *restrictive trade practices.* The definition of *monopolistic trade practices* given by the MRTP Act can be summarized as follows. *Monopolistic practices* are the practices that are adopted by a firm or a group of firms by virtue of their dominance that are detrimental to public interest. *Monopolistic practices* includes (*i*) maintaining price at an unreasonably high level, limiting production and controlling the supply to maintain unreasonably high level of price, (*ii*) any policy, practice or act that reduces both the current and potential competition, and (*iii*) limiting capital investment and technical development and allowing the quality of product to deteriorate.

3. Including 15 others, viz., Larsen & Toubro, Modis, Bajaj, Mafatlals, Chidambaram, Hindustan Lever, United Breweries, TVS Iyenger, ITC, Sri Ram, ACC, Oswal Agro, Mahindra & Mahindra, Essar, and Kirloskars (placed in the order of their assests).

Restrictive Trade Practices *Restrictive trade practices* are, in contrast, related to the practices or acts that prevent, restrict or distort competition. Restrictive trade practices refer generally to the practices that are adopted by a small group of dominant firms with 'agreement', arrangement or understanding to prevent the growth of competition. This system is known as 'cartelization' of the industry. The restrictive trade practices by way of agreements, arrangements and understanding, listed by the MRTP Act are many and the list is long. These include mainly (*i*) restricting the sale of goods to or purchase of goods from specified persons and in specific manner, (*ii*) requiring the purchase of some goods to purchase some other goods, i.e., tie-in-sale, (*iii*) requiring the dealers and the wholesalers not to deal in the product of an y other company, (*iv*) restricting the areas for sales, and (*v*) restricting the dealers in granting discounts, rebates and concessions and tying the sale to specified service sellers.

The *other forms of restrictive trade practices* include (*i*) collusion among firms or formation of cartels, (*ii*) price discrimination between different groups of buyers, (*iii*) predatory pricing to eliminate competitors, (*iv*) tie-up sales of goods in high and low demand, (*v*) forcing full-line purchase, and (*vi*) area restriction.

42.2.4 Amendments in MRTP Act

The MRTP Act (1969) was amended significantly in 1982, 1984, 1985 and 1991 to bring it in tune with the changing conditions of the industrial sector, though the basic structure of the Act remains the same.

In the **1982-amendment,** the definition of dominant firm was changed from one having a market share of one-third to one having a market share of one-fourth of the total. In addition, the share of a firm in the total installed capacity for the production of a commodity was also made a consideration in deciding on a dominant firm. Also, export part of the total sale was excluded for the purpose of calculating the market dominance.

In the **1984-amendment,** the major amendments included (*i*) unfair trade practices were brought under purview of the Act, (*ii*) certain provisions of Companies Act relating to acquisition and transfer of shares were transferred to the MRTP Act, and (*iii*) the central government was given powers to delink the interconnected companies in certain cases.

In the **1985-amendment,** the asset-holding limit for the purpose of determining the dominant status of a firm was raised from `20 crore tò100 crore.

The **1991-amendments** in the MRTP Act were made through the Industrial Policy change announced on July 24, 1991. This amendent made drastic changes in the Act. The major amendments are summarized below.

(*i*) The provision regarding the concentration of economic power was repealed, except where it was detrimental to common cause. This shifted the emphasis of the Act from monopolies as such to their restrictive and unfair trade practices.

(*ii*) The provision of seeking government permission for (*a*) capacity expansion, (*b*) establishment of new undertakings, (*c*) merger, amalgamation and take over, and (*d*) appointment of certain directors was repealed too.

The MRTP Act is implemented by the Monopolies and Restrictive Trade Practices Commission (MRTP Commission) set up on the recommendation of the Monopoly Inquiry Commission (1965) for the purpose. The MRTP Commission is empowered to inquire

into monopolistic, restrictive and unfair trade practices of the companies on (*a*) reference made by the Centre and the state governments, (*b*) on the application from the Director-General of Investigation, and (*c*) on complaints from a trade or consumer association with a membership of 25 or more. The MRTP Commission enjoys the power of a Civil Court. It has the power to summon a company or business house, inquire into its business activities and practices, ask it to submit its business accounts and to produce evidence contrary to the complaint against it. It has the power to dismiss a case against a company and initiate action in accordance with the provisions of the Act.

