Lesson 1

INTRODUCTION TO MANAGERIAL ECONOMICS

The heart of Managerial economics is the micro economic theory. Much of this theory was formalized in a textbook written more than 100 years ago by Professor Alfred Marshall of Cambridge University. The world has changed a great deal since Marshall's ideas were developed. Yet, basic micro economic principles such as supply and demand, elasticity, short-run and long-run shifts in resource allocation, diminishing returns, economies of scale, and pricing according to marginal revenue and marginal cost continue to be important tools of analysis for managerial decision makers.

Economics is divided into two broad categories: Micro and Macro.

Microeconomics is the study of the economic behavior of individual decision-making units. It has a great relevance to Managerial Economics. On the other hand, **Macroeconomics** is the study of the total or aggregate level of output, income, employment, consumption, investment, and prices for the economy viewed as a whole. In economics, the key term is **Scarcity**. In the presence of a limited supply relative to demand, countries must decide how to allocate their scarce resources. This decision is central to the study of economics:

- What to produce?
- How to produce?
- And for whom to produce?

These are the well known what, how and for whom questions found in the introductory chapter of all economics textbooks.

DEFINITION OF MANAGERIAL ECONOMICS

Joel Dean, author of the first managerial economics textbook, defines managerial economics as "the use of economic analysis in the formulation of business policies".

Douglas - "Managerial economics is the application of economic principles and methodologies to the decision-making process within the firm or organization."

Pappas & Hirschey - "Managerial economics applies economic theory and methods to business and administrative decision-making."

Salvatore - "Managerial economics refers to the application of economic theory and the tools of analysis of decision science to examine how an organization can achieve its objectives most effectively."

The meaning of this definition can best be examined with the aid of Figure 1-1

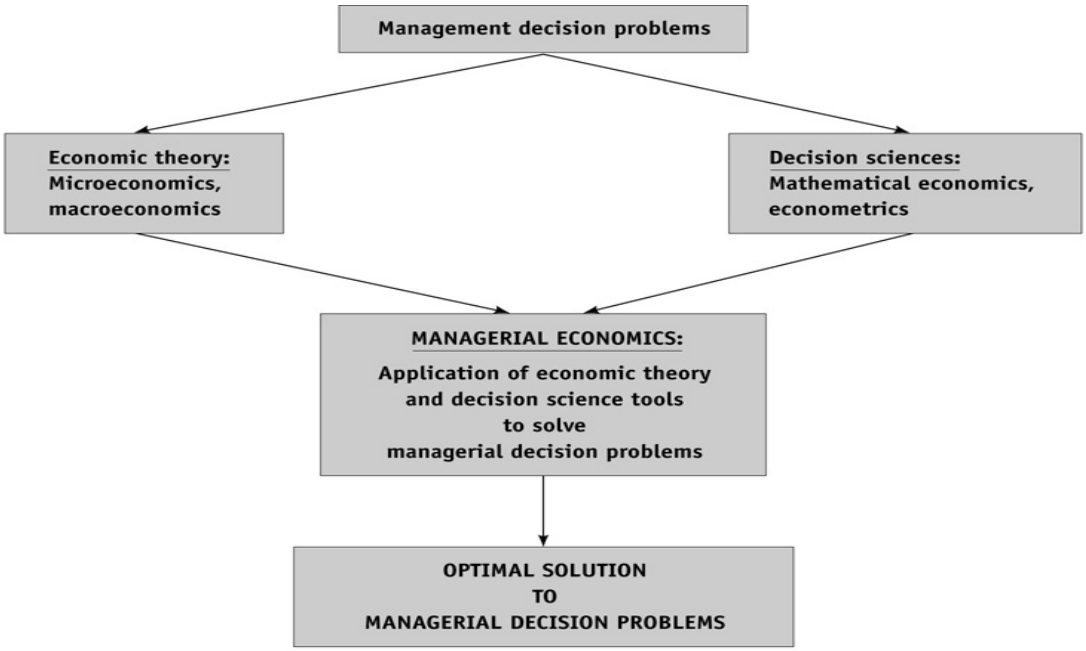


FIGURE 1-1 The Nature of Managerial Economics*
* Managerial economics *refers to the application of economic theory and decision science tools to find the optimal solution to managerial decision problems.*

RELATIONSHIP TO ECONOMIC THEORY

Economic theories seek to predict and explain economic behavior. Economic theories usually begin with a model. For example, the theory of the firm assumes that the firm seeks to maximize profits, and on the basis of that it predicts how much of a particular commodity the firm should produce under different forms of market structure. The profit-maximization model accurately predicts the behavior of firms, and, therefore, we accept it. Thus, the methodology of economics is to accept a theory or model if it predicts accurately.

RELATIONSHIP TO THE DECISION SCIENCES

Managerial economics is also closely related to the decision sciences. These use the tools of mathematical economics and econometrics to construct and estimate decision models aimed at determining the optimal behavior of the firm**. Mathematical economics** is used to formalize the economic models in equational form postulated by economic theory. Econometrics then applies statistical tool (particularly regression analysis) to real-world data to estimate the models postulated by economic theory and for forecasting.

SCOPE OF MANAGERIAL ECONOMICS

Managerial economics has applications in both profit and not-for-profit sectors. For example, an administrator of a nonprofit hospital seeks to provide the best medical care possible given limited medical staff, beds and equipment. Using the tools and concepts of managerial economics, the administrator can determine the optimal allocation of these limited resources. In short, managerial economics helps managers arrive at a set of operating rules that help in the efficient use of scarce human and capital resources. By following these rules, businesses, educational institutions, hospitals, other nonprofit organizations, and government agencies are able to meet their objectives efficiently.

THEORY OF THE FIRM

The theory of firm is the center-piece and central theme of Managerial economics. A firm is an organization that combines and organizes resources for the purpose of producing goods and/or services for sale.

The model of business is called the theory of the firm. In its simplest version, the firm is thought to have profit maximization as its primary goal. Today, the emphasis on profits has been broadened to include uncertainty and the time value of money. In this more complete model, the primary goal of the firm is long-term expected value maximization.

EXPECTED VALUE MAXIMIZATION

The value of the firm is the present value of all expected future profit of the firm. Future profits must be discounted at an appropriate interest rate .to the present because a dollar of profit in the future is worth less than today. This model can be expressed as follows:

Formally the wealth or value of the Firm = Present Value of Expected Future Profits

$$PV = \frac{\pi_1}{(1+r)^1} + \frac{\pi_2}{(1+r)^2} + \cdots + \frac{\pi_n}{(1+r)^n} = \sum_{t=1}^{n} \frac{\pi_t}{(1+r)^t}$$

Here, π1, π2, . . . πn represent expected profits in each year, t, and r is the appropriate interest, or discount, rate.

$$\text{Value of Firm} = \sum_{t=1}^{n} \frac{\pi_t}{(1+r)^t} = \sum_{t=1}^{n} \frac{TR_t - TC_t}{(1+r)^t}$$

CONSTRAINTS AND THE THEORY OF THE FIRM

Managerial decisions are often made in light of constraints imposed by technology, resource scarcity, contractual obligations, laws, and regulations. Organizations frequently face limited availability of essential inputs, such as skilled labor, raw materials, energy, specialized machinery, and warehouse space.

LIMITATIONS OF THE THEORY OF THE FIRM

Some critics question why the value maximization criterion is used as a foundation for studying firm behavior. The theory of the firm which postulates that the goal of the firm is to maximize wealth or the value of the firm has been criticized as being much too narrow and unrealistic.

Hence, broader theories of the firm have been purposed. The most prominent among these are:

- Sales maximization (Adequate rate of profit)
- Management utility maximization (Principle-agent problem)
- Satisfying behavior

These alternative theories, or models, of managerial behavior have added to our understanding of the firm. Still, none can replace the basic value maximization model as a foundation for analyzing managerial decisions.

DEFINITIONS OF PROFIT

- **Business or Accounting Profit**: Total revenue minus the explicit or accounting costs of production.

- **Economic Profit**: Total revenue minus the explicit and implicit costs of production.

THEORIES OF PROFIT

- Risk-Bearing Theories of Profit
- Frictional Theory of Profit
- Monopoly Theory of Profit
- Innovation Theory of Profit
- Managerial Efficiency Theory of Profit

Lesson 2

ECONOMIC OPTIMIZATION PROCESS

Optimization is mainly concerned with finding maximum and minimum points, also known as optimum points of a function. Applications include finding optimum values for functions such as profit, cost, revenue, production and utility. These functions which are to be maximized or minimized are called objective function.

Examples

- Consumers maximize utility by purchasing an optimal combination of goods
- Firms maximize profit by producing and selling an optimal quantity of goods
- Firms minimize their cost of production by using an optimal combination of inputs

Just as there is no single "best" purchase decision for all customers at all times, there is no single "best" investment decision for all managers at all times. When alternative courses of action are available, the decision that produces a result most consistent with managerial objectives is the **optimal decision.** The process of arriving at the best managerial decision is the goal of economic optimization and the focus of managerial economics.

MAXIMIZING THE VALUE OF THE FIRM

In managerial economics, the primary objective of management is assumed to be maximization of the value of the firm. This value maximization objective which we have introduced in our lesson 1, is expressed as:

Maximizing the above equation is a complex task that involves consideration of future revenues, costs, and discount rates. For many day-to-day operating decisions, managers typically use less complicated, partial optimization techniques.

EXPRESSING ECONOMIC RELATIONSHIPS

Common ways of specifying Economic functions are:

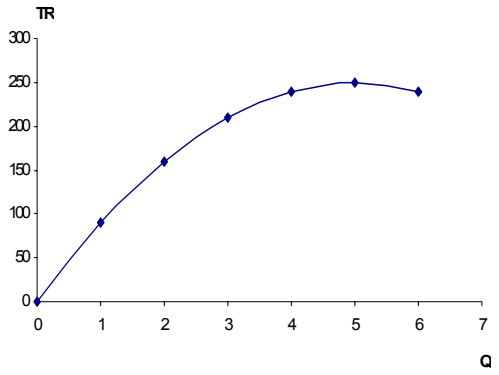
- Set form
- Functional form
- Graphs
- table

Tables:

Q	0	1	2	3	4	5	6
TR	0	90	160	210	240	250	240

$S = \{(a,b)/\ a \varepsilon Q \text{ and } b \varepsilon R\}$ (set form)

Equations: $TR = 100Q - 10Q^2$ (Functional Form)



Tables are the simplest and most direct form for presenting economic data. When these data are displayed electronically in the format of an accounting income statement or balance sheet, the tables are referred to as **spreadsheets**. When the underlying relation between economic data is simple, tables and spreadsheets may be sufficient for analytical purposes. In such instances, a simple **graph** or visual representation of the data can provide valuable insight.

Complex economic relations require more sophisticated methods of expression. An **equation** is an expression of the functional relationship among economic variables.

TOTAL, AVERAGE, AND MARGINAL RELATIONS

Total, average, and marginal relations are very useful in optimization analysis. The relationship between total, average and marginal concepts is extremely important in optimization analysis. A marginal relation is the change in the dependent variable caused by a one-unit change in an independent variable. For example, marginal revenue is the change in total revenue associated with a one-unit change in output; marginal cost is the change in total cost following a one-unit change in output; and marginal profit is the change in total profit due to a one-unit change in output.

REVENUE RELATIONS

Price and Total Revenue

Total Revenue = Price × Quantity

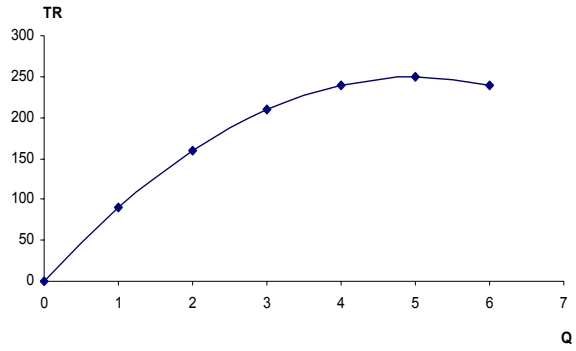
Marginal Revenue

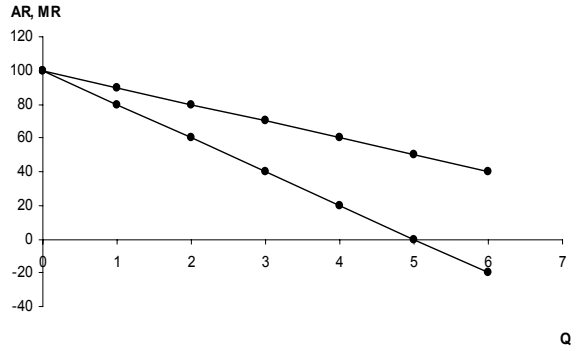
Change in total revenue associated with a one-unit change in output.

Revenue Maximization

Quantity with highest revenue, MR = 0.

Q	TR	AR	MR
0	0	-	-
1	90	90	90
2	160	80	70
3	210	70	50
4	240	60	30
5	250	50	10
6	240	40	-10





PROFIT RELATIONS

Total and Marginal Profit

- Total Profit (π) = Total Revenue - Total Cost.
- Marginal profit is the change in total profit due to a one-unit change in output, $M\pi = MR - MC$.

Profit Maximization

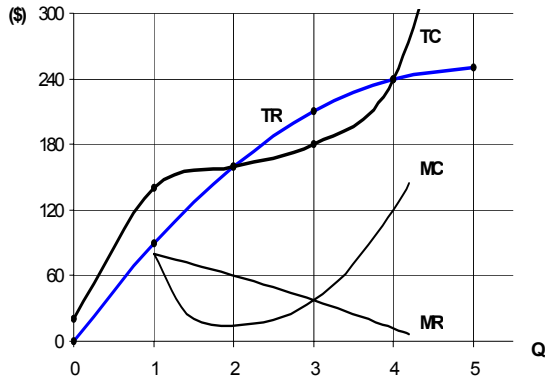
- Profit is maximized when $M\pi = MR - MC = 0$ or $MR = MC$, assuming profit declines as Q rises.

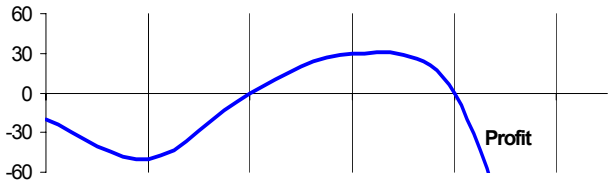
Marginal v. Incremental Profits

- Marginal profit is the gain from producing one more unit of output (Q).
- Incremental profit is gain tied to a managerial decision, possibly involving multiple units of Q.

PROFIT MAXIMIZATION

Q	TR	TC	Profit
0	0	20	-20
1	90	140	-50
2	160	160	0
3	210	180	30
4	240	240	0
5	250	480	-230





COST RELATIONS

Total Cost
Total Cost = Fixed Cost + Variable Cost.

Marginal and Average Cost
Marginal cost is the change in total cost associated with a one unit change in output.
Average Cost = Total Cost/Quantity

Average Cost Minimization

- Average cost is minimized when MC = AC.
- Reflects efficient production of a given output level.

Total Cost (TC) = Fixed Costs (FC) + Variable Costs (VC)

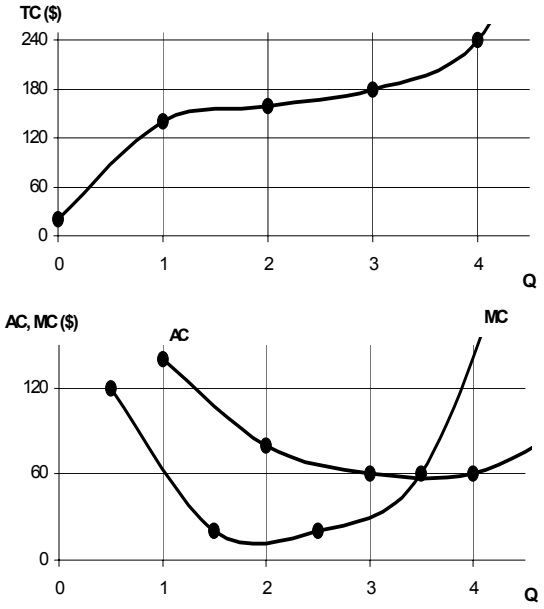
FC = a
VC = $bQ + Q^2$
TC = $a + bQ + Q^2$

Marginal Costs (MC) = dTC/dQ

MC = $b + 2Q$

Average Total Cost (ATC) = Total Cost/Q

ATC = $(a + bQ + Q^2)/Q$ so that: ATC = $a/Q + b + Q$



GEOMETRIC RELATIONSHIPS

- The slope of a tangent to a total curve at a point is equal to the marginal value at that point
- The slope of a ray from the origin to a point on a total curve is equal to the average value at that point
- A marginal value is positive, zero, and negative, respectively, when a total curve slopes upward, is horizontal, and slopes downward
- A marginal value is above, equal to, and below an average value, respectively, when the slope of the average curve is positive, zero, and negative

Lesson 3

ECONOMIC OPTIMIZATION WITH CALCULUS

MARGINAL ANALYSIS IN DECISION MAKING

The marginal analysis is one of the most important concepts in managerial economics in general and in optimization analysis in particular. According to marginal analysis, the firm maximizes profits when marginal revenue equals marginal cost. Marginal cost (MC) is defined as the- change in total cost per unit change in output and is given by the slope of the TC curve. Marginal analysis gives clear rules to follow for optimal resource allocation. As a result, geometric relations between totals and marginals offer a fruitful basis for examining the role of marginal analysis in managerial decision making.

Geometric relations between totals and marginals offer a fruitful basis for examining the role of marginal analysis in economic decision making. Managerial decisions frequently require finding the maximum value of a function. For a function to be at a maximum, its marginal value (slope) must be zero. Evaluating the slope, or marginal value, of a function, therefore, enables one to determine the point at which the function is maximized.

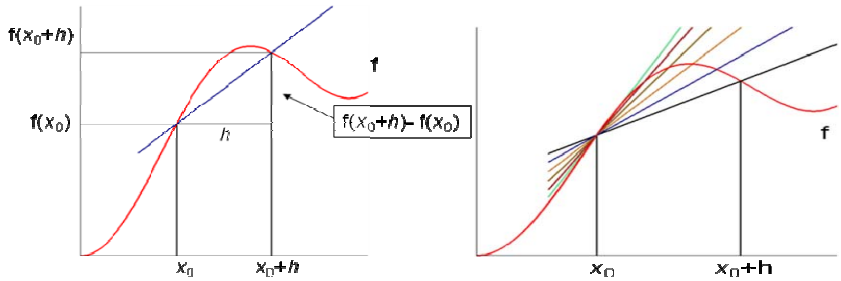
TANGENTS AS LIMITS OF SECANT LINES

Slope is a measure of the steepness of a line and is defined as the increase (or decrease) in height per unit of movement along the horizontal axis. The slope of a straight line passing through the origin is determined by dividing the *Y* coordinate at any point on the line by the corresponding *X* coordinate. Using Δ (read *delta*) to designate change:

$$\textbf{Slope} = \Delta Y/\Delta X$$

The marginal relation has a similar geometric association with the total curve. The slope of a nonlinear curve varies at every point on the curve. Slopes of nonlinear curves are typically found geometrically by drawing a line tangent to the curve at the point of interest and determining the slope of the tangent. A **tangent** is a line that touches but does not intersect a given curve.

The basic problem that leads to differentiation is to compute the slope of a tangent line of the graph of a given function f at a given point x_0. The key observation, which allows one to compute slopes of tangent lines, is that the tangent is a certain limit of secant lines as illustrated in the figure below. A secant line intersects the graph of a function f at two or more points. $x=x_0$ and $x=x_0 + h$. As h approaches 0, the secant line in question approaches the tangent line at the point $(x_0, f(x_0))$.



CONCEPT OF THE DERIVATIVE

A marginal value is the change in a dependent variable associated with a 1– unit change in an independent variable. If Y = f(X). Using Δ to denote change, it is possible to express the change in the value of the independent variable, X, by the notation ΔX and the change in the dependent variable, Y, by ΔY. The ratio $\Delta Y/\Delta X$ is a general specification of the marginal concept. A derivative is a precise specification of marginal relation.

A marginal value is the change in a dependant variable associated with a 1-unit change in an independent variable.

$$\text{Marginal } Y = \Delta Y/\Delta X$$

Finding a derivative involves finding the value of the ratio $\Delta Y/\Delta X$ for extremely small changes in X. In symbols:

$$dY/dX = \text{Lim } \Delta Y/\Delta X \quad \text{as} \quad \Delta X \to 0$$

Lim of the slope of the Secant line as $\Delta X \to 0$ = slope of the Tangent

The derivative of a function is a very precise measure of its slope or marginal value at a particular point. The terms derivative and marginal are interchangeable. Thus maxima or minima of a function occur where its derivative or marginal value is equal to zero. Geometrically, this corresponds to the point of graph, where the curve has zero slope.

RULES OF DIFFERENTIATION

1- Constant Function Rule: The derivative of a constant, Y = f(X) = a, is zero for all values of a (the constant).

$$Y = f(X) = a$$

$$\frac{dY}{dX} = 0$$

2- Power Function Rule: The derivative of a power function, where a and b are constants, is defined as follows.

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

$$\frac{dY}{dX} = b \cdot aX^{b-1}$$

3- Sum-and-Differences Rule: The derivative of the sum or difference of two functions, U and V, is defined as follows.

$$U = g(X) \qquad V = h(X) \qquad Y = U \pm V$$

$$\frac{dY}{dX} = \frac{dU}{dX} \pm \frac{dV}{dX}$$

4- Product Rule: The derivative of the product of two functions, U and V, is defined as follows:

$$U = g(X) \qquad V = h(X) \qquad Y = U \cdot V$$

$$\frac{dY}{dX} = U\frac{dV}{dX} + V\frac{dU}{dX}$$

5- Quotient Rule: The derivative of the ratio of two functions, U and V, is defined as follows:

$$U = g(X) \qquad V = h(X) \qquad Y = \frac{U}{V}$$

$$\frac{dY}{dX} = \frac{V\left(dU/dX\right) - U\left(dV/dX\right)}{V^2}$$

6- Chain Rule: The derivative of a function that is a function of X is defined as follows:

$$Y = f(U) \qquad U = g(X)$$

$$\frac{dY}{dX} = \frac{dY}{dU} \cdot \frac{dU}{dX}$$

DERIVATIVE OF A DERIVATIVE

Geometrically, the derivative refers to the slope of the function, while the second derivative refers to the *change* in the slope of the function. The value of the second derivative can thus be used to determine whether we have a maximum or a minimum at the point at which the first derivative (slope) is zero. The rule is *if the second derivative is positive, we have a minimum, and if the second derivative is negative, we have a maximum.*

Given objective function Y = f(X)
First Order Condition (F O C)
Find X such that dY/dX = 0
Second Order Condition (S O C)
If $d^2Y/dX^2 > 0$, then X is a minimum.
If $d^2Y/dX^2 < 0$, then X is a maximum.

EXAMPLE 1 MAXIMIZATION
Given the following total revenue (TR) function, determine the quantity of output (Q) that will maximize total revenue:

$TR = 100Q - 10Q^2$
$dTR/dQ = 100 - 20Q = 0$
$Q^* = 5$ and $d^2TR/dQ^2 = -20 < 0$

EXAMPLE 2 MINIMIZATION
Given the following marginal cost function (MC), determine the quantity of output that will minimize MC:

$MC = 3Q^2 - 16Q + 57$
$dMC/dQ = 6Q - 16 = 0$
$Q^* = 2.67$ and $d^2MC/dQ^2 = 6 > 0$

MULTIVARIATE OPTIMIZATION
Objective function $Y = f(X_1, X_2, X_3)$

Find all partials w.r.t X_1, X_2 & X_3 & set them equal to zero

PARTIAL DERIVATIVE:
$\partial Y/\partial X_1 = dY/dX_1$ while keeping X_2 and X_3 constant. Partial derivatives follow the same pattern as the rules that we have described for ordinary derivatives.

OPTIMIZATION OF MULTIVARIATE FUNCTIONS

Example 3
Determine the values of X and Y that maximize the following profit function:
$\pi = 80X - 2X^2 - XY - 3Y^2 + 100Y$

Solution
$\partial\pi/\partial X = 80 - 4X - Y = 0$
$\partial\pi/\partial Y = -X - 6Y + 100 = 0$
Solve simultaneously
$X = 16.52$ and $Y = 13.92$
$\pi = 80(16.52) - 2(16.52)^2 - (16.52)(13.92) - 3(13.92)^2 + 100(13.92)$
$= \$ 1,356.52$

ROLE OF CONSTRAINTS
Solution to economic problems frequently have to be found under constraints e.g. maximizing utility subject to a budget constraint or minimizing costs subject to production quota constraint. Managers frequently face constrained optimization problems, decision situations with limited choice alternatives. e.g.; marketing managers are assumed to maximize sales, subject to the constraint that they do not exceed a fixed advertising budget.
Mathematically, what the constraint does is to narrow down the domain, and hence the range of the objective function.

STEPS IN CONSTRAINED OPTIMIZATION
- Define an objective mathematically as a function of one or more choice variables
- Define one or more constraints on the values of the objective function and/or the choice variables
- Determine the values of the choice variables that maximize or minimize the objective function while satisfying the constraint.

LAGRANGIAN METHOD
Form the Lagrangian function by adding the Lagrangian multiplier and constraint to the objective function and then optimize the Lagrangian function. The solution to the Lagrangian function will automatically satisfy the constraint.

Example 4: Lagrangian Method
Use the Lagrangian method to maximize the following profit function:
$\pi = 80X - 2X^2 - XY - 3Y^2 + 100Y$
Subject to the following constraint:
$X + Y = 12$ (output capacity constraint)

Set the constraint function equal to zero and obtain
$0 = 12 - X - Y$
Form the Lagrangian function
$L = 80X - 2X^2 - XY - 3Y^2 + 100Y + \lambda(12 - X - Y)$

Find the partial derivatives and solve simultaneously

$\partial L/\partial X = 80 – 4X –Y - \lambda = 0$ 1

$\partial L/ \partial Y = – X – 6Y + 100 - \lambda = 0$ 2

$\partial L/ \partial \lambda = 12 - X – Y = 0$ 3

Subtract Eq 2 from Eq 1, we get Eq 4

$- 3X + 5Y -20 = 0$ 4

Solution: $X = 5$, $Y = 7$, and $\lambda = 53$

Interpretation of the Lagrangian Multiplier, λ

The value of Lambda, λ, has an important economic interpretation. It is the marginal effect on the objective-function solution associated with a 1-unit change in the constraint. In our example, $\lambda = 53$, so a 1-unit increase in the output capacity constraint from 12 to 13 units will cause profit to increase by approximately $53 i-e by λ times.

To generalize, a Lagrangian multiplier,λ, indicates the marginal effect of decreasing or increasing the constraint requirement by one unit. This will decrease or increase the optimal value of the objective-function by λ times. So it helps managers to evaluate the potential benefit or cost of relaxing constraints.

Lesson 4

DEMAND ANALYSIS

IMPORTANCE OF DEMAND FOR A FIRM

Demand is one of the most important aspects of managerial economics, since a firm would not be established or survive if a sufficient demand for its product did not exist. That is, a firm could have the most efficient production techniques and the most effective management, but without a demand for its product, it simply would not survive. Many firms go out of business soon after being set up because their expectation of a sufficient demand for their products fails to come up, even with a great deal of advertising. Each year also sees many previously established and profitable firms close as a result of consumers shifting their purchases to different firms and products. Demand is, thus, essential for the creation, survival, and profitability of a firm.

DEMAND

Demand is the quantity of a good or service that customers are willing and able to purchase during a specified period under a given set of economic conditions. The time frame might be a month, or a year. Conditions to be considered include the price of the good in question, prices and availability of related goods, expectations of price changes, consumer incomes, consumer tastes and preferences, advertising expenditures, and so on.

For managerial decision making, a prime focus is on market demand. Market demand is the aggregate of individual demand Individual demand is determined by the value associated with getting and using any good or service and the ability to get it. Both are necessary for effective individual demand. Desire without purchasing power may lead to want, but not to demand.

DIRECT DEMAND

There are two basic models of individual demand. One, known as the theory of consumer behavior, relates to the direct demand for personal consumption products. This model is appropriate for analyzing individual demand for goods and services that directly satisfy consumer desires. This is also labeled as consumer demand.

DERIVED DEMAND

The outputs of engineers, production workers, managers, lawyers, consultants, office business Machines and natural resources are all examples of goods and services demanded not for direct consumption but rather for their use in providing other goods and services. Their demand is derived from the demand for the products they are used to provide. Input demand is called derived demand. This is also sometimes called business demand. The inputs purchased by a Business can be classified into raw materials, energy, labor, and capital, which may be substitutes or complements. .

LAW OF DEMAND

Holding all other things constant (*ceteris paribus*), there is an inverse relationship between the price of a good and the quantity of the good demanded per time period. It is often useful to examine the relationship between the quantity demanded of a commodity per, unit of time and the price of the commodity only. This can be achieved by assuming, for, the moment, that the individual's income, the price of related commodities, and tastes are unchanged. The inverse relationship between the price and the quantity demanded of the commodity per time period is then the individual’s demand schedule for the commodity, and the plot of data (with price on the vertical axis and the quantity on the horizontal axis) gives the corresponding individual's demand curve per time period at lower prices. The inverse relationship between the price of the commodity and the quantity demanded per time period is referred to as the law of demand.

COMPONENTS OF DEMAND

- Substitution Effect
- Income Effect

The reason for the negative slope of d_X, or inverse relationship between P_X and Qd_X, is not difficult to find. When P_X falls, the quantity demanded of the commodity by the individual (Qd_X) increases because the individual substitutes in consumption, commodity X for other commodities (which are now relatively more expensive). This is called **the substitution effect**. In addition, when the price of a commodity falls, a consumer can purchase more of the commodity with a given money income (i.e., .his or her real income increases). This is called the **income effect**. Thus, a fall in P_X leads to an increase in Qd_X (so that d_X is negatively sloped) because of the substitution effect and the income effect.

Assuming that *real income* is constant: If the *relative price* of a good rises, then consumers will try to substitute away from the good. Less will be purchased. If the *relative price* of a good falls, then consumers will try to substitute away from other goods. More will be purchased. The **substitution effect** is consistent with the law of demand.

The *real value* of income is inversely related to the prices of goods. A change in the real value of income: will have a direct effect on quantity demanded if a good is normal and will have an inverse effect on quantity demanded if a good is inferior. The **income effect** is consistent with the law of demand only if a good is normal.

DEMAND CURVE DETERMINATION

Demand curve shows price and quantity relation holding everything else constant.

Change in quantity demanded or movements along the same demand curve

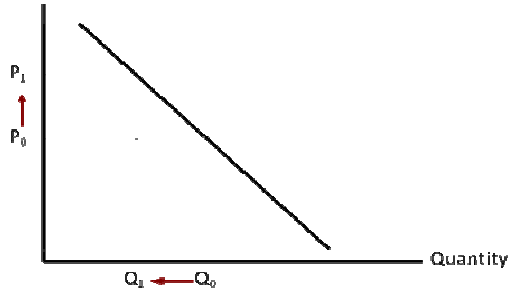
- Quantity demanded falls if price rises.
- Quantity demanded rises if price falls.

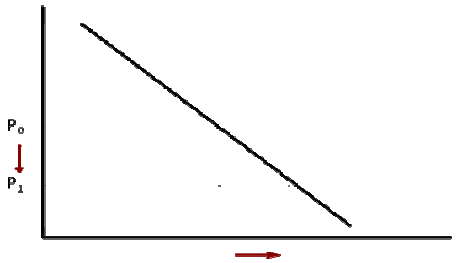
Role of Non-Price Variables

Change in non-price variables will define a new demand curve that is demand curve shifts upwards or downwards.

- Demand increases if a non-price change allows more to be sold at every price.
- Demand decreases if a non-price change causes less to be sold at every price.

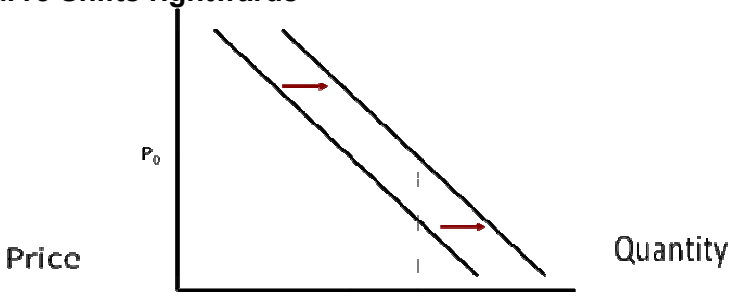
Movement along the same Demand curve ‘



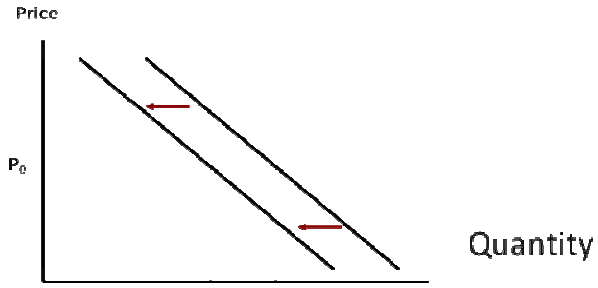


DEMAND SHIFTERS

Entire Demand Curve Shifts rightwards



Entire Demand Curve Shifts leftwards



DEMAND FUNCTION

The demand for a commodity arises from the consumer's willingness and ability to purchase the commodity. Consumer demand theory assumes that the quantity demanded of a commodity is a function of, or depend on, the price of the commodity, the consumer's income, the price of related commodities, and the tastes of the consumer. In functional form, we can express this as:

$QD_X = f(P_X, N, I, P_Y, T)$,Where

QD_X = quantity demanded of commodity X
P_X = price per unit of commodity X
N = number of consumers on the market
I = consumer income
P_Y = price of related (substitute or complementary)commodity
T = consumer tastes

The market demand curve for a commodity is simply the horizontal summation of the demand curves of all the consumers in them market. For example, in the top part of the following figure, the market demand curve for commodity X is obtained by the horizontal summation of the demand curve of individual 1 (d_1) and individual 2 (d_2), on the assumption that they are the only

two consumers in the market. Thus, at P x = $1, the market quantity demanded of 5 units of commodity X is the sum of the 3 units of X demanded by individual 1 and the 2 units of X demanded by individual 2. If there were 100 individuals in the market instead, each with demand curve d_X, the market demand curve for commodity X would be D_X (see the bottom part of Figure). D_X has the same shape as d_X, but the horizontal scale refers to hundreds of units of commodity X. The market demand curve for a commodity shows the various quantities of the commodity demanded in the market per time period (Qd_X) at various alternative prices of the commodity, while holding every thing else constant.