42.3 INDUSTRIAL RELATIONS AND INDUSTRIAL DISPUTES ACT

One major problem in India is disharmonious industrial relations, i.e., lack of healthy relationship between the industrialists and the industrial workers, leading to industrial unrest and loss of man-hours and output. Industrial relations encompass a wide area extending to payment of reasonable wages and salaries, services conditions including payment of bonus, permissible leaves, working hours, working conditions, workers' safety, social security, non-pecuniary benefits. A healthy and harmonious relationship is indispensable for continuity in production, uninterrupted supply of goods and services to the people, continuity in employment status, a higher level of employment, promotion of work-ethos and self-cultivated discipline. Unfortunately, there seems to be a wide gap between what is offered by the industrial undertakings and what labour expects and demands. Therefore, industrial relations in most industrial organizations are not harmonious.

42.3.1 Industrial Relations in India

The industrial relations scenario in India has been very ominous and disconcerting. Industrial unrest and disputes are all pervasive, be it industry-wise, region-wise or sector-wise. The average number of strikes, lock-outs and 'gheraos' increased from 1100 in 1961 to 2700 in 1995. The number of workers involved increased from 500 thousand in 1961 to 1600 thousand in 1992. The recent scenario of the industrial relations in India is revealed by the number of strikes and lock-outs and man-days lost presented in Table 42.1.

Table 42.1 *Industrial Relations Scenario in India: Number of Strikes, Lock-outs and Man-days Lost*

Year	*Strikes*		*Lock-outs*		*Total*	
	Number	*Man-days lost**	*Number*	*Man-days lost**	*Number*	*Man-days lost**
1996	763	7.8	403	12.5	1166	20.3
1997	793	6.3	512	10.7	1305	17.0
1998	665	9.4	432	12.7	1097	22.1
1999	540	10.6	387	16.2	927	26.8
2000	426	12.0	345	16.8	771	28.8
2001						

372
5.6
302
18.2
674
23.8
2002
295
9.7
284
16.9
579
26.6
2003
255
3.2
297
27.1
552
30.3
(Contd...)

2004
236
4.8
241
19.0
477
23.9
2005
227
10.8
229
18.9
456
29.7
2006
243
–
187
–
–
20.3
2007
210
–
179
–
–
27.2
2008
240
–
181
–
–
17.4
*In Million.

Source: Compiled from defferent issues of *Economic Surveys*, Ministry of Finance, Government of India.

As Table 42.1 shows, the number of strikes declined almost continuously after 1996, but the man-days lost first increased and then declined. On the lock-out front, while the number of lock-outs decreased, there was increase in the loss of man-days. The scenario that emerges on the whole is that there was some improvement in the industrial relations after 1996. However, the number of man-days lost increased intermitantly till 2005, though it declined later marginally. This is a disconcerting trend.

42.3.2 Causes of Industrial Disputes

The causes of industrial disputes and unrest are various and complex. However, the main cause of industrial disputes is the confrontation between the management and the labour unions. The growth of labour union power and its role will be discussed later separately. Let us first look at the factors that create the situation of confrontation between the management and the labour unions. The major causes of confrontation can be classified as (*i*) economic factors, (*ii*) conflict of interest, and (*iii*) non-economic factors.

(*i*) *Economic factors.* Economic factors are largely related to wages and allowances. These factors become more relevant in view of 8–9 per cent rate of inflation in the country, expect during the late 1990s when annual average rate of inflation declined to 5–6 per cent per annum. Inflation causes a decline in real wages at the rate at which it grows. This has been a major cause of demand for compensatory increase in wages and salaries, especially due to increase in cost of living. Workers find that compensatory increase in wages (dearness allowance, etc.) is not sufficient to make up the loss. Besides, workers feel that the wages and salaries they are paid are not commensurate with their productivity, hazardous working conditions and the risks involved, e.g., in case of mine workers and those working in cement factories. Nearly one-third of the industrial disputes arise on account of this factor. Also, the issue of *bonus payment* is another economic factor. About 7% industrial unrest is said to have been caused by non-payment of bonus or its delayed payment.

(*ii*) *Conflict of interests.* One of the main causes of disharmonious industrial relations is the conflict of interests of the employers and employees. The conflict arises mainly because of the gap between what employers offer in terms of wages and salary, bonus, working conditions, workers' safety, and non-pecuniary benefits and what workers expects or demand from their employers. The employers aim at minimizing their cost and workers aim at maximizing their earnings and other benefits. Another area of conflict of interest is productivity and efficiency expectations of the employers and the humanitarian concerns, e.g., safety and leisure of the employees.