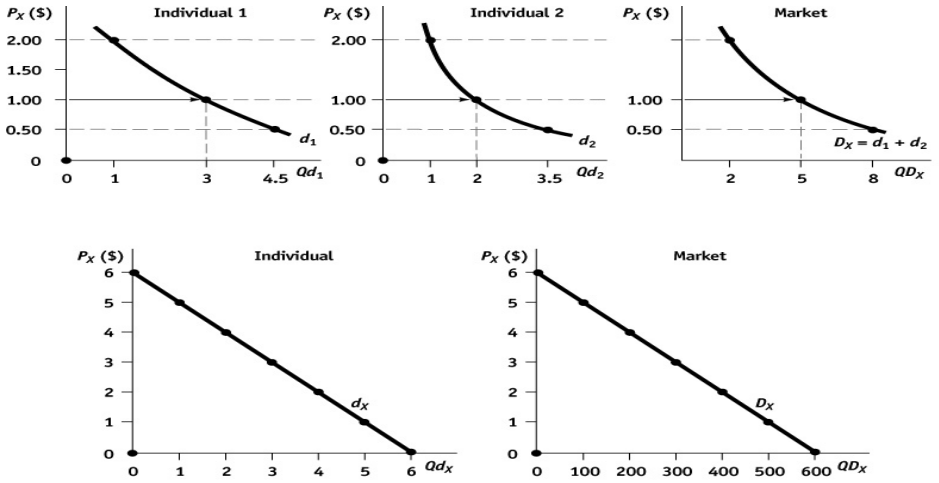


FIGURE 3-2 From Individual to Market Demand The top part of the figure shows that the market demand curve for the commodity, D_X, is obtained from the horizontal summation of the demand curve of individual 1 (d_1) and individual 2 (d_2). The bottom part of the figure shows an individual's demand curve, d_X, and the market demand curve, D_X, on the assumption that there are 100 individuals in the market with demand curves identical to d_X.

In managerial economics we are primarily interested in the demand for a commodity faced by the firm. Since the analysis of the firm is central to managerial economics, we are primarily interested in the demand for a commodity faced by a firm. The demand for a commodity faced by a particular firm depends on the size of the market or industry demand for the commodity, the form in which the industry is organized, and the number of firms in the industry.

The demand for a firm's product also depends on the type of product that the firm sells. If the firm sells durable goods e.g., automobiles, washing machines, and refrigerators that provide services not only during the year when they are purchased but also in subsequent years, or goods that can be stored, the firm will generally face unstable demand than a firm selling nondurable goods. The reason is that consumers can run their cars, washing machines, or refrigerators a little longer by increasing their expenditures on maintenance and repairs, and they can postpone the purchase of a new unit until the economy improves and their income rises.

THE THEORY OF CONSUMER CHOICE MATHEMATICALLY

Behind Demand curve lies the Theory of Consumer Choice. Suppose that a consumer spends all of his or her income on commodities *X* and Y. To reach equilibrium, the consumer must maximize utility (U) subject to his or her budget constraint. That is, the consumer must

$$\textbf{Maximize } U = f(Q_X.\ Q_Y)$$
$$\textbf{Subject to } M = P_X Q_X + P_Y Q_Y$$

This constrained maximization problem can be solved by the Lagrangian multiplier method.

To do so, we first form the Lagrangian function:

$$\mathbf{L = f(Q_X, Q_Y) + \lambda(M - P_X Q_X - P_Y Q_Y)}$$

To maximize L, we find the partial derivative of L with respect to Q_X, Q_Y and λ and set them equal to zero, That is the First-order conditions imply(FOC)

$$\partial L/\partial Q_X = \partial f/\partial Q_X - \lambda P_X = 0$$
$$\partial L/\partial Q_Y = \partial f/ \partial Q_Y - \lambda P_Y = 0$$
$$\partial L/\partial \lambda = M - P_X Q_X - P_Y Q_Y = 0$$

Second Order Partials :SOC

$$\partial MU_X/\partial Q_X < 0 \text{ and } \partial MU_Y/\partial Q_Y < 0$$
$$\partial MU_X/\partial Q_Y \text{ and } \partial MU_Y/\partial Q_X$$
$$\text{Discriminant} = \partial MU_X/\partial Q_X * \partial MU_Y/\partial Q_Y - \partial MU_X/\partial Q_Y * \partial MU_Y/\partial Q_X > 0$$

Solving Equations and solve for *λ* and setting them equal to each other, we get

$$\lambda = \frac{MU_X}{P_X} = \frac{MU_Y}{P_Y}$$

.

$$\mathbf{\lambda = MU_X/P_X = MU_Y/P_Y}$$
$$\mathbf{Or\ MU_X/MU_Y = P_X/P_Y}$$

Where MU_X is the marginal or extra utility that the individual receives from consuming the last unit of commodity X and MU_Y is the marginal utility of *Y.* Thus, Equation postulates that in order to maximize utility subject to the budget constraint (i.e., in order to be in equilibrium), the individual must spend his or her income so that the marginal utility of the last dollar spent on X equals the marginal utility of the last dollar spent on *Y.* Thus, λ is the marginal utility of the last dollar spent on X and *Y* when the consumer is in equilibrium.

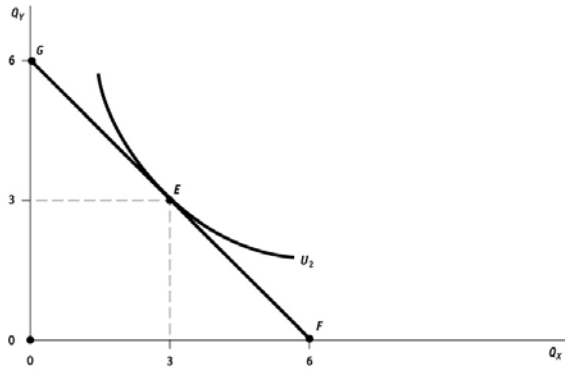


FIGURE 3-9 The Consumer's Equilibrium Given budget line *GF*, the consumer is in equilibrium when he or she consumes 3*X* and 3*Y* (point *E*), where budget line *GF* is tangent to the indifference curve U_2 (the highest indifference curve that the consumer can reach with his or her budget line).

Utility maximization requires:

$$\mathbf{P_X/P_Y = MU_X/MU_Y, \text{ or}}$$
$$\mathbf{MU_X/P_X = MU_Y/P_Y}$$

From this equilibrium condition (which is also called Tangency condition), we get one point on the individual, demand curves for commodity X and commodity *Y.* By changing the price of X and *Y* and repeating the process, we obtain other points of consumer equilibrium, and, by joining these, we can derive the individual's demand curve *for* commodities X and *Y* (i.e., d_X and d_Y).

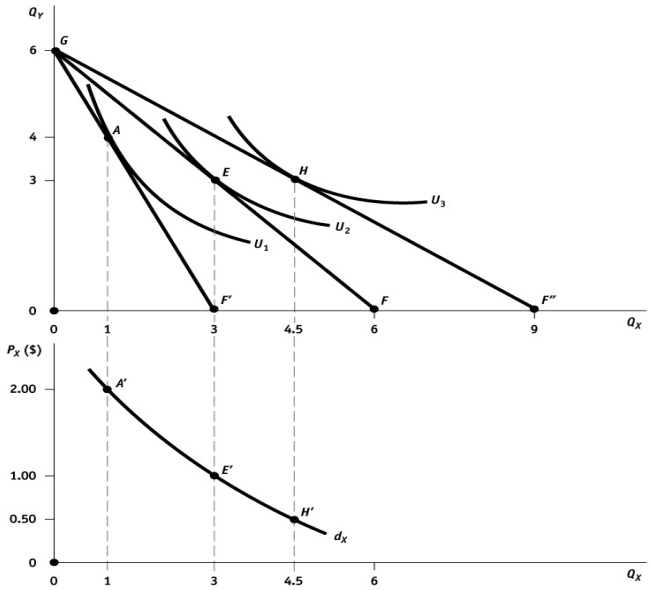


FIGURE 3-10 Derivation of the Consumer's Demand Curve The top panel shows that with $P_X = \$2$, $P_X = \$1$, and $P_X = \$0.67$, we have budget lines GF', GF, and GF'', and consumer equilibrium points A, E, and H, respectively. From equilibrium points A, E, and H in the top panel, we derive points A', E', and H', in the bottom panel. By joining points A', E', and H', we derive d_X, the consumer's demand curve for commodity X.

Lesson 5

SUPPLY ANALYSIS

BASIS FOR SUPPLY

The term **Supply** refers to the quantity of a good or service that producers are willing and able to sell during a certain period under a given set of conditions. Factors that must be specified include the price of the good in question, prices of related goods, the current state of technology, levels of input prices, weather, and so on. The amount of product that producers bring to the market, the supply of the product—depends on all these influences.

LAW OF SUPPLY

A decrease in the price of a good, all other things held constant, will cause a decrease in the quantity supplied of the good. An increase in the price of a good, all other things held constant, will cause an increase in the quantity supplied of the good. Changes in price result in changes in the quantity supplied shown as movement *along* the supply curve, while Changes in non-price determinants result in changes in supply shown as a *shift* in the supply curve.

Among the factors influencing the supply of a product, the price of the product itself is often the most important. Higher prices increase the quantity of output producers want to bring to market. When marginal revenue exceeds marginal cost, firms increase supply to earn the greater profits associated with expanded output. Higher prices allow firms to pay the higher production costs that are sometimes associated with expansions in output. On the other hand, lower prices typically cause producers to supply a lower quantity of output. Just as there are non-price determinants of demand, there are non-price determinants of supply. A change in any one or a combination of these factors will cause the supply curve shift to the right or to the left.

NON-PRICE DETERMINANTS OF SUPPLY

- costs and technology
- prices of other goods or services offered by the seller
- future expectations
- number of sellers
- input prices
- weather conditions

INDUSTRY SUPPLY VERSUS FIRM SUPPLY

Just as in the case of demand, supply functions can be specified for an entire industry or an individual firm. Even though factors affecting supply are highly similar in industry versus firm supply functions, the relative importance of such influences can differ markedly.

Managerial decision making requires understanding both individual firm supply and market supply conditions. Market supply is the aggregate of individual firm supply, so it is ultimately determined by factors affecting firm supply.

SUPPLY CURVE AND SUPPLY FUNCTION

Supply curve shows price and quantity relation holding everything else constant. Along a supply curve, all non-price variables are held constant. A rise in price will increase the quantity supplied, while a fall in price will decrease the quantity supplied. The **supply function** specifies the relation between the quantity supplied and all variables that determine supply. The **supply curve** expresses the relation between the price charged and the quantity supplied, holding constant the effects of all other variables.

SUPPLY CURVE SHIFTS

Supply increases if a non-price change allows more to profitably produced and sold. S-curve shits *downwards (right).* On the other hand, supply decreases if a non-price change causes less to be profitably produced and sold. S-curve shifts *upwards (left).*

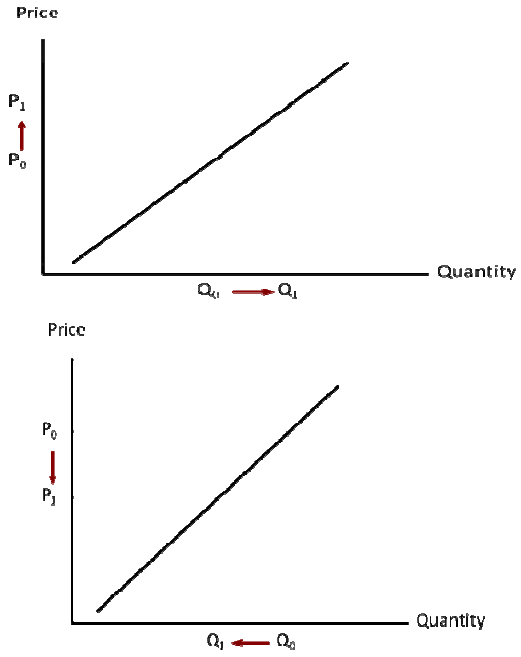
The market **supply function** for a product is a statement of the relation between the quantity supplied and all factors affecting that quantity. The generalized supply function expressed in a supply Equation must lists variables that influence supply. As is true with the demand function, the supply function must be made explicit to be useful for managerial decision making.

Consider the supply function for automobile industry and assume that the supply function has been specified as follows:

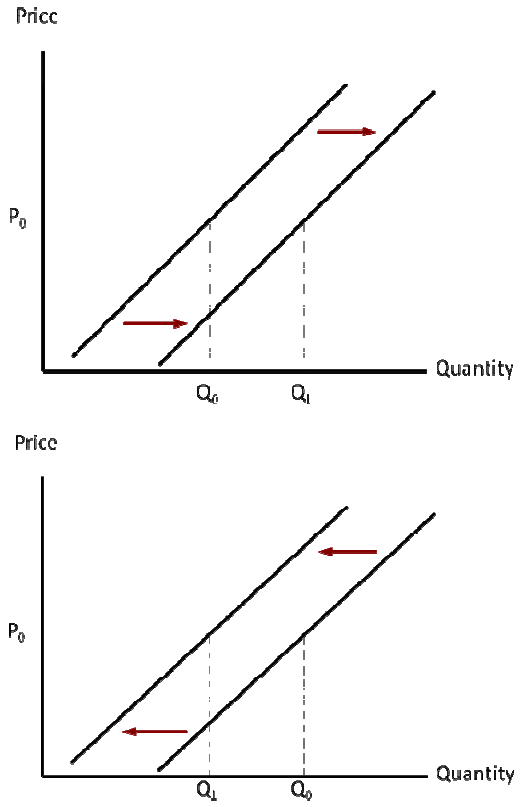
$$Q = b_1P + b_2P' + b_3W + b_4S + b_5E + b_6i$$

The above equation states that the number of new domestic cars supplied during a given period, *Q*, is a linear function of the average price of new domestic cars, *P*; average price of new sport car, *P'*; average hourly price of labor, *W*; average cost of steel, *S*; average cost of energy, *E*; and average interest rate (cost of capital in percent), *i*. The terms *b*1, *b*2, . . . , *b*6 are the parameters of the supply function.

MOVEMENT ALONG THE SUPPLY CURVE



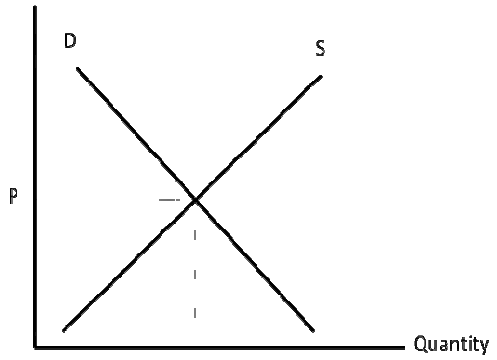
CHANGE IN SUPPLY



MARKET EQUILIBRIUM

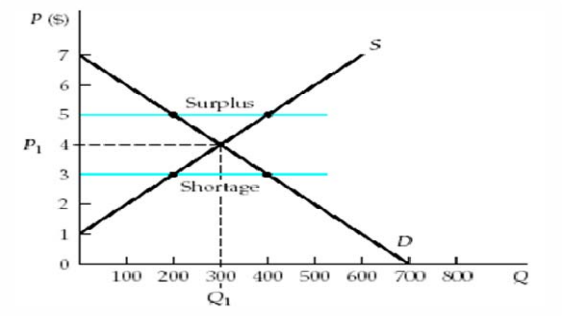
Market equilibrium is determined at the intersection of the market demand curve and the market supply curve. **Equilibrium price** is the price that equates the quantity demanded with the quantity supplied. This price is referred to as the **market equilibrium price**, or the market clearing price, because it just clears the market of all supplied product. **Equilibrium quantity** is the amount that people are willing to buy and sellers are willing to offer at the equilibrium price level.

Market equilibrium describes a condition of perfect balance in the quantity demanded and the quantity supplied at a given price. In equilibrium, there is no tendency for change in either price or quantity.



MARKET DISEQUILIBRIUM

A **surplus** is created when producers supply more of a product at a given price than buyers demand. Surplus describes a condition of excess supply. Conversely, a **shortage** is created when buyers demand more of a product at a given price than producers are willing to supply. Shortage describes a condition of excess demand. Neither surplus nor shortage will occur when a market is in equilibrium, because equilibrium is defined as a condition in which the quantities demanded and supplied are exactly in balance at the current market price. Surplus and shortage describe situations of market disequilibrium because either will result in powerful market forces being exerted to change the prices and quantities offered in the market.



COMPARATIVE STATICS

Equilibrium exists when there is no economic incentive for change in demand or supply. Changing demand or supply affects equilibrium. Comparative Statics is the study of how equilibrium changes with changing demand or supply. This change continues until a new equilibrium is established. A surplus describes an excess in the quantity supplied over the quantity demanded at a given market price. A surplus results in downward pressure on both market prices and industry output. Shortage describes an excess in the quantity demanded over the quantity supplied at a given market price. A shortage results in upward pressure on both market prices and industry output.

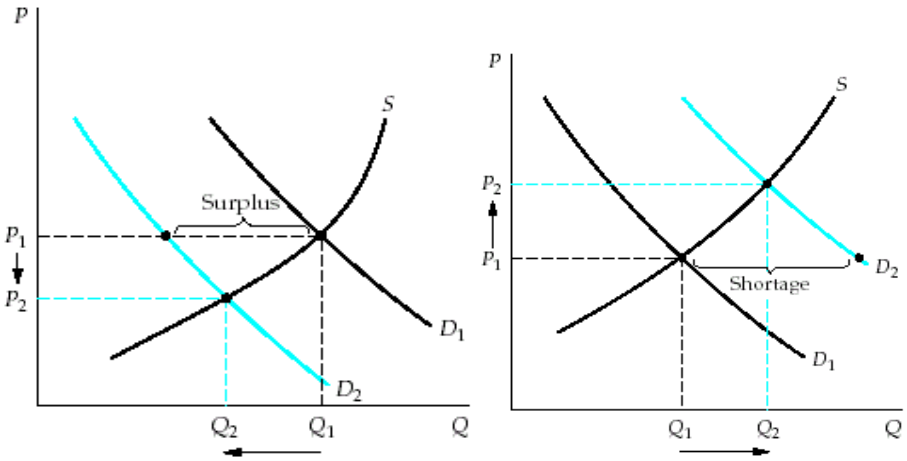
COMPARATIVE STATICS ANALYSIS

The **short run** is the period of time in which:

- sellers already in the market respond to a change in equilibrium price by adjusting variable inputs
- buyers already in the market respond to changes in equilibrium price by adjusting the quantity demanded for the good or service

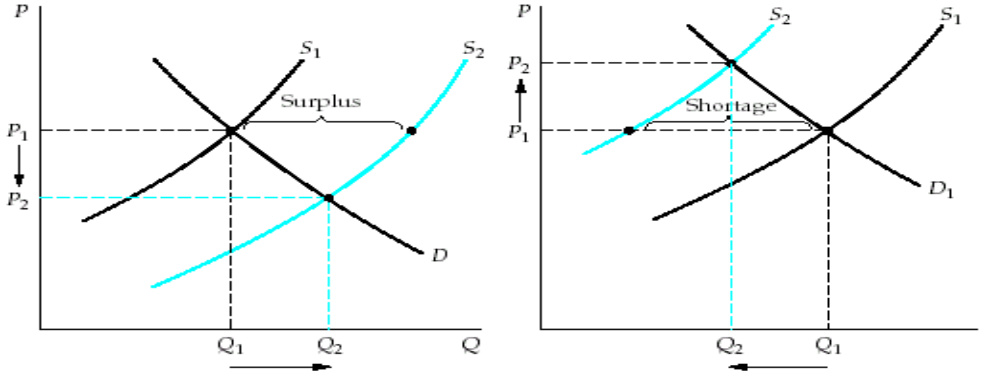
SHORT-RUN ANALYSIS

Comparative statics of Changing Demand: Holding supply conditions constant, demand will vary with changing interest rates. Demand falls with a rise in interest rates; demand increases as interest rates falls.



COMPARATIVE STATICS: CHANGING SUPPLY

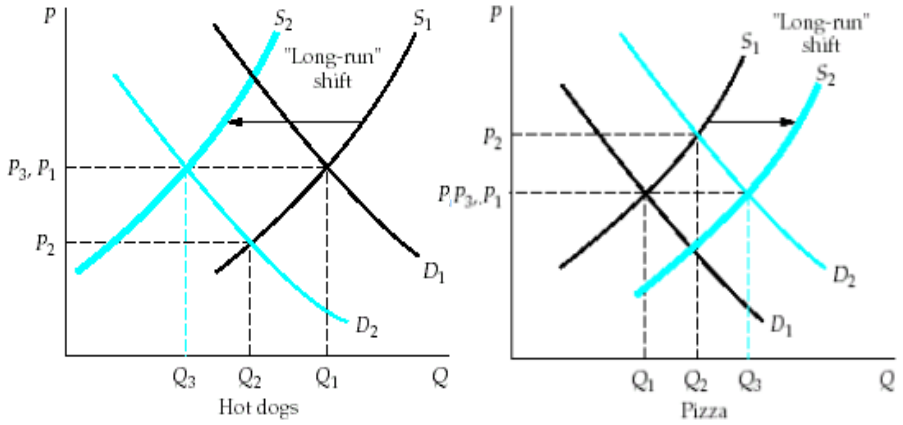
An increase in supply causes equilibrium price to fall and equilibrium quantity to rise while a decrease in supply causes equilibrium price to rise and equilibrium quantity to fall.



LONG RUN ANALYSIS

The **long run** is the period of time in which:

- new sellers may enter a market
- existing sellers may exit from a market
- existing sellers may adjust fixed factors of production
- buyers may react to a change in equilibrium price by changing their tastes and preferences



Initial change in the left panel: decrease in demand from D_1 to D_2, as a result there is a reduction in equilibrium price and quantity (to P_2, Q_2)
Follow-on adjustment:

- movement of resources out of the market
- leftward shift in the supply curve to S_2

Equilibrium price and quantity (to P_3, Q_3)
Initial change in the right panel: increase in demand from D_1 to D_2 that results in an increase in equilibrium price and quantity (to P_2, Q_2)
Follow-on adjustment:

- movement of resources into the market
- rightward shift in the supply curve to S_2

Equilibrium price and quantity (to P_3, Q_3)

Lesson 6

DEMAND SENSITIVITY ANALYSIS

THE ELASTICITY CONCEPT

For useful managerial decision making, the firm must know the sensitivity or responsiveness of demand to changes in factors that make up the underlying demand function. One measure of responsiveness employed not only in demand analysis but throughout managerial decision making is **elasticity**, defined as the percentage change in a dependent variable, *Y*, resulting from a 1 percent change in the value of an independent variable, *X*. The equation for calculating elasticity is:

$$E_p = \frac{\%\Delta \text{ Quantity}}{\%\Delta \text{ Price}}$$

POINT ELASTICITY AND ARC ELASTICITY

Elasticity can be measured in two different ways, point elasticity and arc elasticity. Elasticity measures sensitivity.

Point elasticity shows sensitivity of Y to *small* changes in X. The most widely used elasticity measure is the **price elasticity of demand**, which measures the responsiveness of the quantity demanded to changes in the price of the product, holding constant the values of all other variables in the demand function.

$$\varepsilon_X = \partial Y/Y \div \partial X/X.$$

Arc elasticity shows sensitivity of Y to *big* changes in X.

$$E_X = \Delta Y/\Delta X * (X_2+X_1)/(Y_2+Y_1)$$

Elasticity Varies along Demand Curve:

- As price rises, so too does $|\varepsilon_P|$.
- As price falls, so too does $|\varepsilon_P|$.
- In all cases, $\varepsilon_P < 0$.

OPTIMAL PRICE FORMULA

Price elasticity estimates represent vital information because these data, along with relevant unit cost information, are essential inputs for setting a pricing policy that is consistent with value maximization. This stems from the fact that there is a relatively simple mathematical relation between marginal revenue, price, and the point price elasticity of demand. Given any point price elasticity estimate, relevant marginal revenues can be determined easily. When this marginal revenue information is combined with pertinent marginal cost data, the basis for an optimal pricing policy is derived.

TR = P*Q

MR = dTR/dQ = dP*Q/dQ Applying the product rule of derivatives, we get:

= P + Q * dP/dQ

= P + P[Q/P * dp/dQ]

= P + P(1/E)

MR = P[1 + (1/E)]

Profit maximization requires that

MR = MC

P[1 + (1/E)] = MC

P = MC / P[1 + (1/E)] Optimal Price

MR and ε_P are directly related by the expression:

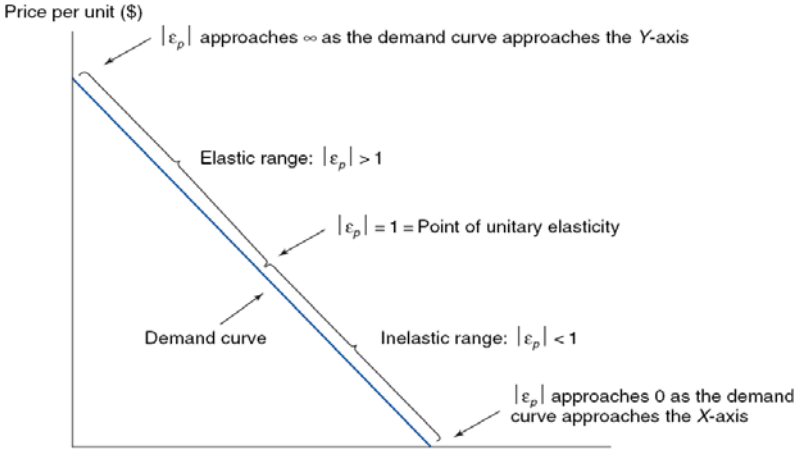
MR = P/[1+(1/ ε_P)].
Optimal P* = MC/[1+(1/ ε_P)].

MARGINAL REVENUE, TOTAL REVENUE, AND PRICE ELASTICITY

There are simple, direct relations between price elasticity, marginal revenue, and total revenue. It is worth examining such relations in detail, given their importance for pricing policy. The relationship between price and revenue depends on elasticity. Why? By itself, a price fall will reduce receipts … BUT because the demand curve is downward sloping, the drop in price will also increase quantity demanded. The question is which effect will be stronger? One of the most important features of price elasticity is that it provides a useful summary measure of the effect of a price change on revenues. Depending on the degree of price elasticity, a reduction in price can increase, decrease, or leave total revenue unchanged. A good estimate of price elasticity makes it possible to accurately estimate the effect of price changes on total revenue. For decision- making purposes, three specific ranges of price elasticity have been identified.

- Price cut increases revenue if $|\varepsilon_P| > 1$.
- Revenue constant if $|\varepsilon_P| = 1$.
- Price cut decreases revenue if $|\varepsilon_P| < 1$.

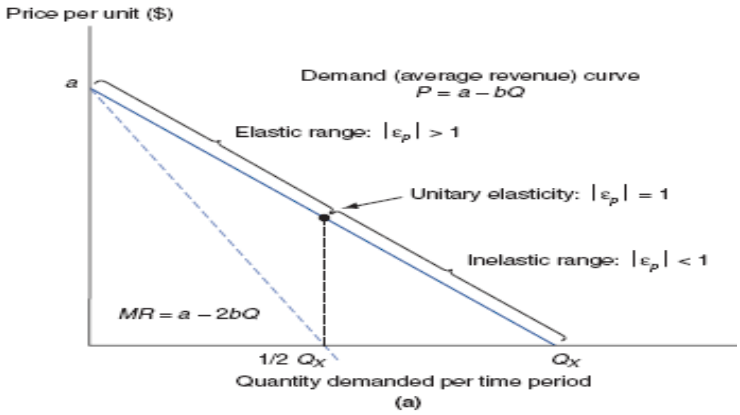
ELASTICITY VARIES ALONG A LINEAR DEMAND CURVE

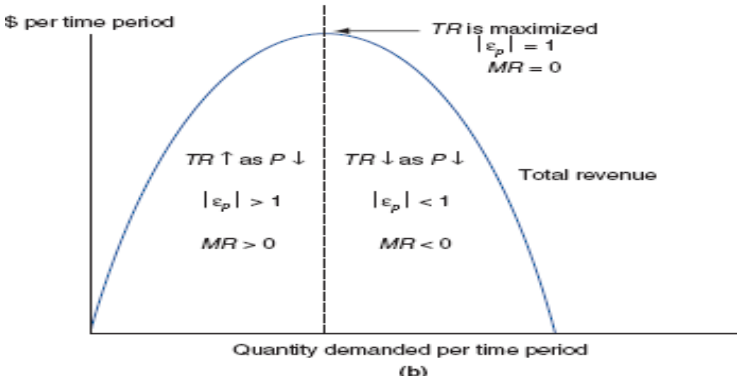


VARYING ELASTICITY AT DIFFERENT POINTS ON A DEMAND CURVE

All linear demand curves, except perfectly elastic or perfectly inelastic ones, are subject to varying elasticities at different points on the curve. In other words, any linear demand curve is price elastic at some output levels but inelastic at others.

- Relative elasticity of demand: $E_p > 1$
- Relative inelasticity of demand: $0 < E_p < 1$
- Unitary elasticity of demand: $E_p = 1$
- Perfect elasticity: $E_p = \infty$
- Perfect inelasticity: $E_p = 0$





AS PRICE DECREASES

- revenue rises when demand is elastic
- revenue falls when it is inelastic
- revenue reaches its peak if elasticity =1

The lower chart shows the effect of elasticity on total revenue.
If P = AR = a – bQ (1)
TR = P*Q
= (a – bQ) * Q
= aQ – bQ2
Then MR = a – 2bQ (2)

The relation between the demand (average revenue) and marginal revenue curves becomes clear when one compares Equations (1) and (2). Each equation has the same intercept *a*. This means that both curves begin at the same point along the vertical price axis. However, the marginal revenue curve has twice the negative slope of the demand curve. This means that the marginal revenue curve intersects the horizontal axis at *1/2QX*, given that the demand curve intersects at *QX*. The above Figure(a) shows that marginal revenue is positive in the range where demand is price elastic, zero where $E_p = -1$, and negative in the inelastic range. Thus, there is an obvious relation between price elasticity and both average and marginal revenue. As shown in Figure (b), price elasticity is also closely related to total revenue.

CONSTANT PRICE ELASTICITY OF DEMAND

Some demand curves have **constant elasticity;** the Demand curve assumes the shape of a rectangular hyperbola (so that TR is constant regardless of price. Such a curve has a nonlinear equation:

$Q = aP^{-b}$

where $-b$ is the elasticity coefficient and equals to –b throughout the demand curve.

FACTORS AFFECTING THE PRICE ELASTICITY OF DEMAND

There are three major influences on price elasticities: (1) the extent to which a good is considered to be a necessity; (2) the availability of substitute goods to satisfy a given need; and (3) the proportion of income spent on the product.

The size of the price elasticity of demand is larger the closer and the greater is the number of available substitutes for the commodity. For example, the demand for sugar is more price elastic than the demand for table salt because sugar has better and more substitutes (honey and saccharine) than salt. Thus, a given percentage increase in the price of sugar and salt elicits a larger percentage reduction per time peri9d in the quantity demanded of sugar than of salt.

In general, the more narrowly a commodity is defined, the greater is its price elasticity of demand because the greater will be- the number of substitutes. For example, the price elasticity for Coke is much greater than the price elasticity for soft drinks in general and still larger than the price elasticity .of demand for all beverages. If a commodity is defined so that it has very close substitutes, its price elasticity of demand is likely to be large indeed and may be close to infinity. For example, if a producer of aspirin tried to increase the price above the general range of market prices for aspirin, he would lose a large portion of his sales as buyers can readily switch most of their purchases to competitors who sell similar products.

Similarly, the demand for “big ticket” items such as automobiles, homes, and vacation travel accounts for a large share of consumer income and will be relatively sensitive to price. Demand for less expensive products, such as soft drinks, movies, and candy, can be relatively insensitive to price. Given the low percentage of income spent on “small ticket” items, consumers often find that searching for the best deal available is not worth the time and effort. Accordingly, the elasticity of demand is typically higher for major purchases than for small ones. The price elasticity of demand for CD players, for example, is higher than that for CDs.

CROSS-PRICE ELASTICITY OF DEMAND

The demand for a commodity also depends on the price of related (i.e., substitute and complementary) commodities. On one hand, if the price of tea rises, the demand for coffee increases (i.e., shifts to the right, and more coffee is demanded at each coffee price) as consumers substitute coffee for tea in consumption. On the other hand, if the price of sugar (a complement of coffee) rises, the demand for coffee declines (shifts to the left so that less coffee is demanded at each, coffee price) because the price of a cup of coffee with sugar is now higher.

We can measure the responsiveness in the demand for commodity X to a change in the price of commodity Y with the cross price elasticity of demand (E_{XY}). This is given by the percentage change in the demand for commodity X divided by the percentage change in the price of commodity, Y, holding constant all the other variables in the demand function, including income and the price of commodity X. As with price and income elasticities, we have point and arc cross-price elasticity of demand. Point cross price elasticity of demand is given by

$$\varepsilon_{PX} = \partial Q_Y/Q_Y \div \partial P_X/P_X.$$
$$\varepsilon_{PX} = \partial Q_Y/\partial P_X * P_X/Q_Y$$

If the value of Ep_X is positive, commodities X and Yare substitutes because an increase in P_Y leads to an increase in Q_X as X is substituted for Y in consumption. Examples of substitute commodities are coffee and tea, coffee and cocoa, butter and margarine, hamburgers and .hot dogs, Coca-Cola and Pepsi, and electricity and gas. On the other hand, if E_{XY} is negative, commodities X and Y are complementary be cause an increase in P_Y leads to a reduction in Q_Y and Q_X. Examples of complementary commodities are coffee and sugar, coffee and cream, hamburgers and buns, hotdogs and mustard and cars and gasoline. The absolute value (i.e., the value without the sign) of E_{XY} measures the degree of substitutability and complementarity's between X and Y. For example, if the cross price elasticity of demand between coffee and tea is found to be larger than that between coffee and cocoa, this means that tea is a better substitute for coffee than cocoa. Finally, if E_{XY} is close to zero, X and Yare independent commodities. This may be the case with books and mutton, cars and muffins, pencils and potatoes, and so on.

The cross price elasticity of demand is a very important concept in managerial decision making. Firms often use this concept to measure the effect of changing the price of a product they sell on the demand of other related products that the firm also sells. For example, the Indus Motors can use the cross-price elasticity of demand to measure the effect of changing the price of Toyotas on the demand for Hondas. Since Toyotas and Hondas are substitutes, lowering the price of the former will reduce the demand for the latter. However, a manufacture of both Lipstick and lip-liner can use cross-price elasticity of demand to measure the increase in the demand for Lipstick that would result if the firm reduced the price of lip-liner.

INCOME ELASTICITY OF DEMAND

For many goods, income is another important determinant of demand. Income is frequently as important as price, advertising expenditures, credit terms, or any other variable in the demand function. This is particularly true of luxury items such as big screen televisions, country club memberships, elegant homes, and so on. In contrast, the demand for such basic commodities as salt, bread, and milk is not very responsive to income changes. These goods are bought in fairly constant amounts regardless of changes in income. Of course, income can be measured in many ways—for example, on a per capita, per household or aggregate basis. Gross national product, national income, personal income, and disposable personal income have all served as income measures in demand studies.

NORMAL VERSUS INFERIOR GOODS

The **income elasticity** of demand measures the responsiveness of demand to changes in income, holding constant the effect of all other variables that influence demand. Income and the quantity purchased typically move in the same direction; that is, income and sales are directly rather than inversely related.

$$\varepsilon_I = \partial Q/Q \div \partial I/I.$$

Normal goods have $\varepsilon_I > 0$.
Inferior goods have $\varepsilon_I < 0$.

This does not hold for a limited number of products termed **inferior goods**. Individual consumer demand for such products as beans and potatoes, for example, is sometimes thought to decline as income increases, because consumers replace them with more desirable alternatives. More typical products, whose individual and aggregate demand is positively related to income, are defined as **normal goods**.

One important use of the income elasticity of demand is in forecasting the change in the demand for the commodity that a firm sells under different economic conditions. On one hand, the demand, for a commodity with low-income elasticity will not be greatly affected (i.e., will not

fluctuate very much) as a result of boom conditions or recession in the economy. On the other hand, the demand for a luxury item such as vacations in the Murree Hills or Swat, will increase very much when the economy is booming and fall sharply during recessionary periods.

USING ELASTICITIES IN MANAGERIAL DECISION MAKING

Elasticities are also used in production and cost analysis to evaluate the effects of changes in input on output as well as the effects of output changes on costs. Factors such as price and advertising that are within the control of the firm are called **endogenous variables**. It is important that management know the effects of altering these variables when making decisions. Other important factors outside the control of the firm, such as consumer incomes, competitor prices, and the weather, are called **exogenous variables**. The effects of changes in both types of influences must be understood if the firm is to respond effectively to changes in the economic environment. For example, a firm must understand the effects on demand of changes in both prices and consumer incomes to determine the price cut necessary to offset a decline in sales caused by a business recession (fall in income). Similarly, the sensitivity of demand to changes in advertising must be quantified if the firm is to respond appropriately with price or advertising changes to an increase in competitor advertising.

Thus, the firm should first identify all the important variables that affect the demand for the product it sells. Then the firm should obtain variable estimates of the marginal effect of a change in each variable on demand. The, firm would use this information to estimate the elasticity of demand for the product it sells with respect to each of the variables in the demand function. These are essential for optimal managerial decisions in the short run and in planning for growth in the long run.

For example, suppose that the ABC Company markets coffee brand X and, estimated the following regression of the demand for its brand of coffee:

$$\mathbf{Q_X = 1.5 - 3.0P_X + 0.8I + 2.0P_Y - 0.6P_S - 1.2A}$$

Where Q_X = sales of coffee brand X in the United States, in millions' of pounds per year
P_X = price of coffee brand X, in dollars per pound
I = personal disposable income, in trillions of dollars per year
P_Y = price of the competitive brand of coffee, in dollars per pound
Ps = price of sugar, in dollars per pound
A = advertising expenditures for coffee brand X, in hundreds of thousands of dollars per year
Suppose also that this year, P_X = \$2, I = \$2.5, P_Y = \$1.80, P_S- \$0.50, and A = \$1. Substituting these values into Equation, we obtain.

$$\mathbf{Q_X = 1.5 - 3(2)+ 0.8(2.5) + 2(1.80) - 0.6(0.50) + 1.2(1) = 2}$$

Thus, this year the firm would sell 2 'million pounds of coffee brand X.,

The firm can use this information to find the elasticity of the demand for coffee brand X with respect to its price, income, the price of competitive coffee brand Y, the price of sugar, and advertising. So that:
E_p = -3[2/2] = -3
E_I = 0.8[2.5/2] = 1
E_{xy} = 2[1.8/2] = 1.8
E_{xs} = -0.6[0.5/2] = -0.15
E_A = 1.2[1/2] = 0.6

Lesson 7

DEMAND ESTIMATION

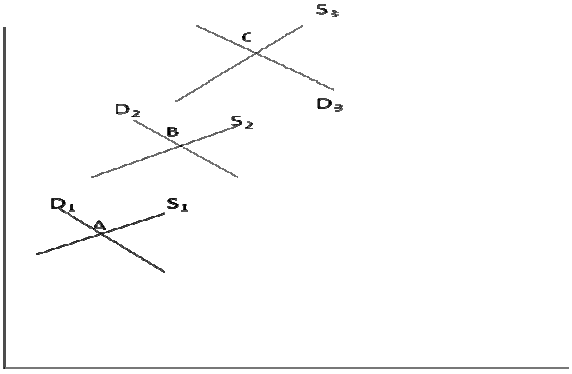
SIMPLE DEMAND CURVE ESTIMATION

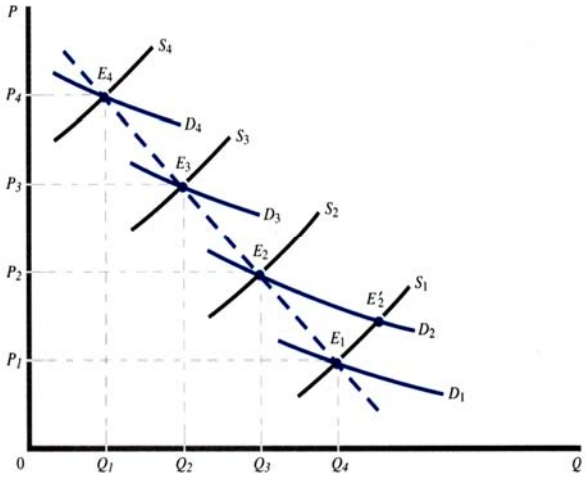
For simple Linear Demand Curves, the best estimation method balances marginal costs and marginal benefits. This means simple linear relations are often useful for demand estimation as straight-line relations can give useful approximations.

THE IDENTIFICATION PROBLEM

The demand curve for a commodity is generally estimated from market data on the quantity purchased of the commodity at various prices over time (i.e., using time series data) or for various consuming units or markets at one point in time (i.e., using cross-section data). However, simply joining the price-quantity observations on a graph does not generate the demand curve for the commodity. The reason is that each price-quantity observation is given by the intersection of a different (but unobserved), demand and supply curve of the commodity. In short, the main problems are:

- Changing Nature of Demand Relations i-e demand relations are dynamic.
- Interplay of Demand and Supply as economic conditions affects demand and supply.
- Shifts in Demand and Supply so that curve shifts can be estimated.
- Simultaneous Relations creates problem as quantity and price are jointly determined. Values for price and quantity are determined & affected by variations in both supply & demand that leads to what is referred to as: “The Simultaneity Problem”. Since both Demand & Supply are shifting the resulting equilibriums do not trace out supply or demand.





In order to derive the demand curve for the commodity from the observed price quantity data points, we should allow the supply curve of the' commodity to shift or to differ, in an unrestricted manner, as shown in above Figure , while we adjust or correct for the shifts or differences in the demand curve, That is, we must adjust or correct for the effect on the demand for the commodity resulting from changes or differences in consumers incomes in the price of related commodities, in consumers' tastes, and in other factors that cause the demand curve of the particular commodity to shift or to be different, so that we can isolate or identify the effect on the quantity demanded of the commodity resulting only from a change in its price. This price - quantity relationship, after correction for all the forces that cause the demand curve, to shift or to be different, gives the true demand curve for the commodity.

By including among the independent or explanatory variables the most important determinants of demand, regression analysis allows the researcher to unravel the independent effects of the various determinant of demand, so as to isolate the effect of the price of the commodity on the quantity demanded of the commodity (i.e., to identify the demand curve for the commodity). Note that nothing is or should be done to correct for shifts or differences in supply. In fact, it is these uncorrected shifts or differences in supply, after having adjusted for shifts or differences in demand, that allow us to derive a particular demand curve. For example, in above Figure point E_2' on demand curve D_2 is derived by correcting the shifts or differences in demand while allowing the supply curve to shift from S_2 to S_1.

MARKETING RESEARCH APPROACH TO DEMAND ESTIMATION

Although Regression analysis is by far the most useful method of estimating demand, marketing research approaches are also used. The most important of these are:

- Consumer Interviews (or survey)
- Consumer Clinic
- Market Experiments

CONSUMER INTERVIEWS OR SURVEYS

Surveys involve questioning a sample of consumers about how they would respond to particular changes in the price of the commodity, incomes, and the price of related commodities, advertising expenditures, credit incentives, and other determinants of demand. These surveys can be conducted by simply stopping and questioning people at a shopping center or by administering sophisticated questionnaires to a carefully constructed representative sample of consumers by trained interviewers.

There are two types of questions that could be a part of a questionnaire.

Specific and closed questions: may be used to obtain specific information and are more generally used in questionnaires. They are seeking Yes or No, or they will ask the respondent to make choices among a set of alternatives given. Such as in KFC and MacDonald use a Lickert scale of 5 about: and the choices are as follows:

a) Food quality
b) Cleanliness
c) Service
d) Atmosphere
e) Staff Behavior

And the choices are as follows:

i. Excellent
ii. Good
iii. Fairly good
iv. Satisfactory
v. Poor

Open-ended questions allow the respondents to answer them in a way they like. An Open-ended question is designed to encourage the consumer to provide an extensive and developmental answer and may be used to reveal attitudes and facts.
In theory, consumer questionnaires can provide a great deal of useful information to the firm. In fact, they are often biased because consumers are either unable or unwilling to provide accurate answers. People are simply very conscious about disclosing their personal information such as age, monthly income and the amount of tax paid per annum.

CONSUMER CLINICS

Another approach to demand estimation is consumer clinics. These are laboratory experiments in which the participants are given a sum of money and asked to spend it in a simulated store to see how they react to changes in the commodity price, product packaging, displays, price of competing products, and other factors affecting demand. Participants in the experiment can be selected so as to closely represent the socioeconomic characteristics of the market of interest. Participants have an incentive to purchase the commodities they want the most because they are usually allowed to keep the goods purchased. Thus, consumer clinics are more realistic than consumer surveys.

Consumer clinics also face serious shortcomings, however. First, the results are questionable because participants know that they are in an artificial situation and that they are being observed. Therefore, they are not likely to act normally, as they would in a real market situation.

MARKET EXPERIMENTS

Unlike consumer clinics, which are conducted under strict laboratory conditions, market experiments are conducted in the actual marketplace. There are many ways of performing market experiments. One method is to select several markets with similar socioeconomic characteristics and change the commodity price in some markets or stores, packaging in other markets or stores, and the amount and type of promotion in still other markets or stores, then record the responses (purchases) of consumers in the different markets. By using census data or surveys for various markets, a firm can also determine the effect of age, gender, level of education, income, family size, and so forth on the demand for the commodity.

The advantages of market experiments are that they can be conducted on a large scale to ensure the validity of the results and that consumers are not aware that they are part of an experiment. Market experiments also have serious disadvantages, however. One of these is

that in order to keep costs down, the experiment is likely to be conducted on too limited a scale and over a fairly short period of time, so that inferences about the entire market and for a more extended period of time are questionable. Extraneous occurrences, such as a strike or unusually bad weather, may seriously bias the results in uncontrolled experiments. Competitors could try to sabotage the experiment by also changing prices and other determinants of demand under their control. They could also monitor the experiment and gain useful information that the firm would prefer not to disclose. Finally, a firm might permanently lose customers in the process of raising prices in the market where it is experimenting with a high price.

Despite these shortcomings, market experiments may be useful to a firm in determining its best pricing strategy and in testing different packaging, promotional campaigns, and product qualities. Market experiments are particularly useful in the process. If introducing a product for which no other data exist. They may also be uses of other statistical techniques used to estimate demand and in providing some of the data required for these other statistical techniques of demand estimation.

REGRESSION ANALYSIS

To understand when the use of regression analysis is appropriate, one must appreciate a basic difference between two broad classes of economic relations. A **deterministic relation** is one that is known with certainty. For example, total profit equals total revenue minus total cost, or $\pi = TR - TC$. Once the levels of total revenue and total cost are known with certainty, total profits can be exactly determined. The profit relation is an example of a deterministic relation.

A **statistical relation** exists between two economic variables if the average of one is related to another, but it is impossible to predict with certainty the value of one based on the value of another. In the earlier example, if $TC = \$10Q$ *on average*, then a one-unit increase in quantity would tend to result in an average $10 increase in total cost. Sometimes the actual increase in total cost would be more than $10; sometimes it would be less. In such circumstances, a statistical relation exists between total costs and output. When a statistical relation exists, the exact or “true” relation between two economic variables is not known with certainty and must be estimated. Perhaps the most common means for doing so is to gather and analyze historical data on the economic variables of interest.

A **time series** of data is a daily, weekly, monthly, or annual sequence of data on an economic variable such as price, income, cost, or revenue. To judge the trend in profitability over time, a firm would analyze the time series of profit numbers. A **cross section** of data is a group of observations on an important economic variable at any point in time. If a firm were interested in learning the relative importance of market share versus advertising as determinants of profitability, it might analyze a cross section of profit, advertising, and market share data for a variety of regional or local markets. To assess the effectiveness of a quality management program, the firm might consider both time-series and cross-section data.

The simplest and most common means for analyzing a sample of historical data is to plot and visually study the data. A **scatter diagram** is a plot of data where the *dependent* variable is plotted on the vertical or *Y*-axis, and the *independent* variable is plotted on the horizontal or *X*-axis. The following Figure shows scatter diagrams that plot the relation between the demand and four different factors that have the potential to influence demand.

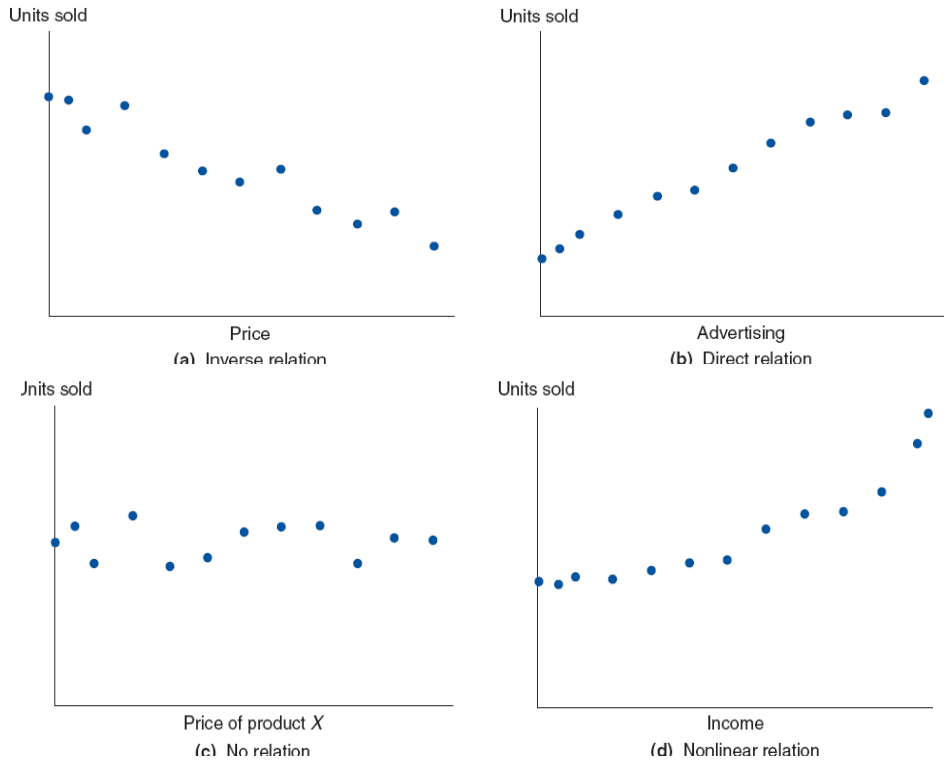


Figure (a) depicts an inverse relation between the quantity sold and price, the dependent variable. In Figure (b), a direct relation between the amount of advertising and demand is shown. No relation is evident between demand and the price of product X (an unrelated product). In panel (d), a nonlinear relation between demand and income is illustrated. Scatter diagrams are analyzed to gain an instinctive "feel" for the data. The method is entirely inductive and intuitive. Although the examination of scatter diagrams has a categorical value as a starting point in the analysis of simple statistical relations, its lack of structure can also limit its value.

Lesson 8

DEMAND ESTIMATION (CONTINUED 1)

REGRESSION ANALYSIS

The term **"regression"** was first used by the geneticist, Francis Galton (1886) who noted that tall fathers have shorter sons and short fathers have taller sons. He described this phenomenon as regression where height tended to "regress" towards the mean height for the population.

THE MODERN INTERPRETATION OF REGRESSION

Regression Analysis is a powerful statistical technique that describes the way in which one important economic variable is related to one or more economic variables.

STEPS IN REGRESSION ANALYSIS

(1) Specify **variables**: Quantity Demanded, Advertising, Income, Price, Other prices, Quality, Previous period demand...
(2) Obtain **data:** Cross sectional vs Time series
(3) Specify functional form of **equation**

Linear $Y_t = a + b X_{1t} + g X_{2t} + u_t$
Multiplicative $Y_t = a X_{1t}^{b} X_{2t}^{g} e_t$
$\ln Y_t = \ln a + b \ln X_{1t} + g \ln X_{2t} + u_t$

(4) Estimate **parameters**
(5) Interpret **results**: economic and statistical

FUNCTIONAL FORM SPECIFICATIONS

Linear Function: Linear Model
Estimation Format:

$$Q_X = a_0 + a_1 P_X + a_2 I + a_3 N + a_4 P_Y + \cdots + e$$

Power Function: $Q_X = a(P_X^{b_1})(P_Y^{b_2})$

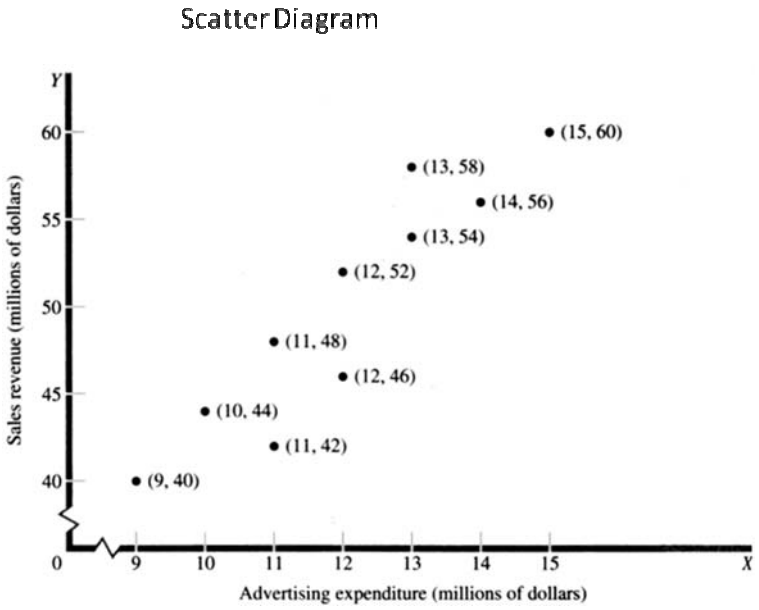
Estimation Format: $\ln Q_X = \ln a + b_1 \ln P_X + b_2 \ln P_Y$

INTRODUCTION TO REGRESSION ANALYSIS

In order to introduce regression analysis; suppose that a manager wants to determine the relationship between the firm's advertising expenditures and its sales revenue. The manager wants to test the hypothesis that higher advertising expenditures lead to higher sales for the firm, and, furthermore, she wants to- estimate the strength of the relationship (i.e., -how much sales increase for each dollar increase in advertising expenditures). To this end, the manager collects data on advertising expenditures and on sales revenue for the firm over the past 10 years. In this case, the level of advertising expenditures (X) is the independent or explanatory variable, while sales revenues (Y) is the dependent variable that the manager seeks to explain. Suppose that the advertising-sales data for the firm in each of the past 10 years that the manager has collected are those in Table.

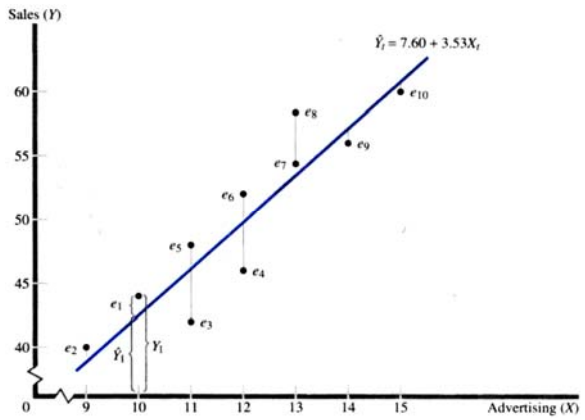
Advertising Expenditures and Sales Revenues Of the Firm (in millions of dollars)

Year	X	Y
1	10	44
2	9	40
3	11	42
4	12	46
5	11	48
6	12	52
7	13	54
8	13	58
9	14	56
10	15	60



If we now plot each pair of advertising-sales values in the above Table as a point on a graph, with advertising expenditures (the independent or explanatory variable) measured along the horizontal axis and sales revenues (the dependent variable) measured along the vertical axis, we get the points (dots) in Figure. This is known as a scatter diagram since it shows the spread of the points in the X - Y plane. From the above Figure (scatter diagram), we see that there is a positive relationship between the level of the firm's advertising expenditures ,and its sales revenues (i.e., higher advertising expenditures are associated with higher sales revenues) and that this relationship is approximately linear.

One way to estimate the approximate linear relationship between the firm's advertising expenditures and its sales revenues is to draw in, by visual inspection, the positively sloped straight line that “best” fits between the data points (so that the data points are about equally distant on either side of the line). By extending the line to the vertical axis, we can then estimate the firm's sales revenues with zero advertising expenditure. The slope of the line will then provide an estimate of the increase in the sales revenues that the firm can expect with each $1 million increase in its advertising expenditures. This will give us a rough estimate of the linear relationship between the firm's sales revenues (Y) and its advertising expenditures (X) in the form of Equation:



The difficulty with the visual fitting of a line to the data points in Figure is that different researchers would probably fit a somewhat different line to the same data points and obtain somewhat different results. Regression analysis is a statistical technique for obtaining the line that best fits the data points according to an objective statistical criterion, so that all researchers looking at the same data would get exactly the same result (i.e., obtain the same line). Specifically, the regression line is the line obtained by minimizing the sum of the squared vertical deviations of each point from the regression line. This method is, therefore, appropriately called the **"ordinary least-squares method or OLS"** in short. The regression line fitted by such a least-squares method is shown in the above Figure.

In this Figure, Y_1 refers to the actual or observed sales revenue of $44 million associated with the advertising expenditures of $10 million in the first year for which the data were collected (see Table). The $\hat{Y}_1$ (reads: Y hat sub 1) shown in the figure is the corresponding sales revenues of the firm estimated from the regression line for the advertising expenditure of $10 million in the first year. The symbol e_1 in the figure is then the corresponding vertical deviation or error of the actual or observed sales revenue of the firm from the sales revenue estimated from the regression line in the first year.

ORDINARY LEAST SQUARES (OLS) MODEL

$$Y_t = a + bX_t + e_t$$

$$\hat{Y}_t = \hat{a} + \hat{b}X_t$$

$$e_t = Y_t - \hat{Y}_t$$

Since there are 10 observation points in our Figure, we have 10 such vertical deviations or errors. These are labeled e_1 to e_{10} in the figure. The regression line shown in Figure is the line that best fits the data points in the sense that the sum of the squared (vertical) deviations from the line is minimum. That is, each of the 10 e values is first squared and then summed. The regression line is the line for which the sum of these squared deviations is a minimum.

Errors arises because

1. Various Explanatory variables are absent
2. Possible errors of measurement in Y
3. Random human behavior that leads to different results under identical conditions

Simple Regression Analysis

1. Calculate the value of *a* and *b*
2. Tests of significance of parameter estimates
3. Confidence interval for the true parameter
4. Overall explanatory power of the regression

The objective of OLS is to determine the slope and intercept that minimize the sum of the squared errors. Thus objective of regression analysis is to obtain the estimates of *a* (the vertical intercept) and *b* (the slope) of the regression line.

$$\sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2 = \sum_{t=1}^{n} (Y_t - \hat{a} - \hat{b}X_t)^2$$

Where $\Sigma^n_{t=1}$ is the sum of all observations, from time period t = 1 to t = n. The estimated values of a and b (that is, â and b) are obtained by minimizing the sum of the squared deviations (i.e., by minimizing the value of Equation).The value of b is given by:

$$\hat{b} = \frac{\sum_{t=1}^{n} (X_t - \overline{X})(Y_t - \overline{Y})}{\sum_{t=1}^{n} (X_t - \overline{X})^2}$$

Where Y' and X' are the mean or average values of the Y_t and the X_t, respectively. The value of a is then obtained from

$$\hat{a} = \overline{Y} - \hat{b}\overline{X}$$

Estimation Example

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Year	**X_t**	**Y_t**	**Y'_t**	**$Y_t - Y'_t = e_t$**	**$(Y_t - Y'_t)^2 = e^2_t$**	**$(X_t - X'_t)^2$**
1	10	44	42.90	1.10	1.2100	4
2	9	40	39.37	0.63	0.3969	9
3	11	42	46.43	-4.43	19.6249	1
4	12	46	49.96	-3.96	15.6816	0
5	11	48	46.43	1.57	2.4649	1
6	12	52	49.96	2.04	4.1616	0
7	13	54	53.49	0.51	0.2601	1
8	13	58	53.49	4.51	20.3401	1
9	14	56	57.02	-1.02	1.0404	4
10	15	60	60.55	-0.55	0.3025	9
n= 10	**$\sum X_t$=120**	**$\sum Y_t$=500**			**$\sum e^2_t$=65.4830**	**$\sum(X_t-X')^2$=30**
	X'=12	**Y'=50**				

$$n=10$$

$$\overline{X} = \sum_{t=1}^{n} \frac{X_t}{n} = \frac{120}{10} = 12$$

$$\overline{Y} = \sum_{t=1}^{n} \frac{Y_t}{n} = \frac{500}{10} = 50$$

$$\sum_{t=1}^{n} X_t = 120 \quad \sum_{t=1}^{n} Y_t = 500$$

$$\sum_{t=1}^{n} (X_t - \bar{X})^2 = 30$$

$$\sum_{t=1}^{n} (X_t - \bar{X})(Y_t - \bar{Y}) = 106$$

$$\hat{b} = \frac{106}{30} = 3.533$$

$$\hat{a} = 50 - (3.533)(12) = 7.60$$

Thus, the value of s_b, is equal to

$$S_b = \frac{\Sigma(Y_t - Y_t)^2}{(n-k)\Sigma(X_t - X')^2} = \frac{65.4830}{(10-2)(30)} = 0.2728 = 0.52$$

Having obtained the value of s_b, we next calculate the ratio b'/s_b. This is called the t statistic or t ratio. The higher this calculated t ratio is, the more confident we are that the true but unknown value of b that we are seeking is not equal to zero (i.e., that there is a significant relationship between advertising and sales). For our sales-advertising example, we have

TESTS OF SIGNIFICANCE: CALCULATION OF THE T STATISTIC

$$t = \frac{\hat{b}}{s_{\hat{b}}} = \frac{3.53}{0.52} = 6.79$$

Degrees of Freedom = (n-k) = (10-2) = 8
Critical Value at 5% level = 2.306

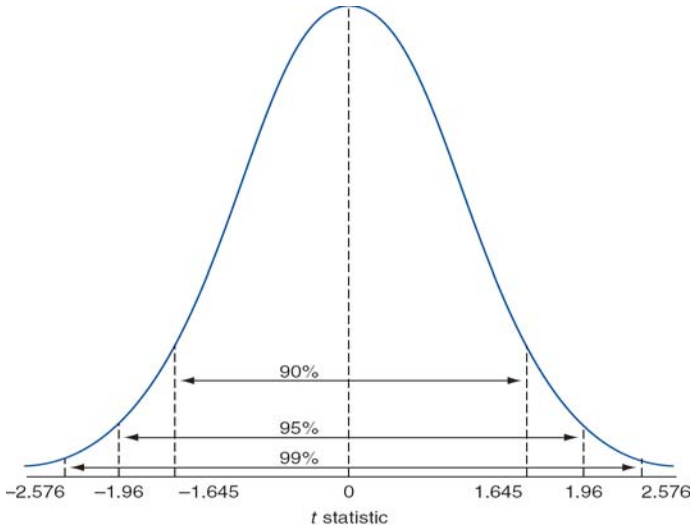
In order to conduct an objective or significance test for b, we compare the calculated t ratio to the critical value of the t distribution with n - k = 10 - 2 = 8 df given by Table (*t* distribution). This ***t*** test of the statistical significance of the estimated coefficient is usually performed at the 5 percent level of significance. Thus, we go down the column headed 0.05 (referring to 2.5 percent of the area or probability in each tail of the t distribution, for a total of 5 percent in both tails) in Table until we reach 8 df. This gives the critical value of t = 2.306 for this two-tailed t test.

Since our calculated value of t = 6.79 exceeds the tabular value of t = 2.306 for the 5 percent level of significance with 8 df, we reject the null hypothesis that there is no relationship between X (advertising) and Y (sales) and accept the alternative hypothesis that there is in fact a significant relationship between X and Y. To say that there is a statistically significant relationship between X and Y at the 5 percent level means that we are 95 percent confident that such a relationship exists. In other words, there is less than 1 chance in 20 (i.e., less than 5 percent chance) of being wrong or accepting the hypothesis that there is, significant relationship between X and Y, when in fact there isn't.
The above concepts can also be used to determine confidence intervals for the true b

coefficient. Thus, using the tabular value of t = 2.306 for the 5 percent level of significance (2.5 percent in each tail) and 8 df in our advertising sales example, we can say that we are 95 percent confident that the true value of b will be between

$$b' \pm 2.306 (s_b)$$
$$3.53 \pm 2.306 (0.52)$$
$$3.53 \pm 1.20$$

That is, we are 95 percent confident that the true value of b lies between 2.33 and 4.73. Similarly, we can say that we are 99 percent confident that the true value of b will be between 3.53 $\pm$ 3.355(0.52), or. 1.79 and 5.27 (the value of t = 3.355 is obtained by going down the column headed 0.01 in Table until we reach 8 dt).



Lesson 9

DEMAND ESTIMATION (CONTINUED 2)

ASSUMPTIONS OF REGRESSION ANALYSIS

Regression Analysis is based on a number of assumptions: these are that the error terms are

- Normally distributed,
- Has zero expected value or mean,
- Has constant variance in each time period and for all values of X i-e (homosceedasticity) equal variance,
- Its value in one time period is unrelated o its value in another period i-e no autocorrelation between any two error terms,
- There is no perfect multicollinearity i-e there are no perfect linear relationship among the explanatory variables.

SPECIFYING THE REGRESSION MODEL

The first step in regression analysis is to specify the variables to be included in the regression equation or model. Product demand, measured in physical units, is the dependent variable when specifying a demand function. The list of independent variables, or those that influence demand, always includes the price of the product and generally includes such factors as the prices of complementary and competitive products, advertising expenditures, consumer incomes, and population of the consuming group. Demand functions for expensive durable goods, such as automobiles and houses, include interest rates and other credit terms; or air conditioners include weather conditions.

The second step in regression analysis is to obtain reliable data. Data must be gathered on total output or demand, measures of price, credit terms, capacity utilization ratios, wage rates, and the like. Obtaining accurate data is not always easy, especially if the study involves time series data over a number of years. Moreover, some key variables may have to be estimated.

Once variables have been specified and the data have been gathered, the functional form of the regression equation must be determined. This form reflects the way in which independent variables are assumed to affect the dependent or *Y* variable. The most common specification is a **linear model**, such as the following demand function:

$$Q = a + bP + cA + dI$$

Here *Q* represents the unit demand for a particular product, *P* is the price charged, *A* represents advertising expenditures, and I is per capita disposable income. Unit demand is assumed to change in a linear fashion with changes in each independent variable. Another common regression model form is the **multiplicative model**:

$$Q = aPbAcId$$

A multiplicative model is used when the marginal effect of each independent variable is thought to depend on the value of all independent variables in the regression equation. For example, the effect on quantity demanded of a price increase often depends not just on the price level, but also on the amount of advertising, competitor prices and advertising, and so on.

TEST OF GOODNESS OF FIT AND CORRELATION

Overall Explanatory Power of the entire Regression

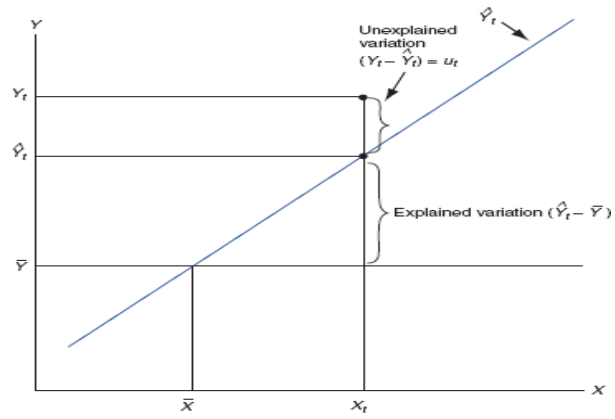
- Coefficient of Determination R^2

- Decomposition of Sum of Squares

TOTAL VARIATION = EXPLAINED VARIATION + UNEXPLAINED VARIATION

$$\sum (Y_t - \bar{Y})^2 = \sum (\hat{Y} - \bar{Y})^2 + \sum (Y_t - \hat{Y}_t)^2$$

Besides testing for the statistical significance of a particular estimated parameter, we can also test for the overall explanatory power of the entire regression. This is accomplished by calculating the coefficient of determination, which is usually denoted by R^2. The coefficient of determination (R^2) is defined as the proportion of the total variation or dispersion in the dependent variable (about its mean) that is explained by the variation in the independent or explanatory variable(s) in the regression. In terms of our advertising sales example, R^2.measures how much of the variation in the firm's sales is explained by the variation in its advertising expenditures.



We can calculate the coefficient of determination (R^2) by defining the total, the explained, and the unexplained or residual variation in the dependent variable, Y. The total variation in Y can be measured by squaring the deviation of each observed value of Y from its me art and then summing. That is,

Total variation in Y = $^{n}\sum_{t=}1(Y_t - Y')^2$

Regression analysis breaks up this total variation in Y into two components: the variation in Y that is explained by the independent variable (X) and the unexplained or residual variation in Y. The explained variation in Y is given by Equation.

Explained variation in Y = $n\sum t=1(Y_t - Y')^2$

The values of Y in Equation are obtained by substituting the various observed values of X (the independent variable) into the estimated regression equation. The mean of Y (Y') is then subtracted from each of the estimated values of Y_t (Y_t). As in dicated by Equation, these differences are then squared and added to get the explained variation in Y

Finally, the unexplained variation in Y is given by Equation

Unexplained variation in Y = $^{n}\sum_{t=1} (Y_t - Y'_t)^2$

That is, the unexplained or residual variation in Y is obtained by first subtracting from each observed value of Y the corresponding estimated value of Y, and then squaring, and summing.

Now, the coefficient of determination, R^2, is defined as the ratio of the explained variation in Y to the ,total variation in Y. That is,

$$\mathbf{R^2 = \frac{\text{Explained variation in Y}}{\text{Total variation in Y}} = \frac{\sum(Y_t - Y'_t)^2}{\sum(Y_t - Y')^2}}$$

Calculations to Estimate the Coefficient of Determination (R^2)							
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Year	Y_t	Y_t-Y'	$(Y_t-Y')^2$	Y'_t	Y'_t-Y'	$(Y'_t-Y')^2$	$(Y_t-Y'_t)^2$
1	44	-6	36	42.90	-7.10	50.4100	1.2100
2	40	-10	100	39.37	-10.63	112.9969	0.3969
3	42	-8	64	46.43	-3.57	12.7449	19.6249
4	46	-4	16	49.96	-0.04	0.0016	15.6816
5	48	-2	4	46.43	-3.57	12.7449	2.4649
6	52	2	4	49.96	-0.04	0.0016	4.1616
7	54	4	16	53.49	3.49	12.1801	0.2601
8	58	8	64	53.49	3.49	12.1801	20.3401
9	56	6	36	57.02	7.02	49.2804	1.0404
10	60	10	100	60.55	10.55	111.3025	0.3025
n=10	**$\sum Y_t$=500 Y'=50**		**$\sum(Y_t-Y')^2$ =440**			**$\sum(Y_t-Y')^2$ =373.8430**	**$\sum(Y_t-Y')^2$ =65.4830**

From the bottom of column 4, we see that the total variation in Y (sales) is $440 million. The explained variation is $373.84 million, as shown at the bottom of column 7. Thus, the coefficient of determination for our advertising sales problem is

$$R^2 = \frac{\$373.84}{\$440} = 0.85$$

This means that 85 percent of the total variation in. the firm's sales is accounted for by the variation in the firm's advertising expenditures.