(*iii*) *Non-economic factors.* Non-economic factors include faulty personnel management, inhuman treatment of labour, ideological conflicts and political interventions. Poor personnel management, especially where workers are maltreated and denied their dues, causes a general

discontentment among the workers and a feeling of anguish and anger. If this situation is allowed to precipitate, it creates a situation for labour unions to take up the issue with management along with other pent-up issues. Also, an unnecessary distance maintained between the management and workers and denial of opportunity to express their feelings and problems hardens workers' attitude towards the management, creates a sense of apathy and resistance to managerial expectations and reduces the sense of belonging. All these culminate into a kind of cold class war. There was a time in India when ideological confrontation—capitalism *vs*. socialism—often politically motivated, used to create a situation of something like class war between the capitalists and the labour, the proletariat.

42.3.3 Labour Unions: Growth and Role

A *labour union* is a body of workers formed to protect and promote the common and individual interest of its members. The unionization of labour creates a formidable force against the powerful class of employers with the purpose of protecting the interest of the working class against the exploitative tendency of the employers.

In India, formation of trade unions is limited to only the organized sector constituting all registered firms, companies and corporations employing 50 or more workers and public sector commercial and non-commercial organizations like railways, roadways, post and telecommunication services, administrative services, etc. The registered trade unions enjoy legal recognition, rights and privileges and protection under the Trade Union Act, 1926.

The workers in agricultural and retail trade sectors are largely unorganized. The growth of trade unions and their collective strength has added to their bargaining power tremendously.

Growth of Trade Unions The growth of trade unions and trade union activities in independent India has been phenomenal. The number of registered trade unions has increased from about 4000 in 1951 to 50,000 in 1989 and the membership of trade unions has increased from 2,00,000 to 62,00,000 during the same period. This means an annual average growth of 1211 unions in the number of trade unions and an annual average growth of 15,790 in their membership.[4] This is, by any standard, a very rapid growth of trade unions and their membership, in view of the fact that, in 1991, only 9.6 per cent of the *main workers* were employed in the organized sector. Of the remaining 90.4 per cent working labour force, 34.0 per cent were employed in the unorganized agricultural sector, and 56.4 per cent were self-employed in 1991.[5]

The growth of trade unions and their numerical strength, on the one hand, and resistance of the management even against the genuine demands of the labour unions and, to a considerable extent, the exploitative tendency of the management on the other have been the main causes of industrial disputes.

42.3.4 Settlement of Industrial Disputes

Industrial disputes leading to strikes and lock-outs create tremendous social and economic costs in terms of breakdown of supply system, general inconvenience to the society, loss of job, output and social welfare. Therefore, all possible efforts are made, including

4. There is, perhaps, no parallel example in the world.

5. *Basic Statistics Relating to the Indian Economy,* Centre for Monitoring Indian Economy (CMIE), August 1993, Table 9.3.

government intervention, to settle the disputes and to establish a harmonious relationship between the employers and employees. The measures of settling industrial disputes can be classified as (*i*) non-statutory methods, and (*ii*) statutory methods. These two methods are generally adopted in the same sequence.

Non-statutory methods In general, efforts are first made to settle the disputes amicably through *non-statutory methods*. Non-statutory methods of resolving industrial disputes include all the methods that are used to resolve the problem out of court. In the first place, efforts are made to make an amicable settlement of the disputes through mutual discussions and negotiations between the two parties. Where mutual negotiations fail to work or mutual trust is lacking, a third party, e.g., a government representative, mediates to bring about a negotiation between the management and labour unions on the matters of dispute. Another non-judiciary way of settlement of industrial disputes is *voluntary arbitration*. The provision for *voluntary arbitration* was made under the Code of Discipline adopted by the central organization of employers and labour unions in 1958 and Industrial Truce Resolution of 1962. The voluntary arbitration has certain merits: (*i*) it is less time-consuming, (*ii*) it is less expensive, and (*iii*) it promotes goodwill, mutual trust and confidence between the two parties.

Statutory Method The recourse to statutory methods is taken when non-statutory methods are not agreeable or they fail to provide a solution agreeable to both the parties. Statutory methods are methods of litigation through the procedure laid down by the Industrial Disputes Act. The provisions and procedure of settlement of dispute by the Industrial Disputes Act is discussed below in some detail.

42.3.5 Industrial Disputes Act

The procedure of settlement of industrial disputes has been laid down by the Industrial Disputes Act, 1947, amended from time to time. The most significant amendment in this Act was made in 1982. The Industrial Disputes Act provides the machinery and a detailed procedure of dispute settlement. The industrial dispute settlement machinery includes the appointment of the following authorities.