From the bottom of column 4, we see that the total variation in Y (sales) is $440 million. The explained variation is $373.84 million, as shown at the bottom of column 7. Thus, the coefficient of determination for our advertising sales problem is

$$R^2 = \frac{\$373.84}{\$440} = 0.85$$

This means that 85 percent of the total variation in. the firm's sales are accounted for by the variation in the firm's advertising expenditures.

COEFFICIENT OF CORRELATION

in simple regression analysis the square root of the coefficient of determination (R^2) is the (absolute value of the) coefficient of correlation, which is denoted by r. That is,

$$\mathbf{r = \sqrt{R^2}}$$

This is simply a measure of the degree of association or co variation that exists between variables X and Y. For our advertising-sales example,

$$\mathbf{r = \sqrt{R^2} = \sqrt{0.85} = 0.92}$$

This means that, variables X and Y vary together 92 percent of the time. The Coefficient of Correlation (*r*) is a measure of the strength of the relationship between two variables.

- It requires interval or ratio-scaled data.
- It can range from -1.00 to 1.00
- Values of -1.00 or 1.00 indicate perfect and strong correlation.
- Values close to 0.0 indicate weak correlation.
- Negative values indicate an inverse relationship and positive values indicate a direct relationship.

Equation can be transformed into a linear relation using logarithms and then estimated by the least squares technique. Thus, Equation 3.13 is equivalent to

$$\log Q = \log a + b \_ \log P + c \_ \log A + d \_ \log I$$

Given the multiplicative or log-linear form of the regression model, these coefficient estimates can also be interpreted as estimates of the constant elasticity of Y with respect to X, or the percentage change in Y due to a one percent change in X. Much more will be said about elasticity later in the book, but for now it is worth noting that multiplicative or log-linear models imply constant elasticity.

THE MULTIPLE REGRESSION MODEL

When the dependent variable that, we seek to explain is hypothesized to depend on more than one independent or explanatory variable, we have multiple regression analysis. For example, the firm's sales revenue may be postulated to depend not only on the firm's advertising expenditures (as examined in Section) but also on its expenditures on quality control the regression model can then be written as:

$$Y=a+b_1X_1+b_2X_2$$

Where Y is the dependent variable referring to the firm's sales revenue, X_1 refers to the firm's advertising expenditures, and X_2 refers to its expenditures on quality control. The coefficients a, b_1 and b_2 are the parameters to be estimated.

In our sales-advertising and quality control problem we postulate that both b_1 and b_2 are positive, or that the firm can increase its sales by increasing its expenditures for advertising and quality control.

The model can also be generalized to any number of independent or explanatory variables (k'), as indicated in Equation:

$$Y = a+b_1X_1+b_2X_2+\ldots\ldots+b'_kX_k'$$

The only assumptions made in multiple regression analysis in addition to those made for simple regression analysis are that the number of independent or explanatory variables in the regression be smaller than the number of observations and that there be no perfect linear correlation among the independent variables.

Yearly Expenditures on Advertising and Quality Control, and Sales of the Firm (in millions of dollars)										
Year (t)	1	2	3	4	5	6	7	8	9	10
Advertising (X_1)	10	9	11	12	11	12	13	13	14	15
Quality control (X_2)	3	4	3	3	4	5	6	7	7	8

Sales revenue (Y)	44	40	42	46	48	52	54	58	56	60

The process of estimating the parameters or coefficients of a multiple regression equation is, in principle, the same as in simple regression analysis, but since the calculations are much more complex and time consuming, they are always done with computers. Since determining b_1, b_2, is very tedious, a software package such as TSP, SPSS, Excel or MINITAB is recommended.
Using computer, For example, if we regress the firm's sales (Y) on its expenditures for advertising (X_1) and quality control (X_2) using the data in above Table, We can write the following regression equation:

$Y'_t = 17.944 + 1.873X_{1t} + 1.915X_{2t}$
t statistic (2.663) (2.813)

The critical value of *t @ 5%* α, 7df is: 2.365, we conclude that both parameters are statistically different from zero.

ADJUSTED COEFFICIENT OF DETERMINATION

$$\overline{R}^2 = 1 - (1 - R^2)\frac{(n-1)}{(n-k)}$$

However, in order to take into consideration that the number of degrees of freedom declines as additional independent or explanatory variables are included in the regression, we calculate the adjusted R^2. Where n is the number of observations or sample data points and k is the number of parameters or coefficients estimated.

THE *F* STATISTIC
Both the coefficient of determination, R2, and corrected coefficient of determination, – R2, provide evidence on whether or not the proportion of explained variation is relatively “high” or “low.” However, neither tells if the independent variables as a group explain a statistically significant share of variation in the dependent Y variable. The F statistic provides evidence on whether or not a statistically significant proportion of total variation in the dependent variable has been explained.
The value of the F statistic is give by

$$F = \frac{\text{Explained variation} / (k-1)}{\textbf{Total variation} / (n-k)}$$

Where, as usual, .n is the number of observations and k is the number of estimated parameters or coefficients in the regression. It is because the F statistic is the ratio of two variances that this test is often referred to as the "analysis of variance." The F statistic can also be calculated in terms of the coefficient of determination as follow:

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

Using the values of $R^2 = 0.930154$, n = 10, and k = 3 for our example, we obtain:

- F = 46.61 : Calculated Value
- F = 4.74 : Critical Value @ 5% level

The F test is used to determine whether a given F statistic is statistically significant. Performing

F tests involves comparing F statistics with critical values from a table of the F distribution. If a given F statistic exceeds the critical value from the F distribution table, the hypothesis of no relation between the dependent Y variable and the set of independent X variables can be rejected.

PROBLEMS IN REGRESSION ANALYSIS

- Multicollinearity: Two or more explanatory variables are highly correlated.
- Heteroskedasticity: Variance of error term is not independent of the Y variable.
- Autocorrelation: Consecutive error terms are correlated.

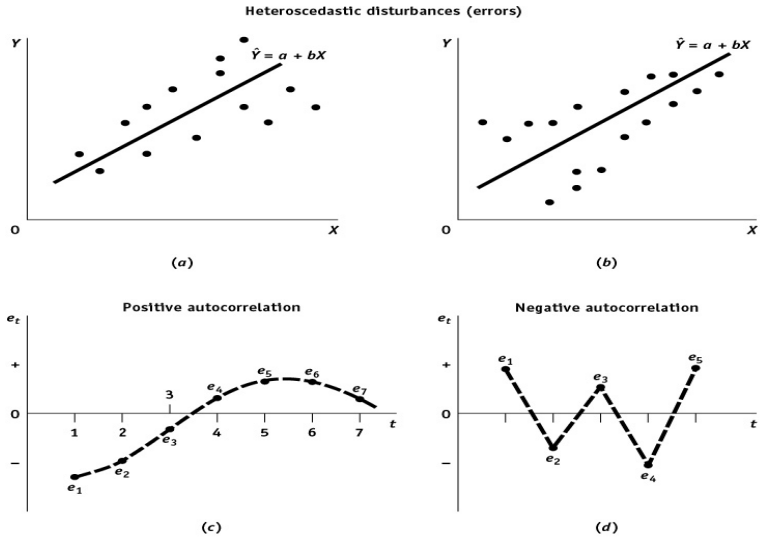


FIGURE 4-5 Heteroscedastic and Autocorrelated Disturbances Part (*a*) shows heteroscedastic disturbances, where the size of the error or residual increases with the size of the value of *X*. Part (*b*) shows the opposite pattern of heteroscedastic disturbances (which is less common). Part (*c*) shows positive autocorrelation (i.e., a positive or negative error in one period is followed by another positive or negative error term, respectively, in the following period). Part (*d*) shows negative autocorrelation (which is less common).

DURBIN-WATSON STATISTIC

- Test for Autocorrelation
- *d* ranges between 0 and 4

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

If d = 2, autocorrelation is absent.

Lesson 10

DEMAND FORECASTING

FORECASTING

The field of organizational forecast, born in the 1950s, is reaching maturity. Most business decisions are made in the face of risk or uncertainty. Accurate business forecasting is a value-added undertaking. A firm must decide how much of each product to produce, what price to charge, and how much to spend on advertising, and it must also plan for the growth of the firm. All these decisions are based on some forecast of the level of future economic activity in general and demand for the firm's product in particular. The aim of economic forecasting is to reduce the risk or uncertainty that the firm faces in its short-term operational decision making and in planning for its long-term growth. The accuracy of any forecast is subject to the influence of controllable and uncontrollable factors.

MACROECONOMIC APPLICATIONS

Predictions of economic activity at the national or international level, e.g., inflation or employment, involves predicting aggregate measures of economic activity at the international, national, regional, or state level. Predictions of gross domestic product (GDP), unemployment, and interest rates by “blue chip” business economists capture the attention of national media, business, government, and the general public on a daily basis.

MICROECONOMIC APPLICATIONS

Predictions of company and industry performance, e.g., business profits, forecasting the demand and sales of the firm's product usually begins with a macroeconomic forecast of the general level of economic activity for the economy as a whole, or gross national product. The reason for this is that the demand and sales of most goods and services are strongly affected by business conditions. For example, the demand and sales of new automobiles, new houses, electricity, and most other goods and services rise and follow with the general level of economic activity. The firm uses these macro-forecast of general economic activity as inputs for their micro-forecasts of the industries and, firm's demand and sales.

In contrast with macroeconomic forecasting, microeconomic forecasting involves the prediction of disaggregate economic data at the industry, firm, plant, or product level. Unlike predictions of GDP growth, which are widely followed in the press, the general public often ignores microeconomic forecasts of the demand for new cars, or production costs for cosmetic prices.

FORECASTING TECHNIQUES

Approaches to forecasting

- **Qualitative forecasting** is based on judgments expressed by individuals or groups
- **Quantitative forecasting** utilizes significant amounts of data and equations

Quantitative techniques can be naïve or causal. **Naïve forecasting** projects past data into the future without explaining future trends. **Causal or explanatory** forecasting attempts to explain the functional relationships between the variable to be estimated (the dependent variable) and the variable or variables that are responsible for the changes (the independent variable).

The most commonly applied forecasting techniques can be divided into the following broad categories:

- Qualitative analyses
- Time-Series Analysis
- Trend analysis and projection

- Exponential smoothing
- Econometric methods

The best forecast methodology for a particular task depends on the nature of the forecasting problem. When making a choice among forecast methodologies, a number of important factors must be considered. It is always worth considering the distance into the future that one must forecast, the lead time available for making decisions, the level of accuracy required, and the quality of data available for analysis, the stochastic or deterministic nature of forecast relations, and the cost and benefits associated with the forecasting problem.

QUALITATIVE ANALYSIS

Qualitative Analysis includes:

1. Expert Opinion or Opinion Poll
2. Survey Techniques

.

1. EXPERT OPINION OR OPINION POLL

Expert Opinion or Opinion Poll can be further categorized as follows:

i. **Executive polling or expert opinion:** The firm can poll its top management from its sales, production, finance, and personnel departments on their views on the sales outlook for the firm during the next quarter or year. Although these personal insights are to a large extent subjective, by averaging the opinions of the experts who are most knowledgeable about the firm and its products, the firm hopes to arrive at a better forecast than would be provided by these experts individually. The most basic form of qualitative analysis forecasting is **personal insight**, in which an informed individual uses personal or company experience as a basis for developing future expectations. Although this approach is subjective, the reasoned judgment of informed individuals often provides valuable insight. When the informed opinion of several individuals is relied on, the approach is called forecasting through **panel consensus**. The panel consensus method assumes that several experts can arrive at forecasts that are superior to those that individuals generate. Direct interaction among experts can help ensure that resulting forecasts embody all available objective and subjective information. Although the panel consensus method often results in forecasts that represent the collective wisdom of consulted experts, it can be unfavorably affected by the forceful personality of one or a few key individuals. A related approach, the **delphi method**, has been developed to counter this disadvantage. In the delphi method, members of a panel of experts individually receive series of questions relating to the underlying forecasting problem. Responses are analyzed by an independent party, who then tries to draw a consensus opinion by providing feedback to panel members in a manner that prevents direct identification of individual positions. To avoid a bandwagon effect (whereby the opinions of some experts might be overshadowed by some dominant personality in their midst), the Delphi method is used.

ii. **Sales force polling**: This is a forecast of the firm's sales in each region and for each product line; it is based on the opinion of the firm's sales force in the field. These are the people closest to the market, and their opinion of future sales can provide valuable information to the firm's top management.

iii. **Consumer intentions polling**: Companies selling automobiles, furniture, household appliances, and other durable goods sometimes poll a sample of potential buyers on their purchasing intentions. Using the results of the poll, the firm can forecast its national sales for different levels of consumers' future disposable Income.

2. SURVEY TECHNIQUES

Survey techniques that skillfully use interviews or mailed questionnaires are an important forecasting tool, especially for short-term projection. Designing surveys that provide unbiased and reliable information is a challenging task. When properly carried out, however, survey research can provide managers with valuable information that would otherwise be unobtainable. Surveys generally use interviews or mailed questionnaires that ask firms, government agencies, and individuals about their future plans. Businesses plan and budget virtually all their expenditures in advance of actual purchase or production decisions. Surveys asking about capital budgets, sales budgets, and operating budgets can thus provide useful forecast information. Government departments that prepare formal budgets also provide a wealth of information to the forecaster. Finally, because individual consumers routinely plan expenditures for such major items as automobiles, furniture, housing, vacations, and education, surveys of consumer intentions often accurately predict future spending on consumer goods. Survey information may be all that is available in certain forecasting situations, as, for example, when a firm is attempting to project new product demand. Although surveys sometimes serve as an alternative to quantitative forecasting techniques, they frequently supplement rather than replace quantitative analysis. Quantitative models generally assume stable consumer tastes. If tastes are actually changing, survey data can suggest the nature and direction of such changes.

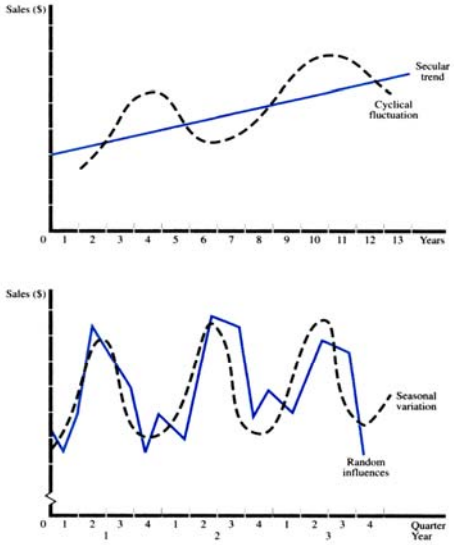
TIME SERIES ANALYSIS

Time series analysis: a naïve method of forecasting from past data by using least squares statistical methods to identify trends, cycles, seasonality and irregular movements. A Time Series is a collection of data recorded over a period of time. The data may be recorded weekly, monthly, or quarterly. This is one of the most frequently used forecasting methods. This method attempts to forecast future values of time series by examining past observations of the data only, on the assumption that the time series will continue to move in the past pattern.

COMPONENTS OF A TIME SERIES

There are four components to a time series:

1. The Secular Trend
2. The Cyclical Variations
3. The Seasonal Variations
4. The Irregular Variations



If we plot most economic time-series data, we discover that they fluctuate or vary over time. This variation is usually caused by secular trends, cyclical fluctuations, seasonal variations, and irregular or random influences. These sources of variation are shown in the above Figure and are briefly explained below:

1. Secular trend refers to a long-run increase or decrease in the data series (the straight solid line in the top panel of Figure). For example, many time series of sales shows rising trends over the years because of population growth and increasing per capita expenditures. Some, such as typewriters, follow a declining trend as more and more consumers switch to personal computers and from personal computers to laptops.

2. Cyclical fluctuations are the major expansions and contractions in most economic time series that seem to recur every several years (the heavy dashed curved line in the top panel of Figure). For example, the housing construction industry follows long cyclical swings lasting 15 to 20 years, while the automobile industry seems to follow much shorter cycles.

3. Seasonal variation refers to the regularly recurring fluctuation in economic activity during each year (the heavy dashed curved line in the bottom panel of Figure) because of weather and social customs. Thus, housing starts used to be much more common in spring and summer than in autumn and winter (because of weather conditions), while retail sales are greatest during the second and last quarter of each year because of weather conditions again. Seasonal variation is a rhythmic annual pattern in sales or profits caused by weather, habit, or social custom such as Basant or religious festivities such as Eid-ul-fitr, Eid-ul-Azha and Ramazan.

4. Irregular or random influences are the variations in .the data series resulting from wars, natural disasters, strikes, or other unique events. These are shown by the solid line segments in the bottom panel of Figure.

TIME SERIES ANALYSIS:

Advantages:

- easy to calculate
- does not require much judgment or analytical skill
- describes the best possible fit for past data
- usually reasonably reliable in the short run

TREND ANALYSIS AND PROJECTION

Trends in Economic Data

Forecasting by trend projection is predicated on the assumption that historical relationships will continue into the future. All such methods use time-series data. Weekly, monthly, or annual series of data on sales and costs, personal income, population, labor force participation rates, and GDP are all examples of economic time series.

All time series, regardless of the nature of the economic variable involved, can be described in terms of a few important underlying characteristics. A secular trend is the long-run pattern of increase or decrease in a series of economic data. Cyclical fluctuation describes the rhythmic variation in economic series that is due to a pattern of expansion or contraction in the overall economy. Seasonal variation, or seasonality, is a rhythmic annual pattern in sales or profits caused by weather, habit, or social custom. Irregular or random influences are unpredictable shocks to the economic system and the pace of economic activity caused by wars, strikes, natural catastrophes, and so on.

LINEAR TREND ANALYSIS

The simplest form of time-series analysis is projecting the past trend by fitting a straight line to the data either visually or, more precisely, by regression analysis. The linear regression model will take the form of

$$S_t = S_0 + b_t \quad (1)$$

Where S_t is the value of the time series to be forecasted for period t, S_0 is the estimated value of the time series (the constant of the regression) in the base period (i.e., at time period t = 0), b is the absolute amount of growth per period, and t is the time period in which the time series is to be forecasted.

TABLE 5-2 Seasonal Demand for (Sales of) Electricity (millions of kilowatt-hours) in a U.S. City, 2003–2006

Time period	2003.1	2003.2	2003.3	2003.4	2004.1	2004.2	2004.3	2004.4
Quantity	11	15	12	14	12	17	13	16
Time period	2005.1	2005.2	2005.3	2005.4	2006.1	2006.2	2006.3	2006.4
Quantity	14	18	15	17	15	20	16	19

For example, fitting a regression line to the electricity sales (consumption) data running from the first quarter of 2003 (t = 1) to the last quarter of 2006 (t = 16) given in the above Table, we get estimated regression Equation.

$$S_t = 11.90 + 0.394t \quad R^2 = 0.50 \quad (2)$$
$$(4.00)$$

Regression Equation indicates that electricity sales in the city in the last quarter of 2002 (that is, S_0) are estimated to be 11.90 million kilowatt hours and increase at the average rate of 0.394 million kilowatt-hours per quarter. The trend variable is statistically significant at better than the 1 percent level (inferred from the value of 4 for the t statistic given in parentheses below the estimated slope coefficient) and "explains" 50 percent in the quarterly variation of electricity consumption in the city (from R^2 = 0.50). Thus, based on the past trend, we can forecast electricity consumption (in million kilowatt-hours) in the city to be

S_{17} = 11.90. +.0.394(17) = 18.60 in the first quarter of 2007
S_{18} = 11.900 + 0.394(18) = 18 .99 in the second quarter of 2007
S_{19} = 11.90 + 0.394(19) = 19.39 in the third quarter of 2007
S_{20} = 11.90 + 0.394(20) = 19.78 in the fourth quarter of 2007

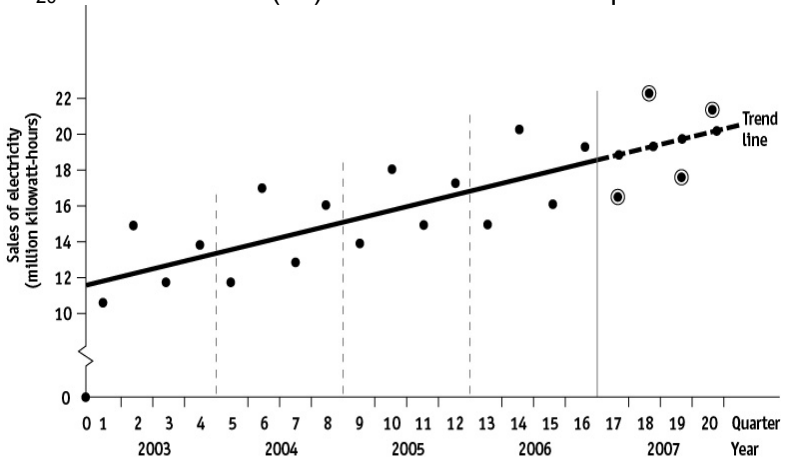


FIGURE 5-2 Forecasting by Trend Projection Electricity sales for the first, second, third, and fourth quarters of 2007 can be read off the extended regression (trend) line (the dots on the dashed portion of the trend line) for quarters 17, 18, 19, and 20, respectively.

These forecasts are shown by the dots on the dashed portion of the trend line extended into 2004 in the above Figure. Note that the forecasted values of electricity sales read off the extended trend line take into consideration only the long-nm trend factor in the data.

Growth Trend Analysis

The constant percentage growth rate model can be specified as:

$$S_t = S_0(1+g)^t \qquad (3)$$

Where g is the constant percentage growth rate to be estimated

To estimate g from Equation, we must first transform the time-series data into their natural logarithms and then run the regression on the transformed data. The transformed regression equation is linear in the logarithms and is given by

$$\ln S_t = \ln S_0 + t \ln (1 + g) \qquad (4)$$

Running regression Equation for the data on electricity sales given in Table transformed into logs, we get

$$\ln S_t = 2.49 + 0.026t \qquad R^2 = 0.50 \qquad (5)$$

$$. \qquad (4.06)$$

In this case, the fit of Equation (5) is similar to that of Equation (2). Because the estimated parameters are now based on the logarithms of the data, however, they must be converted into their anti logs to be able to interpret them in terms of the original data. The antilog of In S_0 = 2.49 is S_0 = 12.06 and the antilog of In (1 + g) = 0.026 gives (1 + g) = 1.026. Substituting these values back into Equation (3), we get

$$S_t = 12.06(1.026)^t \qquad (6)$$

Where S_0 = 12.06 million kilowatt-hours is the estimated sales of electricity in the city in the fourth quarter of 1999..(i.e., at t = 0) and the estimated growth rate is 1.026, or 2.6 percent, per quarter. To estimate sales in any future quarter, we substitute into Equation (5) the value of t for the quarter for which we are seeking to forecast S and solve for S_t Thus,

$S_{17} = 12.06(1.026)^{17} = 18.66$ in the first quarter of 2007
$S_{18} = 12.66(1.026)^{18} == 19.14$ in the second quarter of 2007
$S_{19} = 2.06(1.026)^{19} = 19.64$ in the third quarter of 2007
$S_{20} = 12.06(1.026)^{20} = 20.15$ in the fourth quarter of 2007

These forecasts are similar to those obtained by fitting a linear trend.

Lesson 11

DEMAND FORECASTING (CONTINUED 1)

TIME-SERIES ANALYSIS

Time series data can be represented as:

$$Y_t = f(T_t, C_t, S_t, R_t)$$

Y_t = actual value of the data at time t
T_t = trend component at t
C_t = cyclical component at t
S_t = seasonal component at t
R_t = random component at t

Starting with the example from Lesson # 10, as we have seen, the forecasted value of electricity sales read off from the extended, trend line in Figure take into consideration only the long-run trend factor in the data. The data for the years 2003 to 2006, however, show strong seasonal variation, with sales in the first arid third quarters of each year consistently below the corresponding long-run trend values, while sales in the second and fourth quarters are consistently above the trend values. By incorporating this seasonal variation, we can significantly improve the forecast of electricity sales in the city. We can do this with the ratio-to-trend method or with dummy variables.

TABLE 5-2 Seasonal Demand for (Sales of) Electricity (millions of kilowatt-hours) in a U.S. City, 2003–2006

Time period	2003.1	2003.2	2003.3	2003.4	2004.1	2004.2	2004.3	2004.4
Quantity	11	15	12	14	12	17	13	16
Time period	2005.1	2005.2	2005.3	2005.4	2006.1	2006.2	2006.3	2006.4
Quantity	14	18	15	17	15	20	16	19

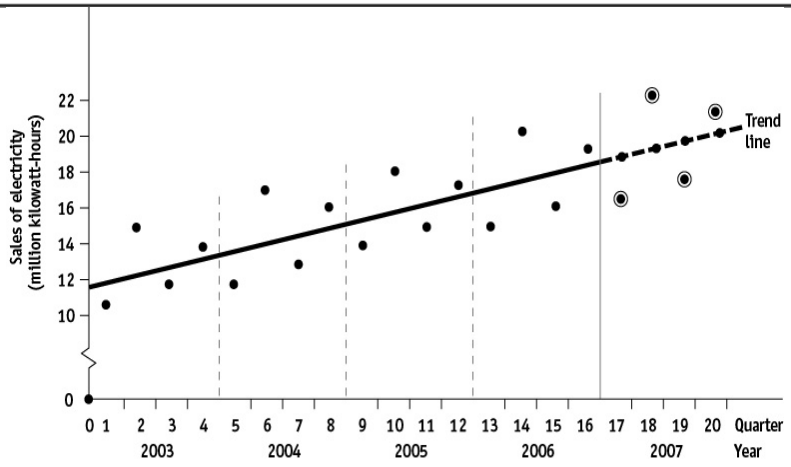


FIGURE 5-2 Forecasting by Trend Projection Electricity sales for the first, second, third, and fourth quarters of 2007 can be read off the extended regression (trend) line (the dots on the dashed portion of the trend line) for quarters 17, 18, 19, and 20, respectively.

To adjust the trend forecast for the seasonal variation by the ratio-to-trend method, we simply find the average ratio by which the actual value of the time series differs from the corresponding estimated trend value in each quarter during the 2003 to 2006 period and then multiply the forecasted trend value by this ratio. The predicted trend value for each quarter in the 2003 to

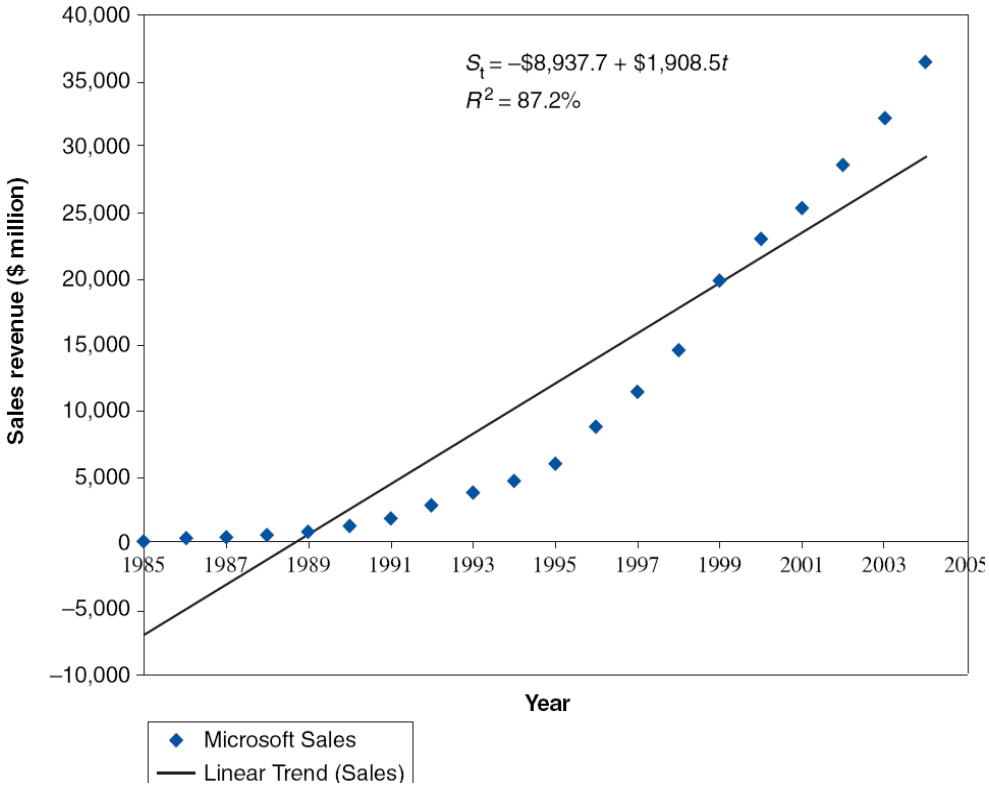
2006 period is obtained by simply substituting the value of t corresponding to the quarter under consideration into Equation and solving for S_t. It is also given in the computer printout for Equation. The above Table shows the calculations for the seasonal adjustment of the electricity sales forecasted for each quarter of 2003 from the extended trend line examined earlier.
Multiplying the electricity sales forecasted earlier (from the simple extension of the linear trend) by the seasonal factors estimated in Table (that is, 0.887 for the first quarter, 1.165 for the second quarter, and so on) we get the following new forecasts based on both the linear trend and the seasonal adjustment:

S_{17} = 18.60(0.887) = 16.50 in the first quarter of 2007
S_{18} = 18.99(1.165) = 22.12 in the second quarter of 2007
S_{19} = 19.39(0.907) = 17.59 in the third quarter of 2007
S_{20} = 19.78(1.042) = 20.61 in the fourth quarter of 2007

These forecasts are shown by the encircled points in Figure. Note that with the inclusion of the seasonal adjustment, the forecasted values for electricity sales closely replicate the past seasonal pattern in the time-series data along the rising linear trend.

TREND PROJECTION
Linear trend analysis assumes a constant period-by-period *unit* change in an important economic variable over time.
Linear Trend:
$S_t = S_0 + b\,t$
b = Growth per time period
Constant Growth Rate
$S_t = S_0 (1 + g)^t$
g = Growth rate
Estimation of Growth Rate
$\ln S_t = \ln S_0 + t \ln(1 + g)$
Microsoft Sales: Linear Trend

Year	t	Sales Revenue ($ Million)	Fitted Sales (Linear)
1985	1	139	7029
1986	2	202	5120
1987	3	345	3212
1988	4	590	1303
1989	5	803	605
1990	6	1183	2513
1991	7	1843	4422
1992	8	2758	6330
1993	9	3753	8239
1994	10	4649	10147
1995	11	5937	12056
1996	12	8671	13964
1997	13	11358	15873
1998	14	14484	17781
1999	15	19747	19690
2000	16	22956	21599
2001	17	25296	23507
2002	18	28635	25416
2003	19	32187	27325



Growth trend analysis assumes a constant period-by-period *percentage* change in an important economic variable over time. Such a forecast model has the potential to better capture the increasing annual sales pattern described by the 1984–2001 Microsoft sales data. This model is appropriate for forecasting when sales appear to change over time by a constant proportional amount rather than by the constant absolute amount assumption implicit in a simple linear model. The constant annual rate of growth model, assuming *annual* compounding, is described as follows:

Sales in *t* Years = Current Sales _ (1 + Growth Rate)*t*
***St* = *S*0 (1 + *g*)*t* (1)**

In words, Equation (1) means that sales in t years in the future are equal to current-period sales, S0, compounded at a constant annual growth rate, g, for a period of t years. Use of the constant annual rate of growth model involves determining the average historical rate of growth in a variable such as sales and then using that rate of growth in a forecast equation such as Equation (1).
The coefficients of the equation can be estimated by using Microsoft sales data for the 1984-2001 period and the least squares regression method. Sales projections are based on a linear trend line, which implies that sales increase by a constant dollar amount each year. In this example, Microsoft sales are projected to grow by $1,407.3 million per year. However, there are important reasons for believing that the true trend for Microsoft sales is nonlinear and that the forecasts generated by this constant change model will be relatively poor estimates of actual values. To see why a linear trend relation may be inaccurate, consider the relation between actual sales data and the linear trend shown in the above Figure.

Microsoft: Linear Trend (1985 to 2004)

$S_t = S_0 + b\ t$

S_t = - 8,937.7 + 1,908.5t
(-4.32) (11.06)
t = 2010 – 1984 = 26
S2010 = - 8,937.7 + 1,908.5(26) = $40,683 million
S2015 = $50,226 million

Microsoft: Growth Trend

Sales in t years = Current Sales * (1 + Growth Rate)t

$S_t = S_0 (1 + g)^t$
g = Growth rate
Estimation of Growth Rate
$\ln S_t$ = 2.260 + 0.128t R^2 = 96.4%
(33.3) (22.67) 34.4% growth rate
S_t = 182.3(1.344)t Antilogs
S2010 = $182.3(1.344)^{26}$ = $397,345million
S2015 = $1,742,465million

The importance of selecting the correct structural form for a trending model can be demonstrated by comparing the sales projections that result from the two basic approaches that have been considered. Remember that with the constant change model, sales were projected to be $24.5 billion in 2005 and $31.6 billion in 2010. Compare these sales forecasts with projections of $158.1 billion in 2005 and $850.0 billion in 2010 for the constant growth rate model. Notice that the difference in the near-term forecasts (2005) is smaller than the difference between longer term (2010) projections. This shows that if an economic time series is growing at a constant rate rather than increasing by a constant dollar amount, forecasts based on a linear trend model will tend to be less accurate the further one forecasts into the future.

Although trend projections provide useful results for some forecasting purposes, shortcomings can limit their usefulness. An obvious problem is that the accuracy of trend projections depends upon a continuation of historical patterns for sales, costs, and profits. Serious forecasting errors resulted when this technique was employed in the periods just prior to unanticipated economic downturns in 1982, 1991 and 2001.

Microsoft Comparison

Year	Linear	Growth
2010	$40.6 billion	$397.3 billion
2015	$50.2 billion	$1,742.5 billion

Microsoft Comparison

Year	Actual

2008	$60.42 billion
2009	$58.43 billion

Annual Report of Microsoft (2009) says "A worldwide economic recession that created the most difficult business environment since the Great Depression, made the fiscal year 2009 a challenging year for the Microsoft Corporation.