1. *Setting-up a Grievance Settlement Authority (GSA).* The Industrial Disputes Act provides that industrial establishments employing 50 or more workers will set up a Grievance Settlement Authority (GSA), known also as Works Committee, within the establishment and the management. The matters of disputes have to first go to the GSA. No reference of such matters shall be made to any other authority unless the decision of the GSA is not acceptable to either of the parties.

2. *Appointment of Conciliation Officers.* The Act provides that the government[6] may appoint, by notification in the *Official Gazette*, persons whom it thinks fit to be charged with the duty of settling the industrial dispute. The conciliation officer brings about a settlement by investigating the dispute and all the related matters. If the conciliation officers fail to settle the dispute, they report the matter to the government on the findings of their inquiry and the actions taken by them.

3. ***Setting up a Board of Conciliation.*** The Industrial Disputes Act provides that the government may set up a Board of Conciliation with the purpose of settling the industrial

6. The Central Government in case of Central Government establishments and State Government in most other cases.

disputes. The Board is constituted of a Chairman, an independent person, and one or two members recommended by each party to represent them. The rest of the procedure is the same as in the case of appointment of the conciliation officers.

4. ***Setting up a Court of Inquiry.*** Where necessary, the government may set up a Court of Inquiry to inquire into the dispute, record the arguments and evidence of both the parties, and report its findings within the specified period, which may be extended, if so required. The Court of Inquiry is not supposed to bring about a settlement of the industrial dispute.

5. ***Appointment of Adjudication Bodies.*** The ultimate statutory method of settling industrial disputes is *adjudication* by a third party appointed by the government. The Industrial Disputes Act provides for a three-tier adjudication machinery consisting of (*i*) Labour Courts, (*ii*) Industrial Tribunals, and (*iii*) National Tribunals. A *labour court* consists of only one person, normally a sitting or an ex-judge of a High Court. The Labour Court adjudicates on the propriety and legality of issues or withdrawal of an order, discharge or dismissal of a workman, and withdrawal of any customary concessions and privileges. *Industrial Tribunal* also is chaired by a sitting or ex-judge of a High Court. It adjudicates on such matters as wages and their period and mode of payment, compensatory and other allowances, hours of work and rest intervals, provisions of leaves with and without wages, bonus, profit-sharing, provident funds and gratuity. *National Tribunal* is set up to adjudicate on the matters of national importance and those pertaining to companies having branches in more than one state, e.g., public utility establishments and non-public utility establishments.

Some Other Provisions of the Industrial Disputes Act In addition to making provision for creating a settlement machinery, the Industrial Disputes Act makes certain other important provisions for the settlement of disputes. The other important provisions of the Act including (*i*) prohibiting *strikes and look-outs*, and (*ii*) preventing log-off and retrenchment, closure of the undertaking, fines and penalties, etc.

SUMMARY

- Free enterprise system gives rise to many undesirable social and economic consequences such as (*a*) growth of monopolies, (*b*) production and supply of prohibited goods, (*c*) exploitations of labour and growth of trade unions. In respect to foreign trade, foreign exchange is black-marketed per personal gains, and so on. Therefore, the government has to enact laws and rules to manage these economic problems.
- Economic legislations refer to the laws and rules enacted for controlling and managing undesirable social and economic effects private business. The main economic legislations discussed here include legislations related to control and regulation of monopolies, trade relations between firms and labour unions, industrial disputes and foreign exchange.
- Monopolies and Restricted Trade Practices (MRTP) Act is intended to prevent the growth of monopolies and regulation of monopolies sanctioned under law. The first MRTP Act of free India was legislated in 1956 and was modified several times till 1985.
- Industrial Relations and Industrial Dispute Act aims at resolving the problems between the firms and labour, finding solution to disputes between them and maintaining good industrial relations.

REVIEW QUESTIONS

1. What are the factors that make economic legislations inevitable? How do you think economic legislations serve the society?
2. How does the growth of monopolies and restrictive trade practices affect the industrial growth of a country and social interest? Why can market system itself not eliminate monopolies and restrictive trade practices?
3. What are the major findings of the Monopoly Inquiry Commission and what are its major

recommendations?

4. What are the major causes of industrial disputes? What is the mechanism of settlement of industrial disputes in India?

5. The tremendous growth of trade unions and their activities in India have done more harm than good to the labour community and to the society. Comment.

CHAPTER

43 Corporate Social Responsibility

CHAPTER OBJECTIVES

The objective of this final chapter is to discuss a controversial issue of social responsibility of business corporations. The main aspects discussed here include the following ones.

• The reason why the issue of corporate social responsibility arises

• Views of economists on the issue of corporate social responsibility and

• The Indian scenario in case of social responsibility of businessmen

43.1 INTRODUCTION

We have noted in the preceding chapter that the government interference with market mechanism and control and regulation of the private business become inevitable due to failure of the market system to fulfill the economic and social aspirations of the society. In general, the economic and social aspirations include optimum utilization of resources, near-full, if not full, employment, equitable distribution of income and wealth, economic growth and stability. The greater the failures of the market system in achieving these goals, the greater the need for government interference, control and regulation of the private business enterprises. Another factor that is attributed to the increasing role of the government in the economic system is the failure of private corporations to recognize, accept and fulfill their social responsibilities. This implies that if businessmen fulfill their social responsibilities, the need for government interference with private business will be considerably reduced, if not eliminated altogether. Now the questions that arise here are: (*i*) What social responsibilities are businessmen supposed to own and fulfill? and (*ii*) Is the social responsibility of businessmen consistent with their basic objective of making maximum profit? We will review in this chapter, the answer which economists and business analysts offer to these questions. In addition, we will look at the response of the businessmen in India to their social responsibilities.

43.2 WHAT ARE CORPORATE SOCIAL RESPONSIBILITIES (CSRs)?

The *question of the social responsibility* of business corporation is essentially an ethical question that combines social welfare with private business-motives of making profit and wealth. The areas of social responsibilities of the businessmen have been charted

out differently by different authors. We present here a consolidated view on the social responsibility of businessmen.

The Research and Policy Committee—a sub-committee of the Committee for Economic Development (1971)—of the USA gave its opinion on the corporate social responsibility in the following words: "Private Business functions by public consent, and its basic purpose is to serve constructively the needs of society—to the satisfaction of the society"1. And, if private business fails to fulfill the aspirations of the society, the social sanction may be withdrawn and legislative control measures may be tightened. The Committee recognizes three layers of social responsibility of the businessmen which in the order of their importance are: (*i*) providing products, jobs and achieving a sustainable economic growth; (*ii*) responding to the changing social values and priorities, and (*iii*) improving the social environment. According to Bowen, social responsibility "refers to the obligations of businessmen to pursue those policies, to make those decisions or to follow those lines of action which are desirable in terms of the objectives and value of our society"2. The World Business Council defines CSR as "continuing commitment by business to behave ethically and contribute to economic development while improving the quality of life of the work force and their families as well as the local community and society as large " and adds that a corporation has to play 'its role as a producer, employer, markets, customer and a citizen in a responsible manner"3.

To look at the corporate social responsibility in specific terms, one needs to recognize the fact that different sections of the society—shareholders, employees, consumers, government and the society as a whole—are associated with private business in different ways. While some sections are related directly and some indirectly and the interest of the different sections is affected by the private business decisions and activities in different ways and in different measures. Therefore, social responsibilities of businessmen have to be different for different sections of the society. The advocates of the corporate social responsibility specify these towards different sections in different terms. The specific responsibilities of businessmen towards (*i*) the shareholders, (*ii*) the employees, (*iii*) the consumers, (*iv*) the government and (*v*) the society as a whole, as enunciated by the economists and business analysts are listed below.

1. **Responsibilities towards shareholders**

(*i*) A reasonable rate of return over time,

(*ii*) The survival and growth of their company, and

(*iii*) Building reputation and goodwill of the company.

2. **Responsibilities towards employees**

(*i*) Fair wages and regular payment of wages,

(*ii*) Good working conditions and safety,

(*iii*) Reasonable work standards and norms,

(*iv*) Labour welfare services—health, education, recreation and accommodation, and

(*v*) Training and promotion,

1. Quoted in Webb, Samuel C., *Managerial Economics, op. cit.*, Ch. 28.

2. Bowen, Howard R., *Social Responsibility of Businessmen* (N.Y., Harper & Row, 1953), p. 6.

3. Quoted from article by Satvik Varma in *Economic Times*, 18th Feb. 2011.

(*vi*) Recognition of and respect for hardwork, efficiency, honesty, sincerity as well as loyalty, and

(*vii*) Efficient redressal of employees' grievances.

3. **Responsibilities towards consumers**

(*i*) Providing goods and services at a reasonable price,

(*ii*) Supplying goods and services of promised quality, durability and service,

(*iii*) Supplying socially harmless products,

(*iv*) Offering an efficient consumer redressal mechanism,

(*v*) Warding off middlemen as far as possible,

(*vi*) Resisting profiteering and black-marketing and

(*vii*) Improving product quality through research and development, inventions and innovations, and improved efficiency.