SEASONAL VARIATIONS

Time series components: **seasonality**

- Need to identify and remove seasonal factors, using moving averages to isolate those factors
- Remove seasonality by dividing data by seasonal facto

Calculation of the Seasonal Adjustment of the Trend Forecast by the Ratio-to-Trend Method

Year	Forecasted	Actual	Actual / Forecasted
2000.1	12.29	11.00	0.895
2001.1	13.87	12.00	0.865
2002.1	15.45	14.00	0.906
2003.1	17.02	15.00	0.881
		Average	**0.887**
2000.2	12.69	15.00	1.182
2001.2	14.26	17.00	1.192
2002.2	15.84	18.00	1.136
2003.2	17.42	20.00	1.148
		Average	**1.165**
2000.3	13.08	12.00	0.917
2001.3	14.66	13.00	0.887
2002.3	16.23	15.00	0.924
2003.3	17.81	16.00	0.898
		Average	**0.907**
2000.4	13.48	14.00	1.039
2001.4	15.05	16.00	1.063
2002.4	16.63	17.00	1.022
2003.4	18.20	19.00	1.044
		Average	**1.042**

RATIO TO TREND METHOD:

Example Calculation for Quarter 1

Trend Forecast for 1996.1 = 11.90 + (0.394)(17) = 18.60

Seasonally Adjusted Forecast for 1996.1 = (18.60)(0.8869) = 16.50

Year	Trend Forecast	Actual	Ratio
1992.1	12.29	11.00	0.8950
1993.1	13.87	12.00	0.8652
1994.1	15.45	14.00	0.9061
1995.1	17.02	15.00	0.8813
Seasonal Adjustment =			0.8869

Ratio to Trend

- S17 = 18.60(0.887) = 16.50 in Q1 0f 1996
- S18 = 18.99(1.165) = 22.12 in Q2 0f 1996
- S19 = 19.39(0.907) = 17.59 in Q3 0f 1996
- S20 = 19.78(1.042) = 20.61 in Q4 of 1996

Lesson 12

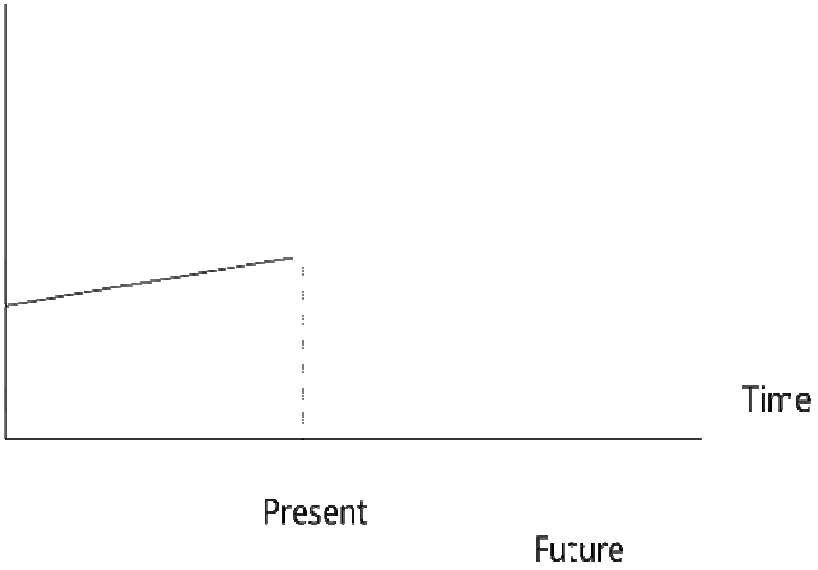
DEMAND FORECASTING (CONTINUED 2)

A wide variety of statistical forecasting techniques can be used to predict unit sales growth, revenue, costs, and profit performance. These techniques range from quite simple to very sophisticated.

SIMPLEST METHOD IS EXTRAPOLATION

In mathematics, extrapolation is the process of constructing new data points outside a discrete set of known data points. Statistical technique of inferring unknown from the known. It attempts to predict future data by relying on historical data, such as estimating the size of a population a few years from now on the basis of current population size and its rate of growth. Extrapolation may be valid where the present circumstances do not indicate any interruption in the long-established past trends. However, a straight line extrapolation (where a short-term trend is believed to continue far in into future) is full of risk because some unexpected factors almost always occur.

Figure (1)



SMOOTHING TECHNIQUES

Other methods of naive forecasting are smoothing techniques. These predict values of a time series on the basis of some average of its past values only. Smoothing techniques are useful when the time series exhibit little trend or seasonal variations but a great deal of irregular or random variation. The irregular or random variation in the time series is then smoothed, and future values are forecasted based on some average of past observations. Smoothing techniques includes:

1. Moving Average
2. Exponential smoothing

Both moving average and exponential smoothing techniques work best when there is :

- no strong trend in series
- infrequent changes in direction of series
- fluctuations are random rather than seasonal or cyclical

MOVING AVERAGE

Moving Average Forecast is the average of data from *w* periods prior to the forecast data point.

$$F_t = \sum_{i=1}^{w} \frac{A_{t-i}}{w}$$

The simplest smoothing technique is the moving average. Here the forecasted value of a time series in a given period (month, quarter, year, etc.) is equal-to the average value of the time series in a number of previous periods. For example, with a three period moving average, the forecasted value of the time series for the next period is given by the average value of the time series in the previous three periods. Similarly, with a five-period moving average, the forecast for the next period is equal to the average for the previous five periods, and so on. The greater the number of periods used in the moving average, the greater is the smoothing effect because each new observation receives less weight. This is more useful the more erratic or random is the time-series data.

For example, columns 1 and 2 in Table present hypothetical data on the market share of a firm for 12 quarters. Note that the data seem to show considerable random variation but no secular or seasonal variations. Column 3 gives the calculated three-quarter moving average. For example, the value of 21.67 for the fourth quarter (the first value in column 3) is obtained by adding the first three values in column 2 and dividing by 3 [i.e., (20 + 22 + 23)/3 = 21.67]. By continuing this way, we forecast the firm's market share to be 21.33 in the thirteenth quarter (this is a real forecast, since no actual data were available for the thirteenth quarter). On the other hand, by averaging the firm's market share in the first five quarters in column 2, we get the five-quarter moving average forecast of 21.4 for the sixth quarter shown in column 6 of the table. This compares with the actual value of 23 in column 2.

Table (1)
Three-Quarter Five-Quarter Moving Average Forecasts and Comparison

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Quarter	**Firm's Actual Market Share (A)**	**Three-Quarter Moving Average Forecast (F)**	**A - F**	**$(A - F)^2$**	**Five-Quarter Moving Average Forecast (F)**	**A – F**	**$(A - F)^2$**
1	20	-	-	-	-	-	-
2	22	-	-	-	-	-	-
3	23	-	-	-	-	-	-
4	24	21.67	2.33	5.4289	-	-	-
5	18	23.00	-5.00	25.0000	-	-	-
6	23	21.67	1.33	1.7689	21.4	1.6	2.56
7	19	21.67	-2.67	7.1289	22.0	-3.0	9.00
8	17	20.00	-3.00	9.0000	21.4	-4.4	19.36
9	22	19.67	2.33	5.4289	20.2	1.8	3.24
10	23	19.33	3.67	13.4689	19.8	3.2	10.24
11	18	20.67	-2.67	7.1289	20.8	-2.8	7.84
12	23	21.00	2.00	4.0000	19.8	3.2	10.24
				Total: 78.3534			**Total: 62.48**

13	-	21.33			20.6		

Although in Table (1) we calculated the three-quarter and the five-quarter moving average forecasts for the firm's market share, moving average forecasts for still other numbers of quarters can be obtained. In order to decide which of these moving average forecasts is better (i.e., closer to the actual data), we calculate the root-mean-square error (RMSE) of each forecast and use the moving average that results in the smallest RMSE (weighted average error in the forecast). RMSE measures the accuracy of a forecasting method. The formula for the root-mean-square error (RMSE) is:

$$RMSE = \sqrt{\frac{\sum (A_t - F_t)^2}{n}}$$

Where A_t is the actual value of the time series in period t, F_t is the forecasted value, and n is the number of time periods or observations. The forecast difference or error (that is, A - F) is squared in order 10 penalize larger errors proportionately more than smaller errors.
For example, column 4 in Table shows At - Ft for the three-quarter moving average forecast in column 3. Column 5 shows $(At - Ft)^2$. The RMSE for the three-quarter moving average forecast in column 3 is obtained by dividing the total of column 5 by 9 (the number of squared forecast errors) and finding the square root. That is,

$$\text{RMSE} = \sqrt{\frac{78.3534}{9}} = 2.95$$

This compares with

$$\text{RMSE} = \sqrt{\frac{62.48}{7}} = 2.99$$

For the five quarter moving average forecast. Thus, the three-quarter moving average forecast is marginally better than the corresponding five-quarter moving average forecast. That is, we are a little more confident in the forecast of 21.33 than 20.6 for the thirteenth quarter.

EXPONENTIAL SMOOTHING
A serious criticism of using simple moving averages in forecasting is that they give equal weight to all observations in computing the average, even though intuitively we might expect more recent observations to be more important. Exponential smoothing overcomes this objection and is used more frequently than simple moving averages in forecasting.

Exponential smoothing is a method for forecasting trends in unit sales, unit costs, wage expenses, and so on. The technique identifies historical patterns of trend or seasonality in the data and then extrapolates these patterns forward into the forecast period. Its accuracy depends on the degree to which established patterns of change are apparent and constant over time. The more regular the pattern of change in any given data series, the easier it is to forecast. Exponential smoothing techniques are among the most widely used forecasting methods in business.

With exponential smoothing, the forecast for period t + 1 (that is, F_t+1) is a weighted average of the actual and forecasted values of the time series in period t. The value of the time series at period t (that is, A_t) is assigned a weight (w) between 0 and 1 inclusive, and the forecast for period t (that is, F_t) is assigned the weight of 1 – w. The greater the value of w, the greater is the weight given to the value of the time series in period t as opposed to previous periods. Thus, the value of the forecast of the time series in period t + 1 is

$$F_{t+1} = wA_t + (1- w) F_t$$

$$0 \leq w \leq 1$$

Two decisions must be made in order to use Equation for exponential smoothing. First, it is necessary to assign a value to the initial forecast (F_t) to get the analysis started. One way to do this is to let F_t, equal the mean value of the entire observed time series data. We must also decide on the value of w (the weight to assign to A_t). In general, different values of w are tried, and the one that leads to the forecast with the smallest root-mean-square error (RMSE) is actually used in forecasting. For example, column 3 in Table shows the forecasts for the firm's market share data given in columns 1 and 2 (the same as in Table) by using the average, market share of the firm over the 12 quarters for which we have data (that is, 21.0) for F_t (to get the calculations started) and w = 0.3 as the weight for A_t. Thus, F_2 (the second value in column 3) is

$$F_2 = 0.3(20) + (1-0.3)\ 21 = 20.7$$

The forecasts for other time periods (rounded off to. the first decimal) are similarly, obtained, until F13= 21.0 for the thirteenth quarter.
On the other hand starting again with the average market share of the firm for the 12 quarters for which we have data (that is, 21.0) for Ft but now using w= 0.5 as the weight for At, we get the exponential forecasts of the firm's market share shown in column 6 of Table (2). Thus, F2 (the second value in column 6) is

$$F_2 = 0.5(20) + (1 - 0.5)21 = 20.5$$

The forecasts for the other time periods are similarly obtained, until F_{13} = 21.5 for the thirteenth quarter.

Table (2)

Exponential Forecasts with w = 0.3 and w = 0.5, and Comparison							
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Quarter	Firm's Actual Market Share (A)	Forecast with w = 0.3 (F)	A - F	$(A - F)^2$	Forecast with w = 0.5 (F)	A – F	$(A - F)^2$
1	20	21.0	-1.0	1.00	21.0	-1.0	1.00
2	22	20.7	1.3	1.69	20.5	1.5	2.25
3	23	21.1	1.9	3.61	21.3	1.7	2.89
4	24	21.7	2.3	5.29	22.2	1.8	3.24
5	18	22.4	-4.4	19.36	23.1	-5.1	26.01
6	23	21.1	1.9	3.61	20.6	2.4	5.76
7	19	21.7	-2.7	7.29	21.8	-2.8	7.84
8	17	20.9	-3.9	15.21	20.4	-3.4	11.56
9	22	19.7	2.3	5.29	18.7	3.3	10.89
10	23	20.4	2.6	6.76	20.4	2.6	6.76
11	18	21.2	-3.2	10.24	21.7	-3.7	13.69
12	23	20.2	2.8	7.84	19.9	3.1	9.61
				TOTAL = 87.19			TOTAL = 101.50
13	-	21.0			21.5		

The root-mean-square error (RMSE) for the exponential forecasts using w = 0.3 is

$$RMSE = \sqrt{\frac{87.19}{12}} = 2.70$$

On the other hand, the RMSE for the exponential forecasts using w = 0.5 is

$$RMSE = \sqrt{\frac{101.5}{12}} = 2.91$$

Thus, we are more confident in the exponential forecast of 21.0 for the thirteenth quarter obtained by using w = 0.3 than in the exponential forecast of 21.5 obtained by using w = 0.5 (see Table). Both exponential forecasts are also better than the three-quarter and the five-quarter moving average forecasts obtained earlier in Section. Since the best exponential forecast is usually better than the best moving average forecast, the former is generally used.

BAROMETRIC METHODS

One way to forecast or anticipate short-term changes in economic activity or turning points in business cycles is to use the index of leading economic indicators. These are time series that tend to precede (lead) changes in the level of general economic activity, much as changes in the mercury in a barometer precede change in weather conditions (hence the name barometric methods). Barometric forecasting, as conducted today is primarily the result of the work conducted at the National Bureau of Economic Research (NBER), a private organization. Today, economic indicator data are published monthly by The Conference Board in *Business*

Cycle Indicators. These monthly data are reported in press and on the Net.
There are three major series: leading, coincident, and lagging indicators. As their names imply, the first tells us where we are going, the second where we are, and the third where we have been. Although the leading indicators series is probably the most important, the other two are also meaningful. The coincident indicators identify peaks and troughs, and the lagging series confirms upturns and downturns in economic activity.
A rise in leading economic indicators is used to forecast an increase in general business activity, and vice versa. For example, an increase in building permits can be used to forecast an increase in housing construction. Less obvious but very important an increase in stock prices, in general, precedes (i.e., it is a leading indicator for) an upturn in general business activity, since rising stock prices reflect expectations by business managers and others that profits will rise. On the other hand, a decline in contracts for plant and equipment usually precedes a slowdown in general economic activity. Thus, leading indicators are used to forecast turning points in the business cycle.

The Business Cycle is a rhythmic pattern of economic expansion and contraction and economic indicators help forecast the economy.

- Leading indicators, e.g., stock prices, building permits
- Coincident indicators, e.g., production, manufacturing & trade sales
- Lagging indicators, e.g., unemployment, change in labor cost

Although we are primarily interested in leading indicators, some time series move in step or coincide with movements in general economic activity and are therefore called coincident indicators. Still others follow or lag movements in economic activity and are called lagging indicators. The relative positions of leading, coincident, and lagging indicators in the business cycle are shown graphically in Figure. The figure shows that leading indicators precede business cycles' turning points (i.e., peaks and troughs), coincident indicators move in step with business cycles, while lagging indicators follow or lag turning points in business cycles.

Figure (2)

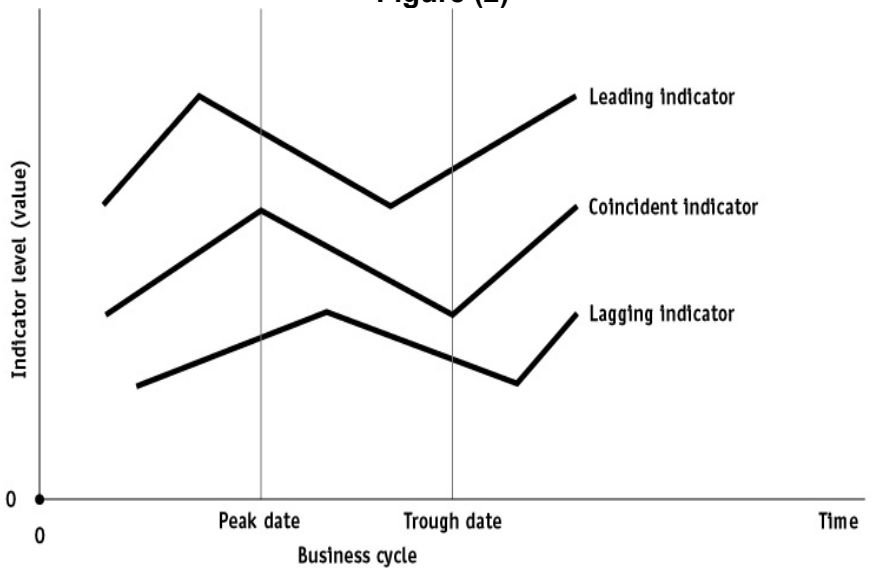


FIGURE 5-3 Economic Indicators *Leading indicators* precede (lead) business cycles' turning points (i.e., peaks and troughs), *coincident indicators* move in step with business cycles, while *lagging indicators* follow or lag turning points in business cycles.

Another method for overcoming the difficulty arising when some of the 10 leading indicators move up and some move down is the diffusion index. Instead of combining the 10 leading indicators into a composite index, the diffusion index gives the percentage of the 10 leading indicators moving upward. If all 10 move up, the diffusion index is 100. If all move down, its value is 0. If only 7 move up, the diffusion index is 70. We usually forecast an improvement in economic activity when the diffusion index is above 50, and we have greater confidence in our forecasting the closer the index is to 100. In general, barometric forecasting employs composite and diffusion indexes rather than individual indicators.

It may be noted at the outset that the barometric technique was developed to forecast the general trend in overall economic activities. This method can be used to forecast the demand prospects of a product, not the actual quantity expected to be demanded, For example, development and allotment of land by the Lahore Development Authority (LDA) to the Eden Developers (a lead indicator) indicate higher demand prospects for cement, bricks, steel and other construction material.

Short List of Leading, Coincident, and Lagging Indicators

LEADING INDICATORS

1. Average weekly hours, manufacturing
2. Initial claims far unemployment insurance
3. Manufacturers new orders, consumer goods and materials
4. Vendor performance, slower deliveries diffusion index
5. Manufacturers new orders, non defense capital goods
6. Building permits, new private housing units
7. Stack prices, 500 common stacks
8. Money supply, M2
9. Interest rate spread, 10-year Treasury bonds less federal funds
10. Index of consumer expectations

COINCIDENT INDICATORS

1. Employees on nonagricultural payrolls
2. Personal income less transfer payments
3. Industrial production
4. Manufacturing and trade sales

LAGGING INDICATORS

1. Average duration of unemployment, weeks
2. Ratio manufacturing and trade inventories to sales
3. Change in labor cast per unit of output, manufacturing
4. Average prime rate charged by banks
5. Commercial and industrial loans outstanding
6. Ratio consumer installment credit to personal income
7. Change in consumer price index for services

COMPOSITE INDEXES

Composite index of 10 leading indicators
Composite index of 4 coincident indicators
Composite index of 7 lagging indicators

Although the composite and diffusion indexes of leading indicators are reasonably good tools for predicting turning points in business cycles, they face a number of shortcomings. One is that on several occasions they forecasted a recession that failed to occur. The variability in lead time can also be considerable. More importantly, barometric forecasting gives little or no indication of the magnitude of the forecasted change in the level of economic activity (i.e., it provides only a qualitative forecast of turning points). Thus, while barometric forecasting is certainly superior to time series analysis and smoothing techniques (naive methods) in forecasting short term turning points in economic activity, it must be used together with other methods (such as econometric forecasting) to forecast the magnitude of change in the level of economic activity.

Lesson 13

DEMAND FORECASTING (CONTINUED 3)

ECONOMETRIC METHODS

Econometric methods combine economic theory with statistical tools to analyze economic relations. The firm's demand and sales of a commodity, as well as many other economic variables, are increasingly being forecasted with econometric models. The characteristic that distinguishes econometric models from other forecasting methods is that they seek to identify and measure the relative importance (elasticity) of the various determinants of demand or other economic variables to be forecasted. By attempting to explain the relationship being forecasted, econometric forecasting allows the manager to determine the optimal policies for the firm.

Econometric forecasting techniques have several advantages over alternative methods.

ADVANTAGES OF ECONOMETRIC METHODS

- Econometric methods force the forecaster to make explicit assumptions about the linkages among the variables in the economic system being examined. In other words, the forecaster must deal with causal relations. This produces logical consistency in the forecast model and increases reliability.

- Another advantage of econometric methods is that the forecaster can compare forecasts with actual results and use insights gained to improve the forecast model. By feeding past forecasting errors back into the model, new parameter estimates can be generated to improve future forecasting results.

- The type of output provided by econometric forecasts is another major advantage. Because econometric models offer estimates of actual values for forecasted variables, these models indicate both the direction and magnitude of change.

- Finally, the most important advantage of econometric models relates to their ability to explain economic phenomena

Econometric forecasting frequently uses the best features of other forecasting techniques, such as trend and seasonal variations, smoothing techniques, and leading indicators. Econometric forecasting models range from single-equation models of the demand that the firm faces for its product to large, multiple equation models describing hundreds of sectors and industries of the economy. Although the concern here is with forecasting the demand for a firm's product, macro forecasts of national income and major sectors of the economy are often used as inputs or explanatory variables in simple single-equation demand models of the firm. Therefore, we discuss both types of forecasting.

SINGLE-EQUATION MODELS

Many managerial forecasting problems can be adequately addressed with single-equation econometric models. The first step in developing an econometric model is to express relevant economic relations in the form of an equation. When constructing a model for forecasting the regional demand for portable personal computers, one might hypothesize that computer demand (*C*) is determined by price (*P*), disposable income (*I*), population (*Pop*), interest rates (*i*), and advertising expenditures (*A*). A linear model expressing this relation is

$$C = a0 + a1P + a2I + a3Pop + a4i + a5A \qquad (1)$$

The next step in econometric modeling is to estimate the parameters of the system, or values of the coefficients, as in Equation (1). The most frequently used technique for parameter estimation is the application of least squares regression analysis with either time-series or cross-section data.

Once the model coefficients have been estimated, forecasting with a single-equation model consists of evaluating the equation with specific values for the independent variables. An econometric model used for forecasting purposes must contain independent or explanatory variables whose values for the forecast period can be readily obtained.

REGRESSION ANALYSIS: a procedure commonly used by economists to estimate consumer demand with available data. In our Demand Estimation lectures we have discussed the Regression analysis in detail.
Two types of regression:

- cross-sectional: analyze several variables for a single period of time
- time series data: analyze a single variable over multiple periods of time

INTERPRETING THE REGRESSION RESULTS:
Coefficients:

- Negative coefficient shows that as the independent variable (X_n) changes, the variable (Y) changes in the opposite direction
- Positive coefficient shows that as the independent variable (X_n) changes, the dependent variable (Y) changes in the same direction
- Magnitude of regression coefficients is a measure of elasticity of each variable

Steps for analyzing regression results

- Check coefficient signs and magnitudes
- Compute implied elasticities
- Determine statistical significance using the appropriate test

MULTIPLE-EQUATION SYSTEMS

Although forecasting problems can often be analyzed with a single-equation model, complex relations among economic variables sometimes require use of multiple-equation systems. Variables whose values are determined within such a model are *endogenous*, meaning originating from within; those determined outside or external to, the system are referred to as *exogenous*. The values of endogenous variables are determined by the model; the values of exogenous variables are given externally. Endogenous variables are equivalent to the dependent variable in a single-equation system; exogenous and predetermined variables are equivalent to the independent variables.

Multiple-equation econometric models are composed of two basic kinds of expressions, identities and behavioral equations. **Identities** express relations that are true by definition. The statement that profits (Π) equal total revenue (TR) minus total cost (TC) is an example of an identity:

$$\Pi = TR - TC \qquad (2)$$

Profits are *defined* by the relation expressed in Equation (2). The second group of equations encountered in econometric models, **behavioral equations**, reflects hypotheses about how variables in a system interact with each other. Behavioral equations may indicate how individuals and institutions are expected to react to various stimuli. Perhaps the easiest way to illustrate the use of multiple-equation systems is to examine a simple three-equation forecast

model for equipment and related software sales for a personal computer retailer. Equation (1) expressed a single-equation model that might be used to forecast regional demand for personal computers. However, total revenues for a typical retailer usually include not only sales of personal computers but also sales of software programs (including computer games) and sales of peripheral equipment (e.g., monitors, printers). The three equations are:

Example (1)

$$\textbf{a)} \quad S_t = b_0 + b_t R_t + u_1$$
$$\textbf{b)} \quad P_t = c_0 + c_1 C_{t-1} + u_2$$
$$\textbf{c)} \quad TR_t = S_t + P_t + C_t$$

Where
S = software sales, P = peripheral sales
C = Computer sales, TR = total revenue
t = current time period, t – 1 = previous time period
u1 and u2 are error terms

Equations (a) and (b) are behavioral equations. Equation (a) hypothesizes that current period software sales are a function of the current level of total revenues; Equation (b) hypothesizes that peripheral sales depend on previous-period personal computer sales. The last equation in the system, Equation (c), is an identity. It defines total revenue as being the sum of software, peripheral equipment, and personal computer sales.

Stochastic disturbance terms in the behavioral equations, *u*1 and *u*2, are included because hypothesized relations are not exact. So long as these stochastic elements are random and their expected values are zero, they do not present a barrier to empirical estimation of system parameters. If error terms are not randomly distributed, parameter estimates will be biased, and the reliability of model forecasts will be questionable. Large error terms, even if they are distributed randomly, reduce forecast accuracy.

REDUCED-FORM EQUATIONS

To forecast next year's software and peripheral sales and total revenue as represented by this illustrative model, it is necessary to express *S*, *P*, and *TR* in terms of variables whose values are known or can be estimated at the moment the forecast is generated. In other words, each endogenous variable (*St*, *Pt*, and *TRt*) must be expressed in terms of the exogenous and predetermined variables (*Ct*-1 and *Ct*). Such relations are called **reduced-form equations** because they reduce complex simultaneous relations to their most basic and simple form. Consider the manipulations of equations in the system necessary to solve for *TR* via its reduced-form equation. Substituting Equation (a) and (b) into Equation (c), we get:

$$\textbf{d)} \quad TRt = b0 + b1TRt + c0 + c1Ct\text{-}1 + Ct$$
$$\textbf{e)} \quad (1 - b1)TRt = b0 + c0 + c1Ct\text{-}1 + Ct$$
$$\textbf{f)} \quad TR_t = \frac{b_0 + c_0 + c_1 C_{t-1} + C_t}{(1 - b_1)}$$

Collecting terms and isolating TR in Equation (e) results in Equation (f). Equation f now relates current total revenues to previous-period and current-period personal computer sales. Assuming that data on previous-period personal computer sales can be obtained and that current-period personal computer sales can be estimated by using Equation (1), Equation (f) provides a forecasting model that accounts for the simultaneous relations expressed in this simplified multiple-equation system. In real-world situations, it is likely that personal computer sales

depend on the price, quantity, and quality of available software and peripheral equipment. Then *S*, *P*, and *C*, along with other important factors, may all be endogenous.

Example (2) **Multiple Equation Model of GNP**

$$C_t = a_1 + b_1 GNP_t + u_{1t}$$
$$I_t = a_2 + b_2 \pi_{t-1} + u_{2t}$$
$$GNP_t \equiv C_t + I_t + G_t$$

Reduced Form Equation

$$GNP_t = \frac{a_1 + a_2}{1 - b_1} + \frac{b_2 \pi_{t-1}}{1} - b_1 + \frac{G_t}{1 - b_1}$$

Since the endogenous variables of the system (i.e., C_t, I_t, and GNP_t) are both determined by and in turn determine the value of the other endogenous variables in the model (i.e., they also appear on the right-hand side of the first two Equations), we cannot use the ordinary least-squares technique (OLS) to estimate the parameters of the structural equations (the a's and the b's in these Equations). More advanced econometric techniques are required to obtain unbiased estimates of the coefficients of the model. These are beyond the scope of this course. By assuming that these coefficients are correctly estimated by the appropriate estimation technique, we can show how the above simple macro model can be used for forecasting the values of the endogenous variables. To do this, we substitute the first two Equations into the third Equation (the definitional equation) and solve. This will give an equation for GNP_t that is expressed only in terms of π_{t-1} and G_t (the exogenous variables of the system).

The discussion, so far, has focused on what is referred to as a structural econometric model. That is, the econometrician uses a blend of economic theory, mathematics, and information about the structure of the economy to construct a quantitative economic model. The econometrician then turns to the observed data—the facts—to estimate the unknown parameter values and turn the economic model into a structural econometric model. The term "structural" refers to the fact that the model gets its structure, or specification, from the economic theory that the econometrician starts with. The idea, for example, that spending on clothing and shoes is determined by household income comes from the core of economic theory.

Actually, no econometric model is ever truly complete. All models contain variables the model cannot predict because they are determined by forces "outside" the model. For example, a realistic model must include personal income taxes collected by the government because taxes are the wedge between the gross income earned by households and the net income (what economists call disposable income) available for households to spend. The taxes collected depend on the tax rates in the income tax laws. But the tax rates are determined by the government as a part of its **fiscal policy** and are not explained by the model. If the model is to be used to forecast economic activity several years into the future, the econometrician must include anticipated future tax rates in the model's **information** base. That requires an assumption about whether the government will change future income tax rates and, if so, when and by how much. Similarly, the model requires an assumption about the **monetary policy** that the central bank (the **Federal Reserve System** in the United States) will pursue, as well as

assumptions about a host of other such "outside of the model" (or exogenous) variables in order to forecast all the "inside of the model" (or endogenous) variables.

JUDGING FORECAST RELIABILITY

Tests of Predictive Capability: consistency between test and forecast sample suggests predictive accuracy.

Correlation Analysis: High correlation indicates predictive accuracy. In analyzing a model's forecast capability, the correlation between forecast and actual values is of substantial interest. The formula for the simple correlation coefficient, *r*, for forecast and actual values, *f* and *x*, respectively, is

$$r = \sigma_{fx} / \sigma_f \sigma_x$$

Where σ_{fx} is the covariance between the forecast and actual series, and σf and σx are the sample standard deviations of the forecast and actual series, respectively. Basic spreadsheet and statistical software readily provide these data, making the calculation of *r* a relatively simple task. Generally speaking, correlations between forecast and actual values in excess of 0.99 (99 percent) are highly desirable and indicate that the forecast model being considered constitutes an effective tool for analysis.

In cross-section analysis, in which the important trend element in most economic data is held constant, a correlation of 99 percent between forecast and actual values is rare. When unusually difficult forecasting problems are being addressed, correlations between forecast and actual data of 90 percent or 95 percent may prove satisfactory. By contrast, in critical decision situations, forecast values may have to be estimated at very precise levels. In such instances, forecast and actual data may have to exhibit an extremely high level of correlation, 99.5 percent or 99.75 percent, to generate a high level of confidence in forecast reliability.

SAMPLE MEAN FORECAST ERROR ANALYSIS

Further evaluation of a model's predictive capability can be made through consideration of a measure called the **sample mean forecast error**, which provides a useful estimate of the average forecast error of the model. It is sometimes called the root mean squared forecast error and is denoted by the symbol *U*. The sample mean forecast error is calculated as:

$$U = 1 \sqrt{(fi - xi)2}$$

Where, *fi* is a forecast value, and *xi* is the corresponding actual value. Deviations between forecast and actual values are squared in the calculation of the mean forecast error to prevent positive and negative deviations from canceling each other out. The smaller the sample mean forecast error, the greater the accuracy associated with the forecasting model.

CHOOSING THE BEST FORECAST TECHNIQUE

To select the best technique, managers must be knowledgeable about the strengths and weaknesses of various forecast methods, the amount and quality of available data, and the human and other costs associated with generating reliable forecasts. All firms and other organizations conduct their activities in an uncertain environment, and probably the major role of forecasting is to reduce this uncertainty. But no forecast, however extensive and expensive, can remove it completely. Managers who use forecasts in their work "need to develop realistic expectations as to what forecasting can and cannot do. Forecasting is not a substitute for management judgment in decision making; it is simply an aid to this process."

Lesson 14

PRODUCTION ANALYSIS AND ESTIMATION

THE ORGANIZATION OF PRODUCTION

Production refers to the transformation of inputs or resources into outputs of goods and services. For example, IBM hires workers to use machinery, parts, and raw materials in factories to produce personal computers. The output of a firm can either be a final commodity (such as a personal computer) or an intermediate product, such as semiconductors (which are used in the production of computers and other goods). The output can also be a service rather than a good. Examples of services are education, medicine, banking, communication, transportation, and many others.

Inputs are the resources used in the production of goods and services. Inputs are classified into labor (including entrepreneurial talent), capital, and land or natural resources. Each of these broad categories, however, includes a great variety of the basic input. For example, labor includes workers, accountants, lawyers, doctors, scientists, and many others. Inputs are also classified as fixed or variable. **Fixed inputs** are those that cannot be readily changed during the time period under consideration. Examples of fixed inputs are the firm's plant and specialized equipment (it takes several years for IBM to build a new factory to produce computer chips to go into its computers). On the other hand **variable inputs** are those that can be varied easily and on very short notice. Examples of variable inputs are most raw materials and unskilled labor.

The time period during which at least one input is fixed is called the **short run**, while the time period when all inputs are variable is called the **long run**. The length of the long run (i.e., the time period required for all inputs to be variable) depends on the industry.

THE PRODUCTION FUNCTION

Just as demand theory centers on the concept of the demand function, production theory revolves around the concept of the production function. A production function is an equation, table, or graph showing the maximum output of a commodity that a firm can produce per period of time with each set of inputs. Both inputs and outputs are measured in physical rather than in monetary units. Technology is assumed to remain constant during the period of the analysis.

$$Q = f(L, K) \qquad (1)$$

Equation (1) reads: The quantity of output is a function of, or depends on, the quantity of labor and capital used in production. Output refers to the number of units of the commodity (say, automobiles) produced, labor refers to the number of workers employed, and capital refers to the amount of the equipment used in production. We assume that all units of Land K are homogeneous or identical.

A production function specifies the maximum output that can be produced for a given amount of input. Alternatively, a production function shows the minimum quantity of input necessary to produce a given level of output. Production functions are determined by the technology available for effectively using plant, equipment, labor, materials, and so on. Any improvement in technology, such as better equipment or a training program that enhances worker productivity, results in a new production function.

TABLE (1)

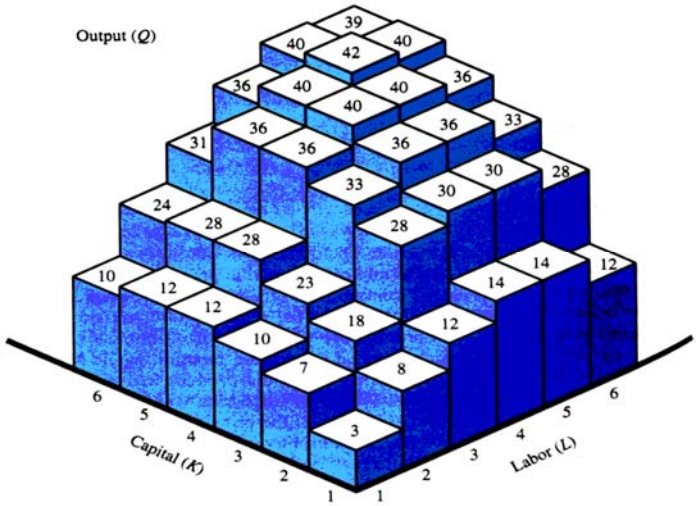
TABLE 6-1 Production Function with Two Inputs

Capital (K)		Labor (L) →						
	6	10	24	31	36	40	39	
	5	12	28	36	40	42	40	
	4	12	28	36	40	40	36	Output (Q)
	3	10	23	33	36	36	33	
↑	2	7	18	28	30	30	28	
K	1	3	8	12	14	14	12	
		1	2	3	4	5	6	

Table (1) gives a hypothetical production function which shows the outputs (the Q's) that the firm can produce with various combinations of labor (L) and capital (K). The table shows that by using 1 unit of labor (1L) and 1 unit of capital (1K), the firm would produce 3 units of output (3Q). With 2L and K, output is 8Q; with 3L and 1K, output is 12Q and so on.