4. **Responsibilities towards government**

(*i*) Regular payment of taxes,

(*ii*) Abiding by laws and regulations,

(*iii*) Resisting bribing bureaucrats and ministers,

(*iv*) Avoiding taking advantage of loopholes in business laws,

(*v*) Cooperating with the government in research and development,

(*vi*) Cooperating with the government in upgradation of environment and

(*vii*) Cooperating with the government in promoting social values.

5. **Responsibilities towards society as a whole**

(*i*) Prevention of environmental pollution,

(*ii*) Preservation of ethical and moral values,

(*iii*) Making provision of health, education and cultural services,

(*iv*) Minimizing ecological imbalance, and

(*v*) Choosing appropriate technology.

43.3 DO BUSINESS CORPORATIONS HAVE SOCIAL RESPONSIBILITIES?

As already mentioned, the question whether business corporations have social responsibilities to fulfill is an ethical question and the answer to that is a matter of opinion and the individual perception on this ethical issue. Therefore, the question of social responsibility of private business has been controversial. We have already noted above that the government committees and a section of business analysts are of the view that businessmen do have certain responsibilities towards the different sections of the society associated with the private business and the society as whole. Here we look at the opposite view.

Friedman and Baumol, two of the greatest economists of the 20th century, are opposed to the view that businessmen have any social responsibility to fulfill. In the opinion of Friedman4, the view that corporations and labour unions should accept social responsibilities

4. Friedman, Milton, *Capitalism and Freedom* (University of Chicago Press, Chicago, 1962).

"shows a fundamental misconception of the character and nature of a free economy". He argues that in a free economy, "there is one and only one social responsibility of business— to use its resources and engage in activities designed to increase its profits so long as it stays within the rules of the game, which is to say, engage in open and free competition without deception or fraud… If businessmen do have social responsibility other than making maximum profit for stockholders how are they to know what it is?" In his opinion, assigning any social responsibility to private entrepreneurs other than profit maximization is 'a fundamentally subvervise doctrine' and it undermines the 'very foundation' of a free society.

Baumol5 is of the view that private business should not be asked to assume the responsibility of fulfilling the social and political goals of the society, nor should they be expected to allocate resources optimally. For, in his opinion, a competitive system automatically rewards efficiency and punishes inefficiency, and where it fails, fiscal measures—taxes and subsidies may be adopted to correct the system and to encourage the business in favour of social goals.

Some other economists argue that even if it is accepted that business corporations have social responsibility, it is extremely difficult to fix it. As Niall Fitzerald has remarked, "corporate social responsibility is a hard-edged business decision" (*TOI*, 28 November 2007).

Fixing corporate social responsibility is extremely difficult because (*i*) social responsibility, being an ethical issue, is difficult to define: it means different things to different persons, (*ii*) it is not easy to fix social responsibility in practicable terms because there are no standard rules, and (*iii*) private firms, even big corporations, do not have control over the market mechanism of resource allocation. In addition, the following arguments are put forward against assigning social responsibility to business corporations.

(*i*) Social responsibility is a matter of public policy, not a matter of business policy;

(*ii*) Businessmen do not have legal powers or social sanction to meddle with social welfare;

(*iii*) Businessmen do not have legal powers to prevent anti-social activities of others;

(*iv*) Managers are not empowered to spend money on social welfare beyond a limit;

(*v*) Expenditure on social welfare often leads to rise in prices;

(*vi*) Businessmen's social responsibility will make them more powerful and will create conflict between the businessmen and the government, and

(*vii*) As regards setting moral values, corporations are not moral agents.

Clearly, the views on whether businessmen have any social responsibility are divergent and arguments against assigning social responsibility to businessmen are equally strong.

The divergence of views, however, should not mean that private business has no social responsibility. It is, of course, difficult to fix social responsibilities for the businessmen and it would be unreasonable to expect the businessmen to give up their profit maximization motive in favour of social interest. However, if one examines the arguments against the social responsibility of businessmen, one would find that they are more emotive than logical, and are not very convincing. For example, look at Friedman's argument that assigning social responsibility to private business is contrary to "the character and nature of a free economy". The validity of this argument is limited in many respects.

5. Baumol, William J., in Melvin Ashen (ed.), *Managing the Socially Responsible Corporations* (Macmillan, New York, 1974), Ch. 4.