The production relationships given in Table (1) are shown graphically in Figure (1), which is three-dimensional. In Figure (1), the height of the bars refers to the maximum output that can be produced with each combination of labor and capital shown on the axes. Thus, the tops of all the bars form the production surface for the firm.

FIGURE (1)



There are two types of Production functions, one is **discrete** production function (shown in Table (1) and Figure (1)) and the other is a continuous production function. A **continuous** production function is one in which inputs can be varied in an unbroken fashion rather than incrementally.

TOTAL, AVERAGE, AND MARGINAL PRODUCT

Total product is the output from a production system. It is synonymous with Q in Equation (1). Total product is the overall output that results from employing a specific quantity of resources in

a given production system. The total product concept is used to investigate the relation between output and variation in only one input in a production function.
By holding the quantity of one input constant and changing the quantity used of the other input, we can derive the total product (TP) of the variable, input. For example, by holding capital constant at 1 unit (i.e., with K = 1) and increasing the units of labor used from zero to 6 units, we generate the total product of labor given by the last row in Table (1) which is reproduced in column 2 of Table (2). We can see that when no labor is used, total output or product is zero. With one unit of labor (1L), total product (TP) is 3. With 2L, TP = 8. With 3L, TP = 12, and so on.

Q=f(L/ K = 1) (2)

From the total product schedule we can derive the marginal and average product schedules of the variable input. The marginal product (MP) of labor (MP_L) is the change in total product, or extra output per unit change in labor used, while the average product (AP) of labor (AP_L) equals total product divided by the quantity of labor used. That is,

$$MP_L = \frac{\Delta TP}{\Delta L}$$

$$AP_L = \frac{TP}{L}$$

Column 3 in Table (2) gives the marginal product of labor (MPL). Since labor increases by 1 unit at a time in column 1, the MP_L in column 3 is obtained by subtracting successive quantities of TP in column 2. For example, TP increases from 0 to 3 units when the first unit of labor is used. Thus, MPL = 3. For an increase in labor 1 from 1L to 2L, TP rises from 3 to 8 units, so that MPL = 5, and so on. Column 4 of Table (2) gives the APL. This equals TP (column 2) divided by L (column 1). Thus, with 1 unit of labor (1L), APL = 3. With 2L, AP_L = 4, and so on.

TABLE 6-2 Total, Marginal, and Average Product of Labor, and Output Elasticity

(1)	(2)	(3)	(4)	(5)
Labor (number of workers)	Output or Total Product	Marginal Product of Labor	Average Product of Labor	Output Elasticity of Labor
0	0	—	—	—
1	3	3	3	1
2	8	5	4	1.25
3	12	4	4	1
4	14	2	3.5	0.57
5	14	0	2.8	0
6	12	–2	2	–1

Column 5 in Table (2) gives the production or output elasticity of labor (E_L). This measure the percentage change in output divided by the percentage change in the quantity of labor used. That is,

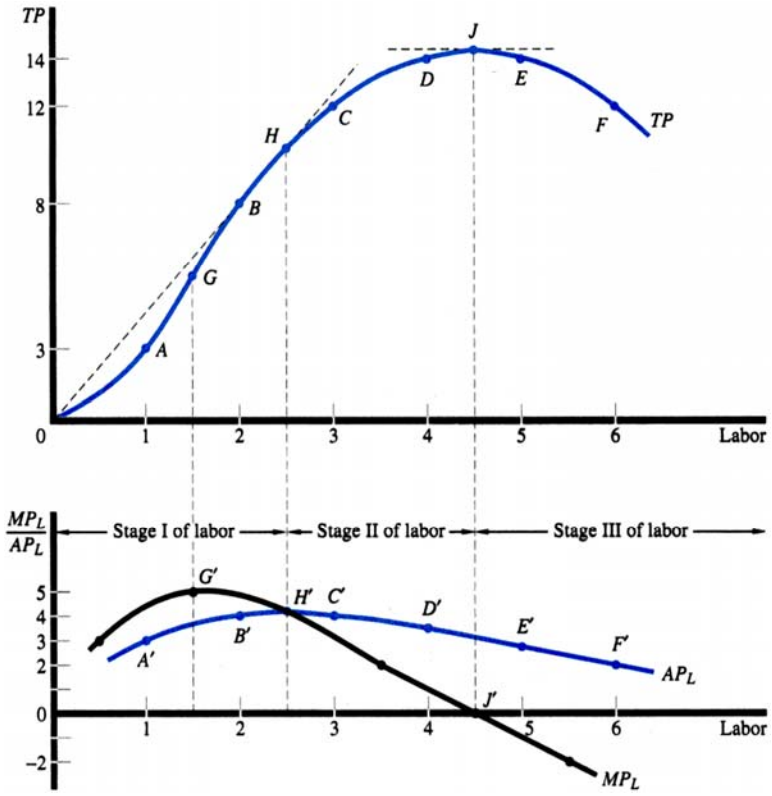
$$E_L = \frac{\%\Delta Q}{\%\Delta L}$$

By rewriting the above equation in a more explicit from and rearranging, we get

$$E_L = \frac{\Delta Q/Q}{\Delta L/L} = \frac{\Delta Q/\Delta L}{Q/L} = \frac{MP_L}{AP_L}$$

That is, output elasticity of labor is equal to the ratio of MP_L to the AP_L. In terms of calculus it can be expressed as: **∂Q / ∂L * L / Q**

FIGURE (2)



In order to show graphically the relationship between the total product, marginal product and average product of labor, we assume that labor time is continuously divisible (i.e., it can be hired for any part of a day). Then the TP, MPL, and AP_L become smooth curves as indicated in Figure (2). The MP_L at a particular point on the TP curve is given by the slope of the TP curve at that point. From Figure (2), we see that the slope of the TP curve rises up to point G (the point of inflection on the TP curve), is zero at point J, and is negative thereafter. Thus, the MPL rises up to point G, is zero at point J, and is negative afterward.

The APL is given by the slope of a ray from the origin to the TP curve. From Figure (2), we see that the slope of the TP curve rises up to point H and falls thereafter but remains positive as long as TP is positive. Thus, the AP_L rises up to point H and falls afterward. Note that at point H the slope of a ray from the origin to the TP curve (or AP_L) is equal to the slope of the TP (or MP_L) curve. So that $MP_L = AP_L$ at point H'. Note also that AP_L rises as long as MP_L is above it and falls when MP_L is below it.

$MP_L = AP_L$ when AP_L is maximum
$MP_L > AP_L$ when AP_L is rising
$MP_L < AP_L$ when AP_L is falling

Three points of interest, G, H, and J, can be identified on the total product curve in Figure (2). Each has a corresponding location on the average or marginal curves. Point G is the inflection point of the total product curve. The marginal product of L (the slope of the total product curve) increases until this point is reached, after which it begins to decrease. This can be seen in the bottom panel of Figure (2), where $\mathbf{MP_L}$ reaches its highest level at G'.

The second point on the total product curve, H, indicates the output at which the average product and marginal product are equal. The slope of a line from the origin to any point on the total product curve measures the average product of L at that point, whereas the slope of the total product curve equals the marginal product.

The third point, J, indicates where the slope of the total product curve is zero and the curve is at a maximum. Beyond J the marginal product of L is negative, indicating that increased use of input L results in a *reduction* of total product. The corresponding point in the bottom panel of Figure (2) is J', the point where the marginal product curve intersects the *X*-axis.

From Figure (2), we can also see that up to point G, the TP curve increases at an increasing rate so that the MP_L rises. Labor is used so scarcely with the 1 unit of capital that the MP_L rises as more labor is used. Past point G, however, the TP curve rises at a decreasing rate so that the MP_L declines. The declining portion of the MP_L curve is a reflection of the **law of diminishing returns.** This postulates that as we use more and more units of the variable input with a given amount of the fixed input, after a point, we get diminishing returns (marginal product) from the variable input. The law of diminishing returns states that the marginal product of a variable factor must eventually decline as more of the variable factor is combined with other fixed resources. The law of diminishing returns is sometimes called the law of diminishing marginal returns to emphasize the fact that it deals with the diminishing marginal product of a variable input factor. The law of diminishing returns cannot be derived deductively. It is a generalization of an empirical regularity associated with every known production system.
In Figure (2), the law of diminishing returns begins to operate after 1.5L is used (after point G' in the bottom panel of Figure (2).

The relationship between the MP_L and AP_L curves in the bottom panel of Figure (2) can be used to define **three stages of production** for labor (the variable input). The range from the origin to the point where the APL is maximum (point H' at 2.5L) is stage I of production for labor. Stage II of production for labor extends from the point where the AP_L is maximum to the point where the MP_L is zero (i.e., from point H' at 2.5L to point J' at 4.5L). The range over which the MP_L is negative (i.e., past point J' or with more than 4.5L) is stage III of production for labor. The rational producer would not operate in stage III of labor, even if labor time were free, because MP_L is negative. In short:

THE THREE STAGES OF PRODUCTION IN THE SHORT RUN:

- Stage I: from zero units of the variable input to where AP_L is maximized (where $MP_L = AP_L$)
- Stage II: from the maximum AP_L curve to where $MP_L = 0$
- Stage III: from where $MP_L=0$ and onwards

In the short run, rational firms should be operating only in Stage II as in Stage III firm uses more variable inputs to produce less output; over-utilizing fixed input; MPL is negative. The question arises why not Stage I? Well the answer is it reflects underutilizing fixed capacity, so the firm can increase output per unit by increasing the amount of the variable input; in stage I MP_K is negative.

OPTIMAL USE OF THE VARIABLE INPUT

How much labor (the variable input in our discussion) should the firm use in order to maximize profits? The answer is that the firm should employ an additional unit of labor as long as the extra revenue generated from the sale of the output produced exceeds the extra cost of hiring the unit of labor (i.e., until the extra revenue equals the extra cost).

The extra revenue generated by the use of an additional unit of labor is called the marginal revenue product of If labor (MRP_L). This equals the marginal product of labor (MP_L) times the marginal revenue (MR) from the sale of the extra output produced. That is,

$$\mathbf{MRP_L = (MP_L)(MR)}$$

On the other hand, the extra cost of hiring an additional unit of labor or marginal resource cost of labor (MRC_L) is equal to the increase in the total cost to the firm re-sulting from hiring the additional unit of labor. That is,

$$\mathbf{MRC_L = \frac{\Delta TC}{\Delta L}}$$

The optimal use of labor requires that:

$$\mathbf{MRP_L = MRC_L}$$

Table (3)

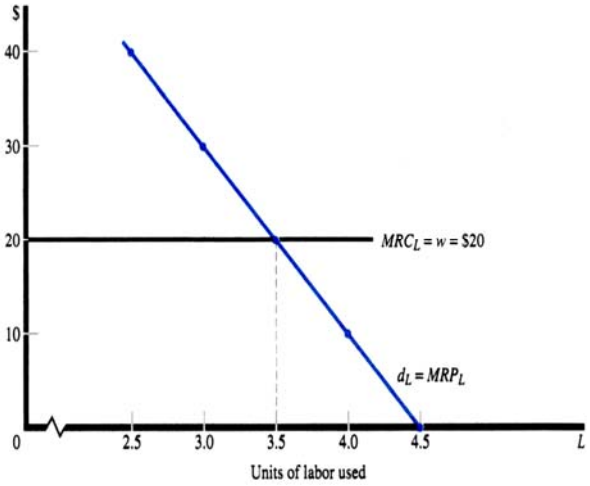
TABLE 6-3 Marginal Revenue Product and Marginal Resource Cost of Labor

(1)	(2)	(3)	(4) = (2) × (3)	(5)
Units of Labor	Marginal Product	Marginal Revenue = *P*	Marginal Revenue Product	Marginal Resource Cost = *w*
2.5	4	$10	$40	$20
3.0	3	10	30	20
3.5	2	10	20	20
4.0	1	10	10	20
4.5	0	10	0	20

Table (3) shows that the optimal use of labor is 3.5 units because only with 3.5L,
$MRP_L = MRC_L = w = \$20$.

The marginal revenue product of labor (MRP_L) schedule in column 4 of Table (3) represents the firm's demand schedule for labor. It gives the amount of labor demanded by the firm at various wage rates. See Figure (3) where $d_L = MRP_L$ is shown as a downward sloping curve.

Figure (3)



For example, if the wage rate per day (w) were $40, the firm would hire 2.5 units of labor because that would be where $MRP_L = MRC_L = W = \$40$. If w = $30, the firm would demand 3 units of labor. If w = $20, the firm would demand 3.5L, and with w = $10, the firm would demand 4L. This is shown in Figure (3), where $d_L = MRP_L$ represents the firm's demand curve for labor; The figure shows that if the wage rate per day (w) were constant at $20, the firm would demand 3.5L, as indicated above.

Lesson 15

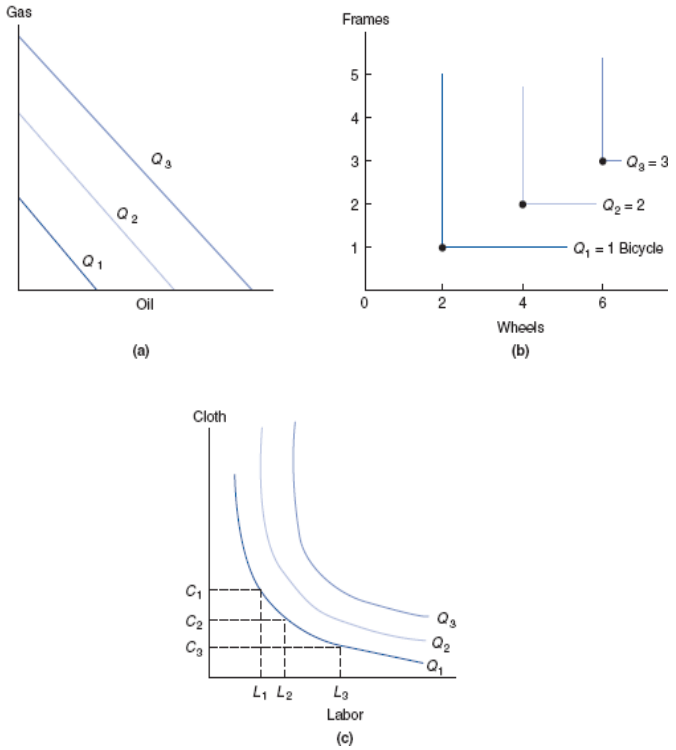
PRODUCTION ANALYSIS AND ESTIMATION (CONTINUED 1)

PRODUCTION WITH TWO VARIABLE INPUTS

PRODUCTION ISOQUANTS

The term **Isoquant**—derived from *iso*, meaning equal, and *quant*, from quantity—denotes a curve that represents the different combinations of inputs that can be efficiently used to produce a given level of output. Efficiency in this case refers to **technical efficiency**, meaning the least cost production of a target level of output. An Isoquant shows the various combinations of two inputs (say, labor and capital) that the firm can use to produce a specific level of output. A higher Isoquant refers to a larger output, while a lower Isoquants refers to a smaller output. Isoquants shapes reveal a great deal about the substitutability of input factors, as illustrated in Figure 1) (a), (b), and (c).

Figure (1)

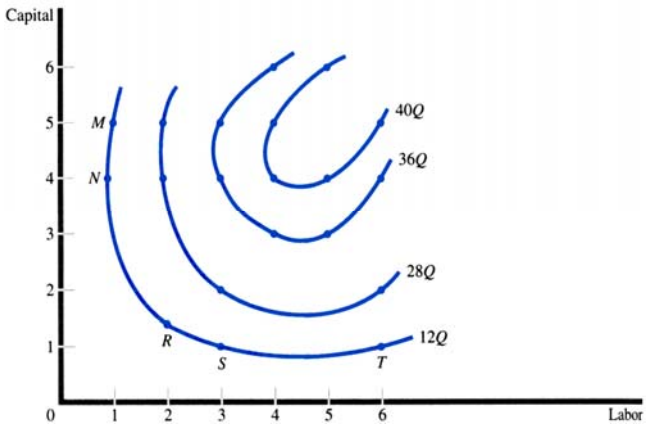


On one hand, the smaller the curvature of an Isoquant, the greater is the degree of substitutability of inputs in production. On the other hand, the greater the curvature of an Isoquant, the smaller is the degree of substitutability.

At one extreme are Isoquants that are straight lines, as shown in the left panel of, Figure (1). In this case, labor and capital are perfect substitutes. That is, the rate at which labor can be substituted for capital in production (i.e., the absolute slope of the Isoquant or MRTS) is constant. This means that labor can be substituted for capital (or vice versa) at the constant rate given by the absolute slope of the Isoquant. And time in a drying process, and fish meal and soybeans used to provide protein in a feed mix.

At the other extreme of the spectrum of input substitutability in production are Isoquants that are at a right angle, as in the right panel of Figure (1). In this case labor and capital are perfect complements. That is, labor and capital must be used in the fixed proportion of 2K/1L. In this case there is zero substitutability between labor and capital in production.
Examples of perfect complementary inputs are certain chemical processes that require basic elements (chemicals) to be combined in a specified fixed proportion, engine and body for automobiles, two wheels and a frame for bicycles, and so on. In these cases, inputs can be used only in the fixed proportion specified (i.e., there is no possibility of substituting one input for another in production).

Figure (2)



Although perfect substitutability and perfect complementarity's of inputs in production are possible, in most cases Isoquants exhibit some curvature (i.e., inputs are imperfect substitutes), as shown in Figure (2). This means that in the usual production situation, labor can be substituted for capital to some degree. The smaller is the degree of curvature of the Isoquant, the more easily inputs can be substituted for each other in production. In addition, when the Isoquant has some curvature, the ability to substitute labor for capital (or vice versa) diminishes as more and more labor is substituted for capital. This is indicated by the declining absolute slope of the Isoquant or marginal rate of technical substitution (MRTS) as we move down along an Isoquant. The ability to substitute one input for another in production is extremely important in keeping production costs down when the price of an input increases relative to the price of another.

MARGINAL RATE OF TECHNICAL SUBSTITUTION

The marginal rate of technical substitution (MRTS) is the amount of one input factor that must be substituted for one unit of another input factor to maintain a constant level of output. Algebraically

MRTS = ∂K/∂L = Slope of an Isoquant (1)

The marginal rate of technical substitution usually diminishes as the amount of substitution increases. In Figure 1 (c), for example, as more and more labor is substituted for cloth, the increment of labor necessary to replace cloth increases. At the extremes, Isoquants may even become positively sloped, indicating that the range over which input factors can be substituted for each other is limited. A classic example is the use of land and labor to produce a given output of grain. At some point, as labor is substituted for land, the farmers will trample the grain. As more labor is added, more land eventually must be added if grain output is to be maintained. The input substitution relation indicated by the slope of a production Isoquant is directly related

to the concept of diminishing marginal productivity. The marginal rate of technical substitution is equal to –1 times the ratio of the marginal products of the input factors [MRTS = $-1(MP_L / MP_K)$].

Along any Isoquant the total differential of the production function must be zero (output is fixed along an Isoquant). Thus, for the production function given by **Q = f (L, K),** setting the total differential equal to zero gives:

∂Q/∂LdL + ∂Q/∂KdK = 0
And, rearranging terms, we get:

∂Q/∂KdK = (-)∂Q/∂LdL

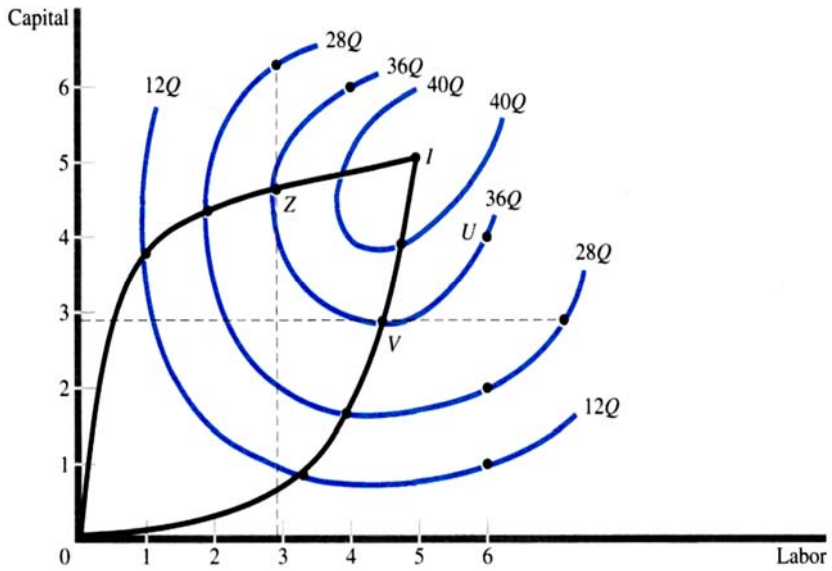
$-MP_L / MP_K = DK / dL$

Or **MRTS = $dK/dL = (-)MP_L/MP_K$ Slope of an Isoquant (2)**

The slope of a production Isoquant such as in Equation (1) and as derived in Equation (2) is equal to ∂K/∂L and is determined by the ratio of the marginal products of both inputs.

RATIONAL LIMITS OF INPUT SUBSTITUTION
RIDGE LINES
It is irrational for a firm to combine resources in such a way that the marginal product of any input is negative, because this implies that output could be increased by using less of that resource. From Equation (2), we can see that if the inputs L and K are combined in proportions such that the marginal product of either factor is negative, then the slope of the production Isoquant will be positive. For a production Isoquant to be positively sloped, one of the input factors must have a negative marginal product. Input combinations lying along a positively sloped portion of a production Isoquant are irrational and would be avoided by the firm.

Figure (3)



While the Isoquants in Figure (2) (repeated in Figure (3)) have positively sloped portions, these portions are irrelevant. That is, the firm would not operate on the positively sloped portion of an Isoquant because it could produce the same level of output with less capital and less labor. In Figure (3), the rational limits of input substitution are where the Isoquants become positively sloped. Limits to the range of substitutability of L for K are indicated by the points of tangency between the Isoquants and a set of lines drawn perpendicular to the Y-axis. Limits of economic substitutability of K for L are shown by the tangents of lines perpendicular to the X-axis. Maximum and minimum proportions of K and L that would be combined to produce each level of output are determined by points of tangency between these lines and the production Isoquants.

It is irrational to use any input combination outside these tangents, or **ridge lines**, as they are called. Such combinations are irrational because the marginal product of the relatively more abundant input is negative outside the ridge lines. The addition of the last unit of the excessive input factor actually reduces output. Obviously, it would be irrational for a firm to buy and employ additional units that cause production to decrease. Ridge lines separate the relevant (i.e., negatively sloped) from the irrelevant (or positively sloped) portions of the Isoquants. In Figure (3), ridge line 0VI joins points on the various Isoquants where the Isoquants have zero slope. The Isoquants are negative sloped to the left of this ridge line and positively sloped to the right. On the other hand, ridge line OZI, joins points where the Isoquants have infinite slope. The Isoquants are negatively sloped to the right of this ridge line and positively sloped to the left. Hence the economic region of production is given by the negatively sloped segment of Isoquants between ridge lines OVI and OZI. The firm will not produce in the positively sloped portion of the Isoquants because it could produce the same level of output with both less labor and less capital.

OPTIMAL COMBINATION OF INPUTS

As an Isoquant shows the various combinations of labor and capital that a firm can use to produce a given level of output. An **Isocost line** shows the various combinations of inputs that a firm can purchase or hire at a given cost. By the use of Isocost and Isoquants, we determine the optimal input combination for the firm to maximize profits.

ISOCOST LINES

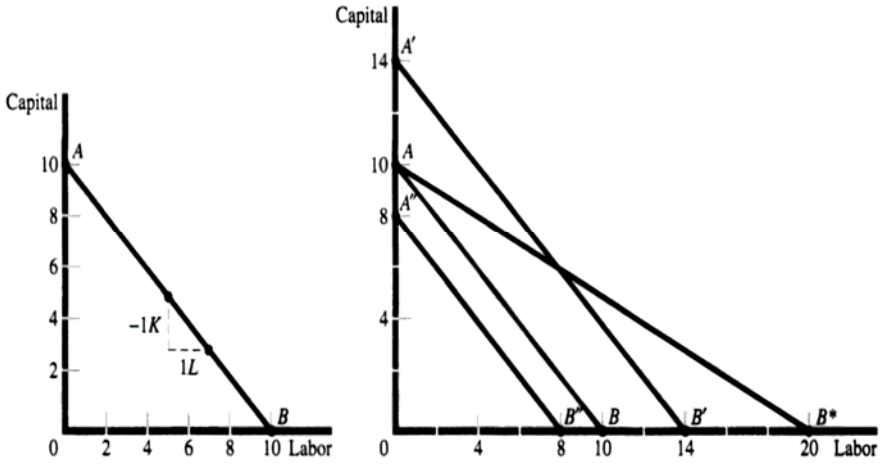
Optimal input proportions can be found graphically for a two-input, single-output system by adding an Isocost curve or budget line, a line of constant costs, to the diagram of production Isoquants. Each point on the Isocost curve represents a combination of inputs, say, L and K, whose cost equals a constant expenditure.

Suppose that a firm uses only labor and capital in production. The total costs or expenditures of the firm can then be represented by

$$C = wL + rK \qquad (3)$$

Where C is total costs, w is the wage rate of labor, L is the quantity of labor used, r is the rental price of capital, and K is the quantity of capital used. Thus, Equation (3) postulates that the total costs of the firm (C) equals the sum of its expenditures on labor (wL) and capital (rK). Equation (3) is the general equation of the firm's Isocost line or equal cost line. It shows the various combinations of labor and capital that the firm can hire or rent at a given total cost. For example, if C = $100, w = $10, and r = $10, the firm could either hire 10L or rent 10K, or any combination of L and K shown on Isocost line AB in the left panel of Figure (4). For each unit of capital the firm gives up, it can hire one additional unit of labor. Thus, the slope of the Isocost line is -1.

Figure (4)



By subtracting wL from both sides of Equation (3) and then dividing by r, we get the general equation of the Isocost line in the following more useful form:

$$K = \frac{C}{r} - \frac{w}{r} L$$

Where C/r is the vertical intercept of the Isocost line and w/r is its slope.
A different total cost by the firm would define a different but parallel Isocost line, while different relative input prices would define an Isocost line with a different slope. For example, an increase in total expenditures to C' = $140 with unchanged w = r = $10 would give Isocost line A' B' in the right panel of Figure (4), with vertical intercept C'/r = $140/$10 = 14K and slope of -w/r = -$10/$10= -1.

$$MRTS = \frac{w}{r}$$

Since the MRTS = MP_L/MP_K, we can rewrite the condition for the optimal combination of inputs as:

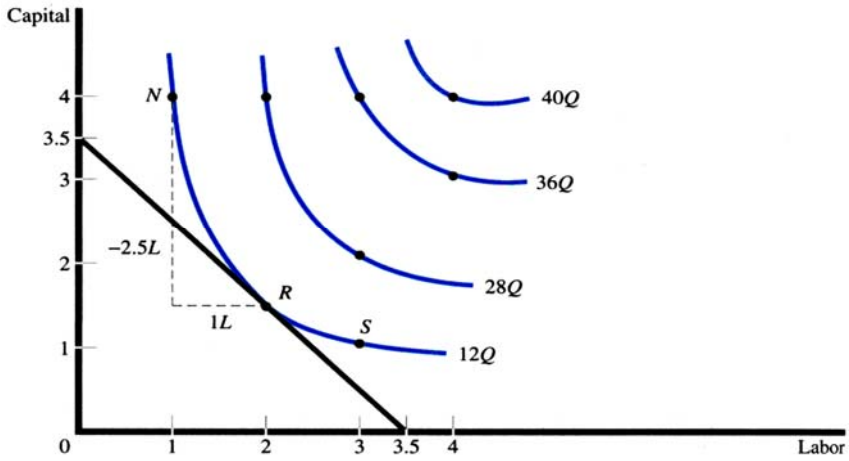
$$\frac{MP_L}{MP_K} = \frac{w}{r}$$

Cross-multiplying, we get:

$$\frac{MPL}{w} = \frac{MPK}{r}$$

Figure (5)

MRTS = -(-2.5/1) = 2.5



At the point of optimal input combination, Isocost and the Isoquant curves are tangent and have equal slope. The slope of an Isocost curve equals *–PX*/*PY* i-e w/r. The slope of an Isoquant curve equals the marginal rate of technical substitution of one input factor for another when the quantity of production is held constant. Therefore, for optimal input combinations, the ratio of input prices must equal the ratio of input marginal products, as is shown:

$$\frac{w}{r} = \frac{MP_L}{MP_K}$$

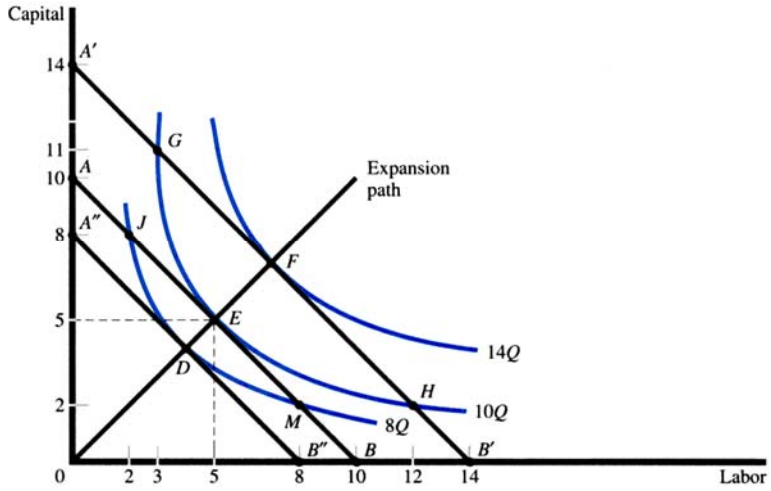
Alternatively, marginal product to price ratio must be equal for each input:

$$MP_X/P_X = MP_Y/P_Y$$

EXPANSION PATH

By connecting points of tangency between Isoquants and budget lines (*points D, E, and F*), an **expansion path** is identified that depicts optimal input combinations as the scale of production expands. For example, line ODEF in Figure (6) is the expansion path for the firm. It shows that the minimum cost of reaching Isoquants 8Q, 10Q and 14Q are $80, $100, and $140, given by the points of tangency of Isoquants and Isoquants (i.e., joining points of optimal input combinations. It also shows that with total cost of $80, $100, and $140 the maximum output that the firm can produce are 8Q, 10Q and 14Q respectively.

Figure (6)


Capital
14
11
10
8
5
2
0
A′
G
A
A″
J
Expansion
path
F
E
D
14Q
H
10Q
M
8Q
B″
B
B′
2
3
5
8
10
12
14
Labor

Lesson 16

PRODUCTION ANALYSIS AND ESTIMATION (CONTINUED 2)

RETURNS TO SCALE

Closely related to the productivity of individual inputs is the question of how a proportionate increase in all inputs will affect total production. **Constant returns to scale** exist when a given percentage increase in all inputs leads to that same percentage increase in output. **Increasing returns to scale** are prevalent if the proportional increase in output is larger than the underlying proportional increase in inputs. If output increases at a rate less than the proportionate increase in inputs, **decreasing returns to scale** are present.

Returns to scale refers to the degree by which output Changes as a result of a given change in the quantity of all inputs used in production. There are three types of returns to scale constant, increasing, and decreasing. If the quantity of all inputs used in production is increased by a given proportion, we have constant returns to scale if output increases in the same proportion increasing returns to scale if output increases by a greater proportion; and decreasing returns to scale if output increases by a smaller proportion. That is, suppose that starting with the general production function.

Q = f (L, K) (1)

We multiply Land K by h, and Q increases by 1, as indicated in Equation (1):

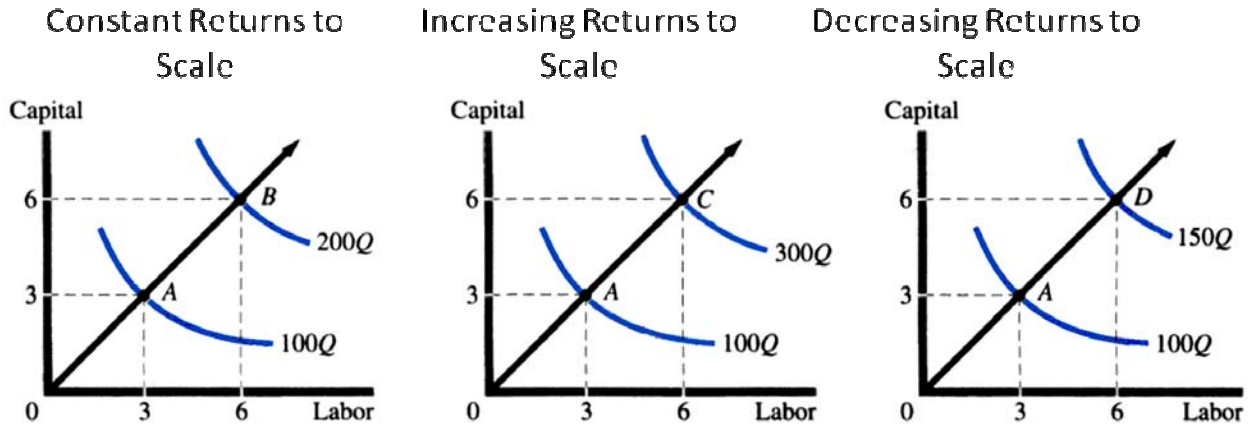
λQ = f(hL, hK) (2)

We have constant, increasing, or decreasing returns to scale, respectively, depending on whether $\lambda = h$, $\lambda > h$, or $\lambda < h$.

For example, if all inputs are doubled, we have constant, increasing, or decreasing returns to scale, respectively, if output doubles, more than doubles, or less than doubles. This is shown in Figure (1). In all three panels of Figure (1) we start with the firm using 3L and 3K and producing 100Q (point A). By doubling inputs to 6L and 6K, the left panel shows that output also doubles to 200Q (point B), so that we have constant returns to scale.

Increasing returns to scale arise because as the scale of operation increases, a greater division of labor and specialization can take place and more specialized and productive machinery can be used. Decreasing returns to scale, on the other hand, arise primarily because as the scale of operation increases, it becomes ever more difficult to manage the firm effectively and coordinate the various operations and divisions of the firm.

Figure (1)



A more general condition is a production function with first increasing, then decreasing, returns to scale. The region of increasing returns is attributable to specialization. As output increases, specialized labor can be used and efficient, large-scale machinery can be used in the production process. Beyond some scale of operation, however, further gains from specialization are limited, and coordination problems may begin to increase costs substantially. When coordination expenses more than offset additional benefits of specialization, decreasing returns to scale set in.

For certain production functions, called **homogeneous production functions**, when each input factor is multiplied by a constant *k*, the constant can be completely factored out of the production function expression. Following a *k*-fold increase in all inputs, the production function takes the form $hQ = k^r f(X, Y, Z)$. The exponent *r* provides the key to returns-to-scale estimation. If $r = 1$, then $h = k$ and the function exhibits constant returns to scale. If $r > 1$, then $h > k$, indicating increasing returns to scale, whereas $r < 1$ indicates $h < k$ and decreasing returns to scale.

Q = f(X,Y,Z)
hQ = f(kX,kY, kZ)
$hQ = k^r f(X,Y,Z)$ Where r is degree of homogeneity (3)

If k = h, then f has constant returns to scale.
If k > h, then f has increasing returns to scale.
If k < h, then f has decreasing returns to scale.

OUTPUT ELASTICITY AND RETURNS TO SCALE

Returns to scale can be accurately determined for any production function through analysis of output elasticities. **Output elasticity**, E_Q, is the percentage change in output associated with a 1 percent change in all inputs and a practical means for returns to scale estimation. Letting *X* represent all input factors,

$$E_Q = \frac{\text{Percentage Change in Output } (Q)}{\text{Percentage Change in All Inputs } (X)} = \partial Q/Q \div \partial X_i/X_i$$

Where *X* refers to capital, labor, energy, and so on, then the following relations hold:

Table (1)

If	Then	Returns to Scale
Percentage change in Q >Percentage change in X	$E\ Q > 1$	Increasing
Percentage change in Q = Percentage change in X	$E\ Q = 1$	Constant
Percentage change in Q < Percentage change in X	$E\ Q < 1$	Diminishing

Thus, returns to scale can be analyzed by examining the relationship between the rate of increase in inputs and the quantity of output produced.

ESTIMATION OF PRODUCTION FUNCTIONS

Types of production functions depending on their degree of power:

- short run, Linear: one fixed factor, one variable factor
  $Q = f(L)_K$
- cubic: increasing marginal returns followed by decreasing marginal returns, all 3 stages of production
  $Q = a + bL + cL^2 – dL^3$
- quadratic: diminishing marginal returns but no Stage I
  $Q = a + bL - cL^2$

Given enough input/output observations, either over time for a single firm or at a single point in time for a number of firms in an industry, regression techniques can be used to estimate the parameters of production functions.

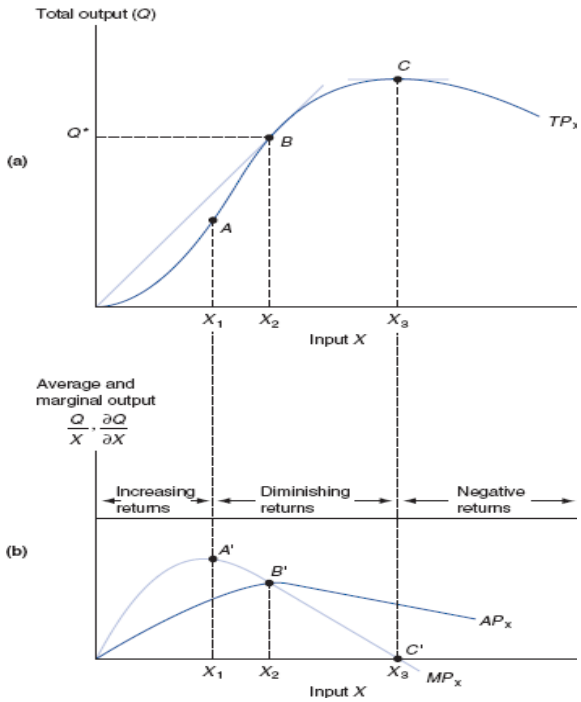
CUBIC PRODUCTION FUNCTIONS

From a theoretical standpoint, the most appealing functional form for production function estimation might be cubic, such as the equation

$$Q = a + bXY + cX^2Y + dXY^2 – eX^3Y – fXY^3 \quad (4)$$

This form is general in that it exhibits stages of first increasing and then decreasing returns to scale. The marginal products of the input factors exhibit a pattern of first increasing and then decreasing returns, as is illustrated in Figure (2).

Figure (2)



Frequently, however, real-world data do not exhibit enough dispersion to indicate the full range of increasing and then decreasing returns. In these cases, simpler functional specifications can be used to estimate production functions. The full generality of a cubic function may be unnecessary, and an alternative linear or log-linear model specification can be usefully applied in empirical estimation.

POWER PRODUCTION FUNCTIONS

One function commonly used in production studies is the **power production function**, a multiplicative relation between output and input that takes the form

$$Q = b_0 X^{b1} Y^{b2} \quad (5)$$

Power functions have properties that are useful in empirical research. Power functions allow the marginal productivity of a given input to depend on the levels of *all* inputs used a condition that often holds in actual production systems. Power functions are also easy to estimate in log-linear form using least squares regression analysis because Equation (5) is mathematically equivalent to:

$$\log Q = \log b0 + b1 \log X + b2 \log Y \quad (6)$$

Returns to scale are also easily calculated by summing the exponents of the power function or, alternatively, by summing the log-linear model coefficient estimates. As seen in Figure (6), if the sum of power function exponents is less than 1, diminishing returns are indicated. A sum greater than 1 indicates increasing returns. If the sum of exponents is exactly 1, returns to scale are constant. Power functions have been successfully used in a large number of empirical production studies since Charles W. Cobb and Paul H. Douglas's pioneering work in the late 1920s. The impact of their work is so great that power production functions are frequently referred to as Cobb-Douglas production functions.

COBB-DOUGLAS PRODUCTION FUNCTION

The production function most commonly used in empirical estimation is the power unction of the form

$$Q = A K^a L^b \quad (7)$$

Where Q, K, and L refer, respectively, to the quantities of output, capital, and labor, and A, a, and b are the parameters to be estimated empirically. Equation (7) is often referred to as the Cobb-Douglas production function in honor of Charles W. Cobb and Paul H. Douglas, who introduced it in the 1920s.

The Cobb-Douglas production function has several useful properties.

- First, the marginal product of capital and the marginal product of labor depend on both the quantity of capital and the quantity of labor used in production, as is often the case in the real world.
- Second, the exponents of K and L (i.e., a and b) represent, respectively, the output elasticity of labor and capital (E_K and E_L), and the sum of the exponents (i.e., a + b) measures the returns to scale. If a + b = 1, we have constant returns to scale; if a + b > 1, we have increasing returns to scale and if a + b < 1, we have decreasing returns to scale.
- Third, the Cobb-Douglas production function can be estimated by regression analysis by transforming it into linear in the logarithms:

$$\ln Q = \ln A + a \ln K + b \ln L \quad (8)$$

- In production theory it is assumed that technology is fixed, however, the data fitted by the researchers may span a period over which technology has changed and improved, one of the independent variable could represent technical change (a time-series) and thus adjust the function to take technology into consideration.
- Finally, the Cobb-Douglas production function can easily be extended to deal with more than two inputs (say, capital, labor, and natural resources or capital, production labor, and non production labor).

The Cobb-Douglas production function can be estimated either from data for a single firm, industry, or nation over time (i.e., using time-series analysis), or for a number of firms, industries, or nations at one point in time (i.e., using cross-sectional data). In either case, the researcher faces three potential difficulties:

1.If the firm produces a number of different products, output may have to be measured in monetary rather than in physical units, and this will require deflating the value of output by the price index in time-series analysis or adjusting for price differences for firms located in different regions in cross-sectional analysis.
2. Only the capital consumed in the production of the output should be counted, ideally. Since machinery and equipment are of different types and ages and productivities, however, the total stock of capital in existence has to be used instead.
3. In time series analysis a time trend is also usually included to take into consideration technological changes over time, while in cross sectional analysis we must ascertain that all firms or industries use the same technology (the best available).
4.Cannot show MP going through all three stages in one specification, also it cannot show a firm or industry passing through increasing, constant, and decreasing returns to scale.

Cobb-Douglas production function is the most frequently used production function both for individual firms and for the entire economy as a whole (aggregate production function). One example is a study by Robert N. Mefford (1986). This study dealt with a sample of plants of multinational consumer goods manufactured. Time series and cross-sectional data were

combined to obtain 127 observations over the period (1975-1982). Management was measured as a performance ranking of on the basis of 3 criteria:

1. output goal attainment
2. cost over- or under -fulfillment
3. quality level of output

The results showed that the Management variable was found statistically significant.
Given Cobb-Douglas function: $\mathbf{Q = AL^{\alpha}K^{\beta}}$

if $\alpha + \beta > 1$, IRTS
if $\alpha + \beta = 1$, CRTS
if $\alpha + \beta < 1$, DRTS

$\mathbf{MP_L = \partial Q/\partial L = \alpha AL^{\alpha - 1}K^{\beta}}$
$\mathbf{MP_K = \partial Q/\partial K = \beta AL^{\alpha} K^{\beta - 1}}$
$\mathbf{MP_L = \partial Q/\partial L = \alpha AL^{\alpha - 1}K^{\beta}}$
$\mathbf{= \alpha . Q/L}$
Since $\mathbf{AP_L = Q/L}$
Hence $\mathbf{AP_L > MP_L}$ **, Similarly**
$\mathbf{AP_K > MP_K}$

$$\varepsilon_L = \partial Q/Q \div \frac{\partial L/L}{L} = \frac{\alpha Q}{Q}\frac{L}{} = \alpha$$

$$\varepsilon_K = \partial Q/Q \div \frac{\partial K/K}{K} = \frac{\beta Q}{Q}\frac{K}{} = \beta$$

Since $\mathbf{MRTS = - MP_L/MP_K}$
$\mathbf{MRTS = -\alpha AL^{\alpha - 1}K^{\beta} / \beta AL^{\alpha}K^{\beta - 1}}$
Using rules for indices:
$\mathbf{MRTS = - \alpha/\beta . K / L}$

Lesson 17

PRODUCTION ANALYSIS AND ESTIMATION (CONTINUED 3)

DUALITY: PROFIT MAXIMIZATION VS COST MINIMIZATION

Individuals maximize utility subject to a budget constraint
Dual problem: individuals minimize the expenditure needed to achieve a given level of utility
Firms minimize cost of inputs to produce a given level of output.
Dual problem: firms maximize output for a given cost of inputs purchased
The difference between profit maximization and cost minimization is simple. Cost minimization requires efficient resource use, as reflected by optimal input proportions. Profit maximization requires efficient resource use and Production of an optimal level of output, as made possible by the optimal employment of all inputs.

PRODUCTION ANALYSIS WITH CALCULUS

The graphical approach provides a useful interpretation of constrained optimisation and enables us to justify some familiar microeconomic results. However, it does not offer a practical way of actually solving such problems, since it is difficult to produce an accurate Isoquant map from any given production function.

The most popular algebraic method for solving this problem is the *Lagrange multiplier method*.

$$L^* = wL + rK + \lambda\left(Y - f\left(K, L\right)\right)$$

$$L^* = f\left(L, K\right) + \lambda\left(C - wL + rL\right)$$

$$\lambda = \frac{\text{marginal benefit of } x_i}{\text{marginal cost of } x_i}$$

We use the **Lagrangian multiplier** method to examine the condition for a firm to be (1) maximizing output for a given cost outlay and (2) minimizing the cost of producing a given output.
Suppose that a firm that uses labor *(L)* and capital *(K)* in production wants to determine the amount of labor and capital that it should use in order to maximize the output *(Q)* produced with a given cost outlay *(C*)*. That is, the firm wants to

Maximize **Q = f(L, K)** **(1)**

Subject to **C* = wL +rK** **(2)**

Where w is the wage of labor and r is the rental price of capital. This constrained maximization problem can be solved by the Lagrangian multiplier method.
To do so, we first form the Lagrangian function:

Z =f (L, K) + λ(C* - wL - rK) **(3)**

To maximize Z, we then find the partial derivatives of Z with respect to L, K, and **λ** and set them equal to zero. That is,

$$\frac{\delta Z}{\delta L} = \frac{\delta f}{\delta L} - \lambda W = 0 \qquad (4)$$

$$\frac{\delta Z}{\delta K} = \frac{\delta f}{\delta K} - \lambda r = 0 \qquad (5)$$

$$\frac{\delta Z}{\delta \lambda} = C^* - wL - rK = 0 \qquad (6)$$

By substituting MP_L for δf / δL and MP_K for δf / δK, transposing w and r to the right of the equals sign, and dividing Equation (4) by Equation (5), we have.

$$\frac{MP_L}{MP_K} = \frac{W}{r}$$

$$\frac{MP_L}{W} = \frac{MP_K}{r}$$

That is, the firm should hire labor and capital so that the marginal product per dollar spent on each input (λ) is equal. This is the first-order condition for output maximization for the given cost outlay or expenditure of the firm. The second-order condition is for the Isoquant to be convex to the origin.

CONSTRAINED COST MINIMIZATION

Suppose, on the other hand, that ,the firm of the previous section wants to determine the amount of labor and capital to use to minimize the cost of producing a given level of output (Q*).The problem would then be

Minimize $C=wL+rK$ (7)

Subject to $Q^* = f(L,K)$ (8)

This constrained cost minimization problem can also be solved by the Lagrangian multiplier method. To do so, we first form the Lagrangian function:

$$Z' = wL + rK + \lambda'[Q^* - f(L, K)] \qquad (9)$$

To minimize Z', we then find the partial derivatives of Z' with respect to L, K, and λ', and set them equal to zero. That is,

$$\frac{\delta Z'}{\delta L} = \frac{\lambda' \delta f}{\delta L} = 0$$

$$\frac{\delta Z'}{\delta K} = r - \frac{\lambda' \delta f}{\delta K} = 0$$

$$\frac{\delta Z'}{\delta \lambda'} = Q^* - f(L, K) = 0$$

By substituting MP_L for δf / δL and MP_K for δf / δK, transposing them to the right of the equal sign, and dividing Equation 6-30 by Equation 6-31, we have

$$\frac{w}{r} = \frac{MP_L}{MP_K}$$

or

$$w / MP_L = r / MP_K \qquad (10)$$

Each term in Equation (10) equals λ' and refers to the marginal cost in terms of labor and capital. That is, Equation (10) assumes that to minimize the costs of producing Q*, the firm should use labor and capital in such a way that the extra cost of producing an additional unit of output is the same whether the firm produces it with more labor or more capital.

PROFIT MAXIMIZATION

In general, the firm will want to determine the amount of labor and capital needed to maximize profits rather than to maximize output or minimize costs. Total profit (л) is

$$\Pi = TR - TC \qquad (11)$$
$$= p * Q - wL - rK \qquad (12)$$

Since Q = f(L, K), we can rewrite the profit function as

$$\Pi = P.\ f(L, K) - wL - rK \qquad (13)$$

To determine the amount of labor and capital that the firm should use in order to maximize profits, we take the partial derivatives of Equation (13) with respect to Land K and set them equal to zero. That is,

$$\frac{\delta л}{\delta L} = \frac{P \delta f}{\delta L} - w = 0$$

$$\frac{\delta л}{\delta K} = \frac{P \delta f}{\delta K} - r = 0$$

Assuming that the price of the final commodity (P) is constant so that it is equal to marginal revenue (MR), we can rewrite Equations 6-38 and 6-39 as

$$(MP_L)(MR) = MRP_L = w \qquad (14)$$
$$(MP_K)(MR) = MRP_K = r \qquad (15)$$

That is, in order to maximize profits, the firm should hire labor and capital until the marginal revenue product of labor equals the wage rate, and until the marginal revenue product of capital is equal to the rental price of capital.

Dividing Equation (14) by Equation (15), we get the following expression:

$$\frac{MP_L}{MP_K} = \frac{w}{r}$$

Cross-multiplying the above expression gives the condition for the optimal combination of inputs given by:

$$\frac{MP_L}{w} = \frac{MP_K}{r}$$

That is, hiring labor and capital so that the above expressions hold which implies optimal input combinations will be satisfied.

ECONOMIC INTERPRETATION OF λ

- A high value of λ indicates that Q could be increased substantially by relaxing the constraint
- A low value of λ indicates that there is not much to be gained by relaxing the constraint
- λ = 0 implies that the constraint is not binding

EXAMPLE: OUTPUT MAXIMIZATION

Maximize

$Q = 12L^{0.5}K^{0.5}$ Subject to: C = 25L + 50K;

When cost constraint is $1414.

1414 - 25L - 50K = 0

Solution:

$L = 12L^{0.5}K^{0.5} + \lambda [1414 - 25L - 50K]$

FOC

$LL = 6L^{-0.5}K^{0.5} - 25\lambda = 0$ (1)

$LK = 6L^{0.5}K^{-0.5} - 50\lambda = 0$ (2)

$L\lambda = 1414 - 25L - 50K = 0$ (3)

K/L = ½ so substitute K = L/2 in Eq (3), we get

1414 = 25L + 50(L/2) or 50L = 1414 so; L = 28.28 and K = 14.14

$Q = 12 (28.28)^{0.5}(14.14)^{0.5} = 12(5.31)(3.76)$

Q = 240 units

Example: Cost Minimization

Minimize

C = 25L + 50K Subject to: $Q = 12L^{0.5}K^{0.5}$

When production constraint is 240 units

Solution:

$L = 25L + 50K + \lambda [240 - 12L^{0.5}K^{0.5}]$

FOC

$LL = 25 - \lambda 6L^{-0.5}K^{0.5} = 0$ (1)

$LK = 50 - \lambda 6L^{0.5}K^{-0.5} = 0$ (2)

$L \lambda = 240 - 12L^{0.5}K^{0.5} = 0$ (3)

K/L = ½ so substitute K = L/2 in Eq (3), we get

$240 = 12L^{0.5}(L/2)^{0.5}$;L = 28.28 and K = 14.14

Therefore C = 25L + 50K = 25(28.28) + 50(14.14) = $1414

Lesson 18

COST ANALYSIS AND ESTIMATION

Cost is an important consideration in managerial decision making, and cost analysis is an essential and major aspect of managerial economics. Cost analysis is made difficult by the effects of unforeseen inflation, unpredictable changes in technology, and the dynamic nature of input and output markets

THE NATURE OF COSTS

EXPLICIT AND IMPLICIT COSTS

Typically, the costs of using resources in production involve both out-of-pocket costs, or explicit costs, and other non-cash costs, called implicit costs. Wages, utility expenses, payment for raw materials, interest paid to the holders of the firm's bonds, and rent on a building is all examples of explicit expenses. The implicit costs associated with any decision are much more difficult to compute. These costs do not involve cash expenditures and are therefore often overlooked in decision analysis.

One crucial distinction in the analysis of costs is between explicit and implicit costs. **Explicit costs** refer to the actual expenditures of the firm to hire, rent, or purchase the inputs; it requires in production. These include the wages to hire labor the rental price of capital, equipment, and buildings, and the purchase price of raw materials and semi finished products. **Implicit costs** refer to the value of the inputs owned and used by the firm in its own production activity. Even though the firm does not incur any actual expenses to use these inputs, they are not free, since the firm could sell or, rent them out to other firms. The amount for which the firm could sell or rent out these owned inputs to other firms represents a cost of production of the firm owning and using them. Implicit costs include the highest salary that the entrepreneur could earn in his or her best alternative employment (say, in managing another firm).

In economics, both explicit and implicit costs must be considered. That is, in measuring production costs, the firm must include the alternative or opportunity costs of all inputs, whether purchased or owned by the firm. The reason is that the firm could not retain a hired input if it paid a lower price for the input than another firm. Similarly, it would not pay for a firm to use an owned input if the value (productivity) of the input is greater to another firm. These **economic costs** must be distinguished from **accounting costs**, which refer only to the firm's actual expenditures or explicit costs incurred for purchased or rented inputs. Accounting or historical costs are important for financial reporting by the firm and for tax purposes. For managerial decision making purposes, however, economic or opportunity costs are the relevant cost concept that must be used.

Suppose that a firm purchased a machine for $1,000. If the estimated life of the machine is 10 years and the accountant uses a straight-line depreciation method (that is, $100 per year), the accounting value of the machine is zero at the end of the tenth year. Suppose, however, that the machine can still be used for (i.e., it would last) another year and that the firm could sell the machine for $120 at the end of the tenth year or use it for another year. The cost of using the machine is zero as far as the accountant is concerned (since the machine has already been fully depreciated), but it is $120 for the economist. Again, incorrectly assigning a zero cost to the use of the machine would be wrong from an economics point of view and could lead to wrong managerial decisions.

OPPORTUNITY COST CONCEPT

Opportunity cost is the foregone value associated with the current rather than next-best use of an asset. In other words, cost is determined by the highest-valued *opportunity* that must be foregone to allow current use. The cost of aluminum used in the manufacture of soft drink containers, for example, is determined by its value in alternative uses. Soft drink bottlers must pay an aluminum price equal to this value, or the aluminum will be used in the production of alternative goods, such as airplanes, building materials, cookware, and so on.

Similarly, if a firm owns capital equipment that can be used to produce either product A or product B, the relevant cost of product A includes the profit of the alternative product B that cannot be produced because the equipment is tied up in manufacturing product A.

The opportunity cost concept explains asset use in a wide variety of circumstances. Gold and silver are pliable yet strong precious metals. As such, they make excellent material for dental fillings. However, when speculation drove precious metals prices skyrocketing during the 1970s, plastic and ceramic materials became a common substitute for dental gold and silver.

More recently, lower market prices have again allowed widespread dental use of both metals. Still, dental customers must be willing to pay a price for dental gold and silver that is competitive with the price paid by jewelry customers and industrial users.

HISTORICAL VERSUS CURRENT COSTS

When costs are calculated for a firm's income tax returns, the law requires use of the actual dollar amount spent to purchase the labor, raw materials, and capital equipment used in production. For tax purposes, **historical cost**, or actual cash outlay, is the relevant cost. Despite their usefulness, historical costs are not appropriate as a sole basis for many managerial. Current costs are typically much more relevant. **Current cost** is the amount that must be paid under prevailing market conditions. Current cost is influenced by market conditions measured by the number of buyers and sellers, the present state of technology, inflation, and so on. For assets purchased recently, historical cost and current cost are typically the same. For assets purchased several years ago, historical cost and current cost are often quite different. Since World War II, inflation has been an obvious source of large differences between current and historical costs throughout most of the world. With an inflation rate of roughly 5 percent per year, prices double in less than 15 years and triple in roughly 22 years. Land purchased for $50,000 in 1970 often has a current cost in excess of $200,000.

REPLACEMENT COST

Although it is typical for current costs to exceed historical costs, this is not always the case. Computers and many types of electronic equipment cost much less today than they did just a few years ago. In many high-tech industries, the rapid advance of technology has overcome the general rate of inflation. As a result, current costs are falling. Current costs for computers and electronic equipment are determined by what is referred to as **replacement cost**, or the cost of duplicating productive capability using current technology. For example, the value of used personal computers tends to fall by 30 to 40 percent per year. In valuing such assets, the appropriate measure is the much lower replacement cost—not the historical cost.

MARGINAL COST VERSUS INCREMENTAL COST

In discussing production costs, we must also distinguish between marginal cost and incremental cost. Marginal cost refers to the change In total cost for a 1-unIt change in output.

For example, if total cost is $140 to produce 10 units of output and $150 to produce 11 units of output, the marginal cost of the eleventh unit is $10. Incremental cost, on the other hand, is a broader concept and refers to the change in total costs from implementing a particular management decision, such as the introduction of a new product line, the undertaking of a new advertising campaign, or production shift.

SUNK COSTS

Inherent in the incremental cost concept is the principle that any cost not affected by a decision is irrelevant to that decision. A cost that does not vary across decision alternatives is called a **sunk cost**; such costs do not play a role in determining the optimal course of action. For example, suppose a firm has spent $5,000 on an option to purchase land for a new factory at a price of $100,000. Also assume that it is later offered an equally attractive site for $90,000. What should the firm do? The first thing to recognize is that the $5,000 spent on the purchase option is a sunk cost that must be ignored. As any costs not affected by available decision alternatives are sunk and irrelevant.

SHORT RUN AND LONG RUN COSTS

The **short run** is the operating period during which the availability of at least one input is fixed. In the **long run**, the firm has complete flexibility with respect to input use. In the short run, operating decisions are typically constrained by prior capital expenditures. In the long run, no such restrictions exist. At least one input is fixed in the short run while all inputs are variable in the long run.

Long-run cost curves are called *planning curves*; short-run cost curves are called *operating curves*. In the long run, plant and equipment are variable, so management can plan the most efficient physical plant, given an estimate of the firm's demand function. Once the optimal plant has been determined and the resulting investment in equipment has been made, short-run operating decisions are constrained by these prior decisions.

FIXED AND VARIABLE COSTS

Fixed costs do not vary with output. These costs include interest expenses, rent on leased plant and equipment, depreciation charges associated with the passage of time, property taxes, and salaries for employees not laid off during periods of reduced activity. Because all costs are variable in the long run, long-run fixed costs always equal zero. **Variable costs** fluctuate with output. Expenses for raw materials, depreciation associated with the use of equipment, the variable portion of utility charges, some labor costs, and sales commissions are all examples of variable expenses. In the short run, both variable and fixed costs are often incurred. In the long run, all costs are variable while fixed cost is a short-run concept.

SHORT-RUN COST CURVES

A **short-run cost curve** shows the minimum cost impact of output changes for a specific plant size and in a given operating environment. Such curves reflect the optimal or least-cost input combination for producing output under fixed circumstances. Wage rates, interest rates, plant configuration, and all other operating conditions are held constant.

Any change in the operating environment leads to a *shift* in short-run cost curves. For example, a general rise in wage rates leads to an upward shift; a fall in wage rates leads to a downward shift. Such changes must not be confused with *movements along* a given short-run cost curve caused by a change in production levels. For an existing plant, the short-run cost curve

illustrates the minimum cost of production at various output levels under current operating conditions. Short-run cost curves are a useful guide to operating decisions.

SHORT-RUN COST CATEGORIES

Both fixed and variable costs affect short-run costs. Total cost at each output level is the sum of total fixed cost (a constant) and total variable cost. Using *TC* to represent total cost, *TFC* for total fixed cost, *TVC* for total variable cost, and *Q* for the quantity of output produced, various unit costs are calculated as follows:

Total Cost = TC = TFC + TVC

Average Fixed Cost = AFC = TFC/Q

Average Variable Cost = AVC = TVC/Q

Average Cost = AC = TC = AFC + AVC

Marginal Cost = MC = ∂TC/∂Q

Marginal cost is the change in cost associated with a one-unit change in output. Because fixed costs do not vary with output, fixed costs do not affect marginal costs. Only variable costs affect marginal costs. Therefore, marginal costs equal the change in total costs *or* the change in total variable costs following a one-unit change in output:

MC = ∂TC/∂Q = ∂TVC/∂Q

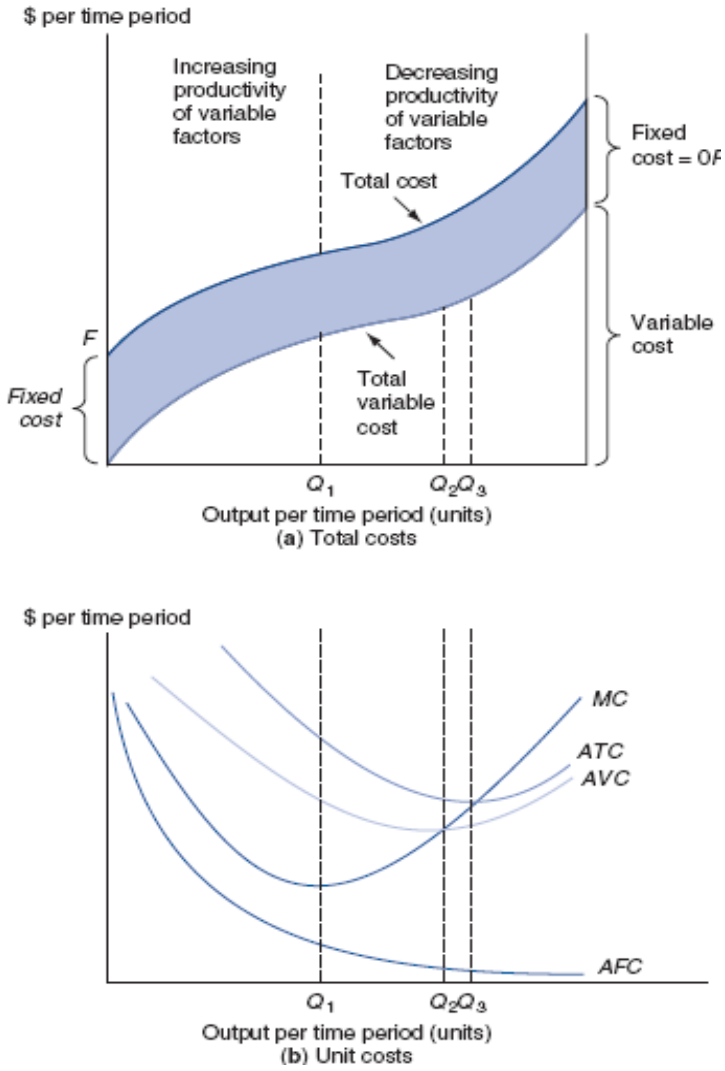
SHORT-RUN COST RELATIONS

Relations among short-run cost categories are shown in Figure 1. Figure 1(a) illustrates total cost and total variable cost curves. The shape of the total cost curve is determined entirely by the total variable cost curve. The slope of the total cost curve at each output level is identical to the slope of the total variable cost curve. Fixed costs merely shift the total cost curve to a higher level. This means that marginal costs are independent of fixed cost.

The shape of the total variable cost curve, and hence the shape of the total cost curve, is determined by the productivity of variable input factors employed. The variable cost curve in Figure 1 increases at a decreasing rate up to output level *Q*1, then at an increasing rate. Assuming constant input prices, this implies that the marginal productivity of variable inputs first increases, then decreases. Variable input factors exhibit increasing returns in the range from 0 to *Q*1 units and show diminishing returns thereafter. This is a typical finding. Past point Q_1, the law of diminishing returns operates, and the *TVC* curve faces upward (concave upwards) or rises at an increasing rate. Since *TC = TFC + TVC*, the *TC* curve has the same shape as the *TVC* but is *OF* (the amount of the *TFC*) level above it at each input level.

The relation between short-run costs and the productivity of variable input factors is also reflected by short-run unit cost curves, as shown in Figure 1(b). Marginal cost declines over the range of increasing productivity and rises thereafter. This imparts the familiar U-shape to average variable cost and average total cost curves. At first, marginal cost curves also typically decline rapidly in relation to the average variable cost curve and the average total cost curve. Near the target output level, the marginal cost curve turns up and intersects each of the *AVC* and *AC* short-run curves at their respective minimum points.

Figure 1



We can explain the U shape of the *AVC* curve as follows. With labor as the only (variable input, *TVC* for any output level (Q) equals the wage rate (w, which is assumed to be fixed) times the quantity of labor (L) used. Thus,

AVC = Average Variable Cost

Since TVC = WL

$AVC = TVC/Q = w/AP_L$

Since the average physical product of labor (AP_L or Q/L) usually rises first, reaches a maximum, and then falls, it follows that the *AVC* curve first falls, reaches a minimum, and then rises. Since the *AVC* curve is U-shaped, the *ATC* curve is also U-shaped. The *ATC* curve continues to fall after the *AVC* curve begins to rise as long as the decline in *AFC* exceeds the rise in *AVC*.

The U shape of the *MC* curve can similarly be explained as follows:

MARGINAL COST

$\Delta TC/\Delta Q = \Delta TVC/\Delta Q = W(\Delta L)/\Delta Q = w/MP_L$

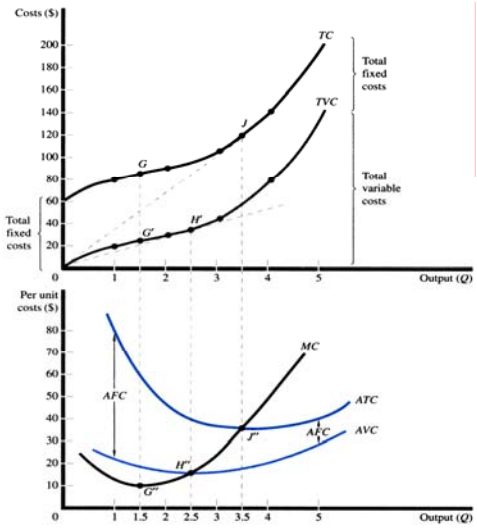
Since the marginal product of labor (MP_L or $\Delta Q/\Delta L$) first rises, reaches a maximum, and then falls, it follows that the *MC* curve first falls, reaches a minimum, and then rises. Thus, the rising portion of the *MC* curve reflects the operation of the law of diminishing returns.

Table 1

Q	TFC	TVC	TC	AFC	AVC	ATC	MC
0	$60	$0	$60	-	-	-	-
1	60	20	80	$60	$20	$80	$20
2	60	30	90	30	15	45	10
3	60	45	105	20	15	35	15
4	60	80	140	15	20	35	35
5	60	135	195	12	27	39	55

Table 1 shows the hypothetical short-run total and per-unit cost schedules of a firm. These schedules are plotted in Figure 2. From, column 2 of Table 1 we see that TFC are $60 regardless of the level of output. TVC (column 3) are zero when output is zero and rise as output rises. Up to point G' (the point of inflection in the top panel of Figure 2), the firm uses little of the variable inputs with the fixed inputs, and the law of diminishing returns is not operating. Thus, the curve faces downward or rises at a decreasing rate. Past point G' (i.e., for output levels greater than 1.5 units in the top panel of Figure 2), the law of diminishing returns operates, and the TVC curve faces upward or rises at an increasing rate. Since TC = TFC + TVC, the TC, curve has the same shape as the TVC curve but is $60 (the amount of the TFC) above it at each output level. These TVC and TC schedules are plotted in the top panel of Figure 2.

Figure 2



RELATIONSHIP BETWEEN AC AND MC CURVES

$d/dQ\ C(Q)/Q = \frac{[C'(Q)^* Q - C(Q).I]}{Q}$

$1/Q[C'(Q) - C(Q)/Q]$ for $Q > 0$

$d/dQ\ C(Q)/Q > 0$ when MC > AC

$d/dQ\ C(Q)/Q = 0$ when MC = AC

$d/dQ\ C(Q)/Q < 0$ when MC < AC

Lesson 19

COST ANALYSIS AND ESTIMATION (CONTINUED 1)

LONG-RUN COST CURVES

Long-Run Total Cost Curves

The long run is the time period during which all inputs are variable. Thus, all costs are variable in the long run (i.e., the firm faces no fixed costs). The length of time of the long run depends on the industry. In some service industries, such as dry cleaning, or a beauty salon or an outlet of ready-made garments for working women, the period of the long run may be only a few months. For others that are capital intensive, such as the construction of a new electricity-generating plant, it may be many years. It all depends on the length of time required for the firm to be able to vary all inputs.

The firm's long-run total cost (LTC) curve is derived from the firm's expansion path and shows the minimum long-run total costs of producing various levels of output. The firm's long-run average and marginal cost curves are then derived from the long-run total cost curve. These derivations are shown in Figure 1.