First*,* this argument is not tenable for the societies which do not have the 'character and nature of a free economy'. Evidently, many countries outside the 'first world' do not have such an economy as the countries like US have. So this argument has a limited applicability.

Secondly*,* Friedman would agree that if growth of monopolies and concentration of economic power damage the very foundation of the free society, some social responsibility has to be fixed with social sanction. As regards the government's power to control the monopoly powers, big business houses influence the political decisions and, in a way do rule the country with their 'invisible hands' and, therefore, the process is the other way round.

Thirdly, as regards the argument that businessmen have the sole objective of profit maximization, some economists[6] object strongly to profit being the sole objective of modern corporations. In their opinion, the profit maximization objective often leads to growth of monopoly and concentration o f economic power, and it is wrong to say that the state will take care of monopolies because large corporations use their money power to bring into power the government of their choice. How can such a government control monopolies in the real sense of the term?

Fourthly, it may be unreasonable to expect businessmen to replace their profit motive with economic welfare of the society, but it is not unreasonable to expect from businessmen to desist from the anti-social activities in which they often indulge, e.g., fleecing the consumer under conditions of scarcity, supplying adulterated, substandard and spurious goods (especially foodstuffs and medicines), black marketing, exploitation of labour (paying them much lower wages than their productivity), not complying with effluent treatment laws, polluting air and water in the residential areas posing risk to human survival, and so on. Many more such malpractices can be cited from the Indian scenario to which we will return shortly. Even the economists like Friedman and Baumol would agree that businessmen should refrain from such anti-social activities in pursuit of their goal of profit. If not *social responsibility*, businessmen must have at least social sensibility, or else they damage the society and their business in the process.

Finally, the world opinion is turning in favour of the view that the businessmen do have some social responsibilities to fulfill. If they ignore their social responsibilities, however defined and determined, they would not only harm the society and harm their own interest but also invite legal measures and problems for their survival. Many factories in residential and unauthorized areas in Delhi were bulldozed by the MCD in 1990s. To have a closer look at the anti-social activities of businessmen, we turn to the Indian scene.

43.4 THE INDIAN SCENARIO

The attention of private business towards social responsibility was drawn in 1978 by the High-Powered Committee on MRTP (Sachar Committee). In its own words, "... in the development of the corporate ethics, we reached a stage where the question of the social responsibility of [private] business can no longer be scoffed at or taken lightly"[7]. It adds, "The company must behave and function as a responsible member of society... The company must accept its obligation to be socially responsible and to work for the larger benefit of the community"[8].

6. Heilbroner, Robert L., et. al. *In the Name of Profit* (Doubleday, N.Y., 1927), pp. 242–45.

7. *Report of the High-Powered Expert Committee on Companies and MRPT Acts*, Government of India, New Delhi, 1978, pp. 9.95.

8. *Ibid*. p. 110.

As regards the attitude of India's businessmen towards their social responsibility, the scene appears to be complex with different scenarios at different levels of business. The scene at the corporate level is different from the one at the small and medium scale business.

At the corporate level, the issue of social responsibility has two dimensions: (*i*) meeting social responsibility in terms of financial contributions towards social welfare activities, and (*ii*) aligning the working of the corporations with the general welfare of the society, i.e., keeping social gains at par with private gains. The corporate sector seems to be fairly well aware of its social responsibilities as far as financial contributions towards social welfare activities are concerned. The big corporations have made significant contribution to the promotion of social welfare activities like building of schools, colleges, charitable hospitals and dispensaries, research and technology institutes, creating special chairs for professors of excellence in the universities, and so on.

However when the issue is examined in the overall social perspective, the working of the corporate sector leaves much to be desired. The corporate sector shows little or no concern for environment and water pollution and safety for human life. Consider, for example, the tragic case of Bhopal gas leak and loss of human life. Exploitation of labour (paying wages less than their productivity) has become an accepted norm of the corporate sector. A much more dangerous trend set in India is the use of money power by big business houses to corrupt the politicians, political environment and bureaucracy for personal gains. India has been rated among the topmost corrupt nations of the world. However, it is unfair to hold business houses alone responsible for the widespread corruption in the country.

At the level of medium and small scale business, things are much worse. Rampant adulteration of consumer eatables is the most heinous crime by a section of the business community. Adulterated mustard oil claimed hundreds of lives and caused physical crippling of over a thousand persons. According to a report of a Committee testing quality of medicines, 63% medicines were found to be adulterated. The Indian market is flooded with substandard imitation of popular brand names (especially of electrical and eletronic goods) and spurious drugs (including life-saving ones). Many cases of cheating the households by fake investment companies have been detected. Recall the case of Subrota Roy who has cheated households to the extent of `17,000 crore. This kind of business has no concern for human life, let alone social responsibility. This is no business: it is a social crime in the name of business. This, however, should not mean that all small and medium scale businessmen are of the same type.

43.5 INDIAN SCENARIO (EXTENSION)

Due to such antisocial business practices, there is now growing awareness in India to formulate rules and regulations for putting the responsibility on business corporations to perform some social functions. It is suggested that India Inc should spend some part of their profits for performing social responsibility. A parliamentary standing committee was set up in 2010 to make its recommendations for assigning social responsibilities to business corporations. The Committee had recommended that companies with a turnover of `1000 crore or net profit of `5 crore or more earmark 2% of their net profit for performing corporate social responsibilities. Although corporates and industry chambers has lobbied against the move, the Ministry of Company Affairs proceeded to finalize the

new Company bill and make a provision for the corporates to assign at least 2% of their net profits for performing social responsibilities.

Besides, during mid-1990s, corporates did realise that they do have to perform some social responsibilities. Not only that, the reputed alumni of IIM-A is planning to set up a thin-thank centre for promoting CSR and to play catalyst role in restructuring the agenda, convincing companies on their social responsibility and to set standards for measuring CSR. With situation turning favourable, the Ministry of Corporate Affairs issued in 2009 voluntary corporate social responsibility (CSR) guide lines for Indian corporates. With corporate attitude turning favourable, the commerce ministry is considering the possibility of making CSR mandatory. It will be mandatory in the sense that if company fails to spend required amount on social responsibility, it will have to provide reasons in its final reports. It may thus be concluded that despite some disagreement by the corporate sector, the need for CSR is being realized strongly and it may be a part of company law in future.

SUMMARY

- As industrialized countries had experienced, working of the free enterprise system led to the emergence of many anti-social and anti-growth elements such as growth of monopolies, exploitation of labour, unemployment. This created the condition for assigning some social responsibilities to the private entrepreneurs.
- The economists have different opinions on the issue whether corporates should be assigned some social responsibility. In the opinion of some economists, specially Paul A. Samuelson and W. J. Baumol, the only objective of business corporation to make maximum profit and they have no any social responsibility. But some other economists have shown through empirical evidence that business corporations do create anti-social and anti-growth factors in the economy and the government has to interfere with the working system of the private entrepreneurs.
- Things are changing in India. The commerce ministry, a section of businessmen and economist agree on the issue of CSR. The government proceeding, though gradually, to frame rules and regulation for making CSR mandatory for business corporations. At present, CSR is in an voluntary form.

REVIEW QUESTIONS

1. Do businessmen have any social responsibilities to fulfill? If your answer is ‘yes’, how will you justify your answer?
2. What are the social responsibilities of business corporations? To what extent do you think businessmen honour and fulfill their social responsibility in India?
3. What social responsibility do business corporations owe to different sections of the society? How do you think business corporations can bring a compromise between the conflicting objectives of

profit maximization and their social responsibilities towards the different sections of the society?

4. Is the concept of social responsibility of business corporations an ethical or legal issue? Elaborate on your answer.

5. In the opinion of some economists, the only social responsibility of business corporations is maximization of profit. Do you agree with their view? Give arguments to support your answer.

6. Write a note on the attitude of Indian businessmen towards their social responsibility and their performance in that regard.

FURTHER READING

Ashen Melvin (ed.), *Managing the Socially Responsible Corporations* (Macmillan, N.Y., 1970).

ChAMberlin, neil W., *The Limits of Corporate Responsibility* (Basic Books, N.Y., 1973).

Committee of Economic Development, *Social Responsibility of Business Corporations,* N.Y., 1971.

FriedMAn, M., *Capitalism and Freedom* (University of Chicago Press, Chicago, 1962).

JACoby, neil h., *Corporate Power and Social Responsibility: A Blueprint for the Future* (Macmillan, N.Y., 1973).

MAnne, h.G., and h.C., WAlliCh, *The Modern Corporation and Social Responsibility,* Amercian Enterprise Institute for Public Policy Research, Washington, D.C., 1972.

sAMuelson, P.A., *The Economic Role of Private Activity* (McGraw-Hill, N.Y., 1970).

steiner, GeorGe A., Business and Society, 2nd edition (Random House, New York).

Index

Document Outline

Table of Contents