The top panel of Figure 1 shows the expansion path of the firm. The expansion path shows the optimal input combinations to produce various levels of output. For example, point A shows that in order to produce 1 unit of output (1Q), the firm uses 4 units of labor (4L) and 4 units of capital (4K). If the wage of labor (w) is $10 per unit and the rental price of capital (r) is also, $10 per unit, the minimum total cost of producing 1Q is:

(4L) ($10) + (4K) ($10) = $80

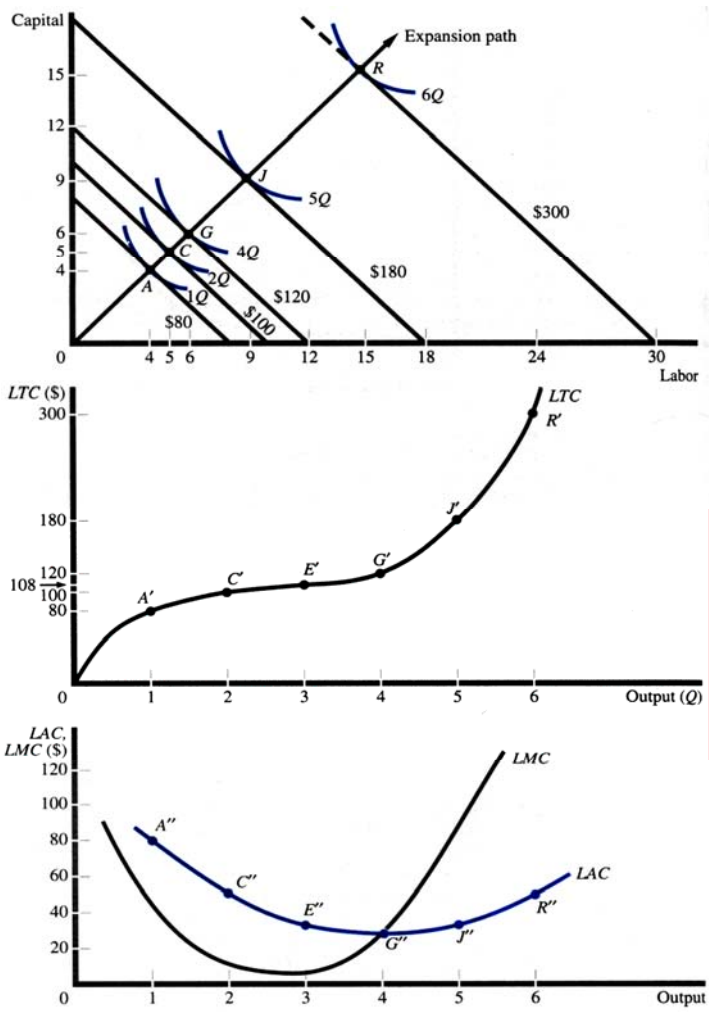
This is shown as point A' in the middle panel, where the vertical axis measures total costs and the horizontal axis measures output. From point C on the expansion path in the top panel, we get point C' ($100), on the LTC curve in the middle panel for 2Q. Other points on the LTC curve are similarly obtained. Note that the LTC curve starts at the origin because there are no fixed costs in the long run.

From the LTC curve we can derive the firm's long-run average cost (LAC) curve. LAC is equal to LTC divided by Q. That is,

$$\text{LAC} = \frac{\text{LTC}}{\text{Q}}$$

For example, the LAC to produce 1Q is obtained by dividing the LTC of $80 (point A' on the LTC curve in the middle panel of Figure 1) by 1. This is the slope of a ray from the origin to point A' on the LTC curve and is plotted as point A" in the bottom panel of Figure 1. Other points on the LAC curve are similarly obtained. Note that the slope of a ray from the origin to the LTC curve declines up to point G' (in the middle panel of Figure 1) and then rises. Thus, the LAC curve in the bottom panel declines up to point G" (4Q) and rises thereafter.

Figure 1



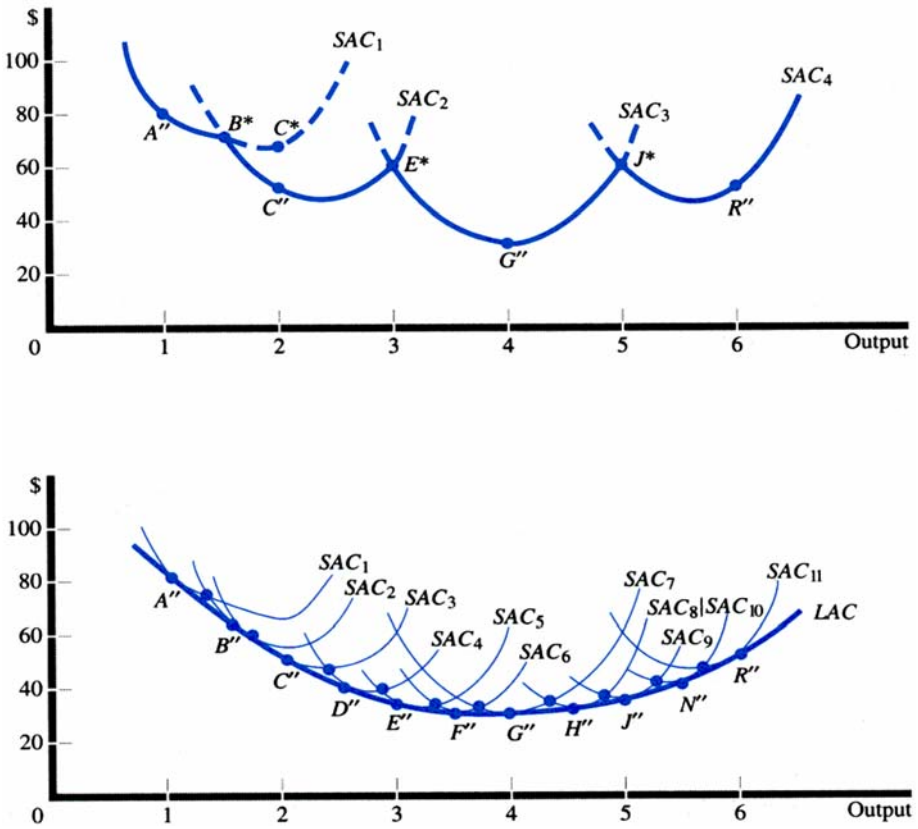
From the LTC curve we can also derive the long-run marginal cost (LMC) curve. This measures the change in LTC per unit change in output and is given by the slope of the LTC curve. That is,

$$\textbf{LMC} = \frac{\Delta \textbf{LTC}}{\Delta \textbf{Q}}$$

Long-run cost curves show the least-cost input combination for producing output assuming an ideal input selection. As in the case of short-run cost curves, wage rates, interest rates, plant configuration, and all other operating conditions are held constant. Any change in the operating environment leads to a shift in long-run cost curves. For example, product inventions and process improvements that occur over time cause a downward shift in long-run cost curves. Such changes must not be confused with movements along a given long-run cost curve caused by changes in the output level. Long-run cost curves reveal the nature of economies or diseconomies of scale and optimal plant sizes. They are a helpful guide to planning decisions.

It is important to keep in mind, however, that while the U shape of the short-run average cost (SAC) curve is based on the operation of the law of diminishing returns (resulting from the existence of fixed inputs in the short run), the U shape of the LAC curve depends on increasing, constant, and decreasing returns to scale, respectively.

Figure 2



LONG-RUN AVERAGE AND MARGINAL COST CURVES

The long-run average cost (LAC) curve shows the lowest average cost of producing each level of output when the firm can build the most appropriate plant to produce each level of output. This is shown in Figure 2. The top panel of Figure 2 is based on the assumption that the firm can build only four scales of plant (given by SAC_1, SAC_2, SAC_3, and SAC_4) while the bottom panel of Figure 2 is based on the assumption that the firm can build many more or an infinite number of cases of plant.

The top panel of Figure 2 shows that the minimum average cost of producing 1 unit of output (1 Q) is $80 and results when the firm operates the scale of plant given by SAC_1 (the smallest scale of plant possible) at point A". The firm can produce 1.5Q at an average cost of $70 by using either the scale of plant given by SAC_1 or the larger scale of plant given by SAC_2 at point B* (see the top panel of Figure 2). To produce 2Q, the firm will use scale of point SAC_2 at point C" ($50) rather than smaller scale of plant SAC_1 at point C* (the lowest point on SAC_1, which refers to the average cost of $67) Thus, the firm has more flexibility in the long run than in the short run. To produce 3Q, the firm is indifferent between using plant SAC_2 or larger plant SAC_3 at point E* ($60). The minimum average cost of producing 4Q ($30) is achieved when the firm operates plant SAC_3 at point G" (the lowest point on SAC3). To produce 5Q, the firm operates either plant SAC_3 or larger plant SAC_4 at point J* ($60). Finally, the minimum cost of producing 6Q is achieved when the firm operates plant SAC_4 (the largest plant) at point R" ($50).

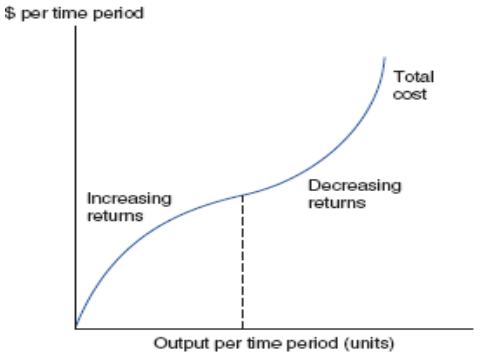
Thus, if the firm could build only the four scales of plant shown in the top panel of Figure 2, the long-run average cost curve of the firm would be A"B*C"E*G"J*R". If the firm could build many more scales of plant, the kinks at points B*, E*, and J* would become less prominent, as shown in the bottom panel of Figure 2. In the limit, as the number of scales of plants that the firm can

build in the long run increases, the LAC curve approaches the smooth curve indicated by the LAC curve in the bottom panels of Figures 1 and 2. Thus, the LAC curve is the tangent or "envelope" to the SAC curves and shows the minimum average cost of producing various levels of output in the long run, when the firm can build any scale of plant. Note that only at point G" (the lowest point on the LAC curve) does the firm utilize the optimal scale of plant at its lowest point. Production systems that reflect first increasing, then constant, then diminishing returns to scale result in U-shaped long-run average cost curves such as the one illustrated in Figure 2 bottom panel.

RELATIONSHIP BETWEEN PRODUCTION AND COST

Cost function is simply the production function expressed in monetary rather than physical units. We assume the firm is a 'price taker' in the input market. If input prices are not affected by the amount purchased, a direct relation exists between long run total cost and production functions. A production function exhibiting first increasing and then decreasing returns to scale is illustrated, along with its implied cubic cost function, in Figure 3. Here, costs increase less than proportionately with output over the range in which returns to scale are increasing but at more than a proportionate rate after decreasing returns set in. A direct relation between production and cost functions requires constant input prices. Similarly the relationship between AC, MC and AP, MP is shown in Figure 4.

Figure 3



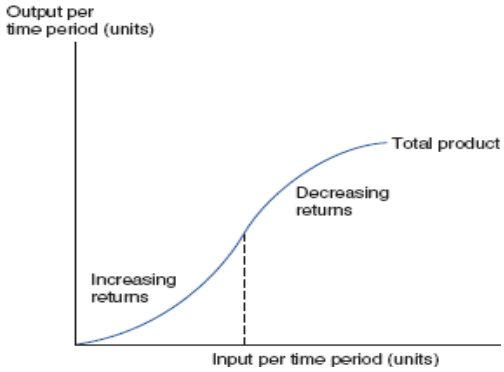
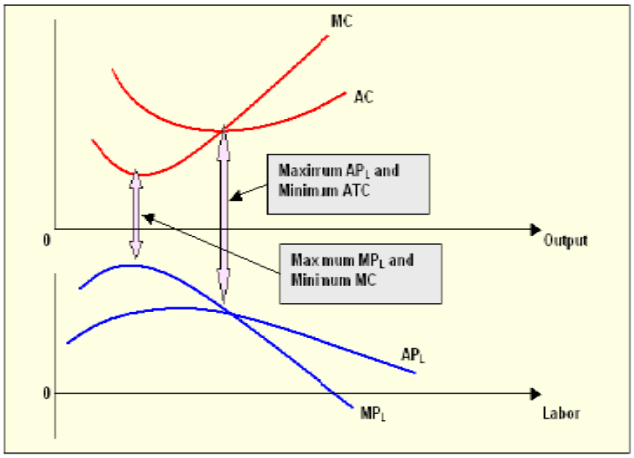


Figure 4



MINIMUM EFFICIENT SCALE

The number of competitors and ease of entry is typically greater in industries with U-shaped long-run average cost curves than in those with L-shaped or downward-sloping long-run average cost curves. Insight on the competitive implications of cost/output relations can be gained by considering the minimum efficient scale concept.

COMPETITIVE IMPLICATIONS OF MINIMUM EFFICIENT SCALE

Minimum efficient scale (MES) is the output level at which long-run average costs are minimized. MES is at the minimum point on a U-shaped long-run average cost curve (output Q = 4 in Figure 2 bottom panel) and at the corner of an L-shaped long-run average cost curve.

Generally speaking, competition is vigorous when MES is low relative to total industry demand. This fact follows from the correspondingly low barriers to entry from capital investment and skilled labor requirements. Competition can be less vigorous when MES is large relative to total industry output because barriers to entry tend to be correspondingly high and can limit the number of potential competitors.

ECONOMIES OF SCALE

Economies of scale exist when long-run average costs decline as output expands. Labor specialization often gives rise to economies of scale. In small firms, workers generally do several jobs, and proficiency sometimes suffers from a lack of specialization. Labor productivity can be higher in large firms, where individuals are hired to perform specific tasks. This can reduce unit costs for large-scale operations. Technical factors can also lead to economies of scale. Large-scale operation permits the use of highly specialized equipment, as opposed to the more versatile but less efficient machines used in smaller firms. Also, the productivity of equipment frequently increases with size much faster than its cost. A 500,000-kilowatt electricity generator costs considerably less than two 250,000-kilowatt generators, and it also requires less fuel and labor when operated at capacity. These economies extend to the cost of capital when large firms have easy access to capital markets and can acquire funds at lower rates.

At some output level, economies of scale are typically exhausted and average costs level out and begin to rise. Increasing average costs at high output levels are often attributed to limitations in the ability of management to coordinate large-scale organizations. Staff overhead also tends to grow more than proportionately with output, again raising unit costs. The current trend toward small to medium-sized businesses indicates that diseconomies limit firm sizes in many industries.

In the bottom panel of Figures 1 and 2, the LAC curve has been drawn as U-shaped. This is based on the assumption that economies of scale prevail at small levels of output and diseconomies of scale prevail at larger levels of output. As pointed out earlier, "economies of scale" refers to the situation in which output grows proportionately faster than inputs. For example, output more than doubles with a doubling of inputs. With input prices remaining constant, this leads to lower costs per unit. Thus, increasing returns of scale are reflected in a declining LAC Curve. On the other hand, decreasing returns to scale refers to the situation where output grows at a proportionately slower rate than the use of inputs. With input prices constant, this leads to higher costs per unit. Thus, decreasing returns to scale are reflected in an LAC curve that is rising. The lowest point on the LAC curve occurs at the output level at which the forces for increasing returns to scale are just balanced by the forces for decreasing returns to scale.

Increasing returns to scale or decreasing costs arise because of technological and financial reasons. At the technological level, economies of scale arise because as the scale of operation increases, a greater division of labor and specialization can take place and more specialized and productive machinery can be used. Specifically, with a large scale operation, each worker can be assigned to perform a repetitive task rather than numerous different ones. This results in increased proficiency and the avoidance of the time lost moving from one machine to another.
Besides the technological reasons for increasing returns to scale or decreasing costs, there are financial reasons that arise as the size of the firm increases. Because of bulk purchases, larger

firms are more likely to receive quantity discounts in purchasing raw materials and other intermediate (i.e., semi-processed) inputs than smaller firms. Large firms can usually sell bonds and stocks more favorably and receive bank loans at lower interest rates than smaller firms. Large firms can also achieve economies of scale or decreasing costs in advertising and other promotional efforts. For all these technological and financial reasons, the LAC curve of a firm is likely to decline as the firm expands and becomes larger.

Decreasing returns to scale, on the other hand, arise primarily because as the scale of operation increases, it becomes ever more difficult to manage the firm effectively and coordinate the various operations and divisions of the firm. The number of meetings the paper work and telephone bills increases more than proportionately to the increase in the scale of operation, and it becomes increasingly difficult for top management to ensure that their directives and guidelines are properly carried out by their subordinates. Thus, efficiency decreases and costs per unit tend to rise.

COST ELASTICITIES AND ECONOMIES OF SCALE

It is often easy to calculate scale economies by considering cost elasticities. **Cost elasticity**, ε_C, measures the percentage change in total cost associated with a 1 percent change in output. Algebraically, the elasticity of cost with respect to output is

$$\varepsilon_C = \frac{\text{Percentage Change in Total Cost (TC)}}{\text{Percentage Change in Output (Q)}}$$

$$\varepsilon_C = \partial TC/TC \div \partial Q/Q$$

$$\varepsilon_C = \partial TC/\partial Q * Q/TC$$

Cost elasticity is related to economies of scale as follows:

If	Then	Which Implies
Percentage change in TC < Percentage change in Q	$\varepsilon_C < 1$	Economies of scale (decreasing AC)
Percentage change in TC = Percentage change in Q	$\varepsilon_C = 1$	No Economies of scale (Constant AC)
Percentage change in TC > Percentage change in Q	$\varepsilon_C > 1$	Economies of scale (Increasing AC) i-e Diseconomies of scale are implied

With a cost elasticity of less than one ($\varepsilon_C < 1$), costs increase at a slower rate than output. Given constant input prices, this implies higher output-to-input ratios and economies of scale. If $\varepsilon_C = 1$, output and costs increase proportionately, implying no economies of scale. Finally, if $\varepsilon_C > 1$, for any increase in output, costs increase by a greater relative amount, implying decreasing returns to scale. To prevent confusion concerning cost elasticity and returns to scale, remember that an inverse relation holds between average costs and scale economies but that a direct relation holds between resource usage and returns to scale. Thus, although $\varepsilon_C < 1$ implies falling *AC* and economies of scale, because costs are increasing more slowly than output, recall from Chapter 7 that an output elasticity greater than 1 ($\varepsilon Q > 1$) implies increasing returns to scale,

because output is increasing faster than input usage. Similarly, diseconomies of scale are implied by $\varepsilon_C > 1$, diminishing returns are indicated when $\varepsilon Q < 1$.

FIRM SIZE AND PLANT SIZE

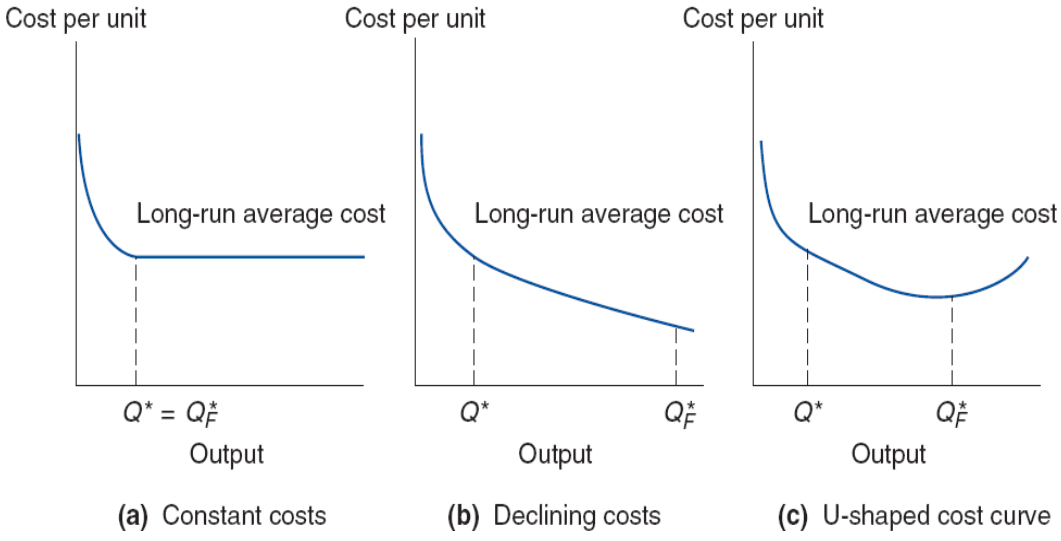
The cost function for a multi plant firm can be the sum of the cost functions for individual plants. It can also be greater or less than this figure. For this reason, it is important to examine the relative importance of economies of scale that arise within production facilities, intra plant economies, and those that arise between and among plants, or multi plant economies of scale.

MULTI-PLANT ECONOMIES AND DISECONOMIES OF SCALE

Multi-plant economies of scale are cost advantages that arise from operating multiple facilities in the same line of business or industry. **Multi-plant diseconomies of scale** are cost disadvantages that arise from managing multiple facilities in the same line of business or industry. To illustrate, assume a U-shaped long-run average cost curve for a given plant, as shown in Figure 2, bottom panel. If demand is sufficiently large, the firm will employ n plants, each of optimal size and producing Q* units of output.

In this case, what is the shape of the firm's long-run average cost curve? Figure 5 shows three possibilities. Each possible long-run average cost curve has important implications for the minimum efficient firm size, Q*F. First, the long-run average cost curve can be L-shaped, as in Figure 5(a), if no economies or diseconomies result from combining plants. Second, costs could decline throughout the entire range of output, as in Figure 5(b), if multi-plant firms are more efficient than single-plant firms. When they exist, such cases are caused by economies of multi-plant operation. For example, all plants may use a central billing service, a common purchasing or distribution network, centralized management, and so on. The third possibility, shown in Figure 5(c), is that costs first decline beyond Q*, the output of the most efficient plant, and then begin to rise. In this case, multi-plant economies of scale dominate initially, but they are later weighed down by the higher costs of coordinating many operating units.

Figure 5



In the real world, the forces for increasing and decreasing returns to scale often operate side by side, with the former prevailing ,at small levels of output (so that the LAC curve declines) and

the latter tending to prevail at much larger levels of output (so that the LAC curve rises). The lowest point on the LAC curve occurs when the forces for, increasing and decreasing returns to scale just balance each other. In the real world, however, the LAC curve is often found to have a nearly flat bottom and to be L-shaped rather than U-shaped. This implies that economies of scale are rather quickly exhausted and constant or near constant returns to scales prevail over a considerable range of outputs in many industries. In these industries, small firms coexist with much larger firms.

There are some industries, however, in which the LAC curve declines continuously as the firm expands output, to the point where a single firm could satisfy the total market for the product or service more efficiently than two or more firms. These cases are usually referred to as "natural monopolies" and often arise in case of such utilities as electricity and public transportation.

Economies of Multi plant Operation: an Example

Consider an Electronics Company XYZ that manufactures industrial control panels. Presently it is producing at a single plant but a multi-plant alternative is being considered. Estimated demand, marginal revenue, and single-plant production plus transportation cost curves for the firm are:

$P = 940 – 0.02Q$
$TR = P*Q = (940 – 0.02Q)\ Q = 940Q – 0.02Q^2$
$MR = \partial TR/\partial Q = 940 – 0.04Q$
$TC = 250{,}000 + 40Q + 0.01Q^2$
$MC = \partial TC/\partial Q = 40 + 0.02Q$

The firm total profit function is:

$\pi = TR – TC = (940 – 0.02Q)\ Q – 250{,}000 - 40Q - 0.01Q^2$
$\pi = - 0.03Q^2 + 900Q – 250{,}000$

Setting marginal revenue = marginal cost and solving for the related output quantity gives:

$MR = MC$
$940 – 0.04Q = 40 + 0.02Q$
$0.06Q = 900$
$Q = 15{,}000$
At $Q = 15{,}000$ $P = 940 – 0.02(15{,}000)$
$= \$640$ &
$\pi = \$6{,}500{,}000$ or $6.5 million

Profits are maximized at Q = 15,000 output level under the assumption of single-plant. In order to get insight regarding the possible advantages of multi-plant operation, the AC function for a single-plant must be examined. To simply matters, we assume that multi-plant production is possible under the same cost conditions. Also assume that there are no other multi-plant economies or diseconomies of scale. The activity level at which AC is minimized is found by setting MC = AC and solving for Q:

$AC = TC/Q = 250{,}000 + 40Q + 0.01Q^2 /Q$
$AC = 250{,}000/Q + 40 + 0.01Q$

$MC = AC$
$40 + 0.02Q = 250{,}000/Q + 40 + 0.01Q$
$Q^2 = 250{,}000/0.01$

So Q = √ 250,000 = 5,000
At Q = 5,000 AC-minimizing activity level might suggest multi-plant production at 3 facilities (plants) will be optimum as:
Optimal No of Plants = Optimal Multi Plant Activity level
Optimal Production per Plant
= 15,000/5000 = 3

But previously MR = MC = $640, however, with multi-plant production, MC will be lowered:
MC = 40 + 0.02Q = 40 + 0.02(5,000) = $140
MR = MC
940 – 0.04Q = 140
Q = 20,000

Optimal No of Plants = Optimal Multi Plant Activity level
Optimal Production per Plant
= 20,000/5,000 = 4 Plants

At Q = 20,000 P = 940 – 0.02(20,000)
= $540
And π = TR – TC = P*Q – 4* TC per plant
= 540(20,000) – 4[250,000 + 40(5000) + 0.01(5,000) 2]
= $8,000,000 or $8 million

ECONOMIES OF SCOPE
Cost analysis focuses not just on how much to produce but also on what combination of products to offer. By virtue of their efficiency in the production of a given product, firms often enjoy cost advantages in the production of related products.

ECONOMIES OF SCOPE CONCEPT
Economies of scope exist when the cost of joint production is less than the cost of producing multiple outputs separately. A firm will produce products that are complementary in the sense that producing them together costs less than producing them individually. In fact, the economies of scope concept explain why firms typically produce multiple products. Companies establish a working relation with an ideal group of prospective customers for stocks, bonds, and other investments. When viewed as a delivery vehicle or marketing device, money market mutual funds may be one of the industry's most profitable financial product lines.

EXPLOITING SCOPE ECONOMIES
Economies of scope are important because they permit a firm to translate superior skill in a given product line into unique advantages in the production of complementary products. Effective competitive strategy often emphasizes the development or extension of product lines related to a firm's current stars, or areas of recognized strength. For example, PepsiCo, Inc., has long been a leader in the soft drink market. Over time, the company has gradually broadened its product line to include various brands of regular and diet soft drinks, Cookies, and other snack foods. PepsiCo can no longer be considered just a soft drink manufacturer. It is a widely diversified beverages and snack Foods Company for whom well over one-half of total current profits come from non–soft drink lines. The economies of scope concept offer a useful means for evaluating the potential of current and prospective lines of business. It naturally leads to definition of those areas in which the firm has a comparative advantage and its greatest profit potential.

Economies of scale have to be distinguished from economies of scope. The latter refer to, the lowering of costs that a firm often experience when it produces two or more products together rather than each alone. A smaller commuter airline, for example, can profitably extend into providing cargo services, thereby lowering the cost of each operation alone. Another example is provided by a firm that produces a second product in order to use the by products (which before the firm had to dispose of at a cost) arising from the production of the first product. Management must be alert to the possibility of profitably extending its product line to exploit such economies of scope.

Lesson 20

COST ANALYSIS AND ESTIMATION (CONTINUED 2)

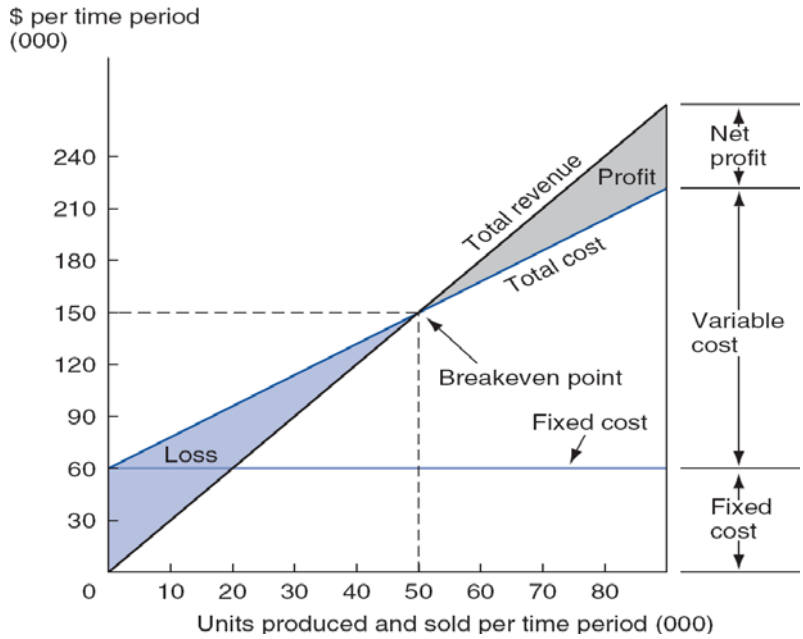
COST-VOLUME-PROFIT ANALYSIS

Cost-volume-profit analysis, sometimes called breakeven analysis, is an important analytical technique used to study relations among costs, revenues, and profits. Both graphic and algebraic methods are employed. For simple problems, simple graphic methods work best. In more complex situations, analytic methods, possibly involving spreadsheet software programs, are preferable.

Cost-Volume-Profit Charts

A basic cost-volume-profit chart composed of a firm's total cost and total revenue curves is depicted in Figure1. Volume of output is measured on the horizontal axis; revenue and cost are shown on the vertical axis. Fixed costs are constant regardless of the output produced and are indicated by a horizontal line. Variable costs at each output level are measured by the distance between the total cost curve and the constant fixed costs. The total revenue curve indicates the price/demand relation for the firm's product; profits or losses at each output are shown by the distance between total revenue and total cost curves.

Figure 1

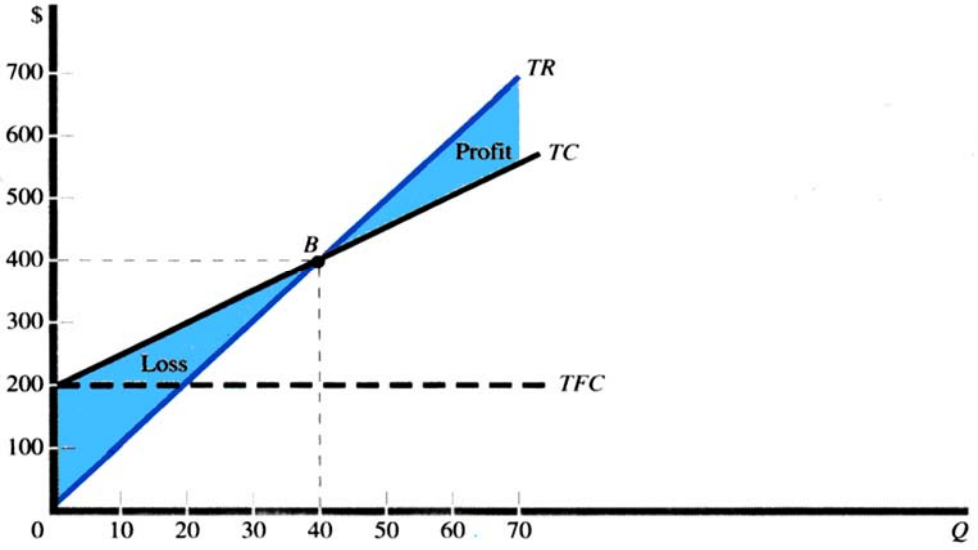


In the example depicted in Figure 1, fixed costs of $60,000 are represented by a horizontal line. Variable costs for labor and materials are $1.80 per unit, so total costs rise by that amount for each additional unit of output. Total revenue based on a price of $3 per unit is a straight line through the origin. The slope of the total revenue line is steeper than that of the total cost line. Below the breakeven point, found at the intersection of the total revenue and total cost lines, the firm suffers losses. Beyond that point, it begins to make profits. Figure 1 indicates a breakeven point at a sales and cost level of $150,000, which occurs at a production level of 50,000 units.

Cost-volume-profit or breakeven analysis examines the relationship among the total revenue, total costs, and total profits of the firm at various levels of output. Cost-volume-profit or breakeven analysis is often used by business executives to determine the sales volume

required for the firm to break even and the total profits and losses at other sales levels. The analysis uses a cost-volume-profit chart in which the total revenue (TR) and the total cost (TC) curves are represented by straight lines, as in Figure 2.

Figure 2



The cost-volume-profit or breakeven chart is a flexible tool to quickly analyze the effect of changing conditions on the firm. For example, an increase in the price of the commodity can be shown by increasing the slope of the TR curve, an increase in total fixed costs of the firm can be shown by an increase in the vertical intercept of the TC curve, and an increase in average variable costs by an increase in the slope of the TC curve. The chart will then show the change in the breakeven point of the firm and the profits or losses at other output or sales levels. Although cost-volume-profit charts can be used to portray profit/output relations, algebraic techniques are typically more efficient for analyzing decision problems. The algebra of cost volume- profit analysis can be illustrated as follows. Let

P = Price per unit sold

Q = Quantity produced and sold

TFC = Total fixed costs

AVC = Average variable cost

On a per-unit basis, profit contribution equals price minus average variable cost ($\pi C = P - AVC$). Profit contribution can be applied to cover fixed costs and then to provide profits. It is the foundation of cost-volume-profit analysis. One useful application of cost-volume-profit analysis lies in the determination of breakeven activity levels. A **breakeven quantity** is a zero profit activity level. At breakeven quantity levels, total revenue (P * Q) exactly equals total costs (TFC + AVC * Q). Total revenue is equal to the selling price (P) per unit times the quantity of output or sales (Q). That is,

TR = (P) (Q) (1)

Total costs equal total fixed costs plus total variable costs (TVC). Since TVC is equal to the average (per-unit) variable costs (AVC) times the quantity of output or sales, we have

TC = TFC + (AVC) (Q) (2)

Setting total revenue equal to total costs and substituting Q_B (the breakeven output) for Q, we have

TR = TC (3)

$(P)(Q_B) = TFC + (AVC)(Q_B)$ (4)

Solving Equation 4 for the breakeven output, Q_B, we get

$(P)(Q_B) - (AVC)(Q_B) = TFC$ (5)

$(Q_B)(P - AVC) = TFC$ (6)

(7) $$Q_B = \frac{TFC}{P - AVC}$$

For example, with TFC = $200, P = $10, and AVC= $5,

$$QB = \frac{\$200}{\$10 - \$5} = 40$$

This is the breakeven output shown on the cost-volume-profit chart in Figure 2. The denominator in Equation 7 (that is, P - AVC) is called the contribution margin per unit because it represents the portion of the selling price that can be applied to cover the fixed costs of the firm and to provide for profits.

More generally, suppose that the firm wishes to earn a specific profit and wants to estimate the quantity that it must sell to earn that profit. Cost-volume-profit or breakeven analysis can be used in determining the target output (Q_T) at which a target profit ($л_T$) can be achieved. To do so, we simply add $л_T$ to the numerator of Equation 7 and have:

(8) $$Q_T = \frac{TFC + л_T}{P\text{-}AVC}$$

For example, if the firm represented in the cost-volume-profit chart in Figure 2 wanted to earn a target profit of $100, the target output would be:

$$Q = \frac{\$200+\$100}{\$10-\$5} = \frac{\$300}{\$5} = 60$$

To see that the output of Q = 60 does indeed lead to the target profit ($л_T$) of $100, note that

TR = (P) (Q) = ($10) (60) = $600
TC = TFC + (AVC) (Q) = $200 + ($5) (60) = $500
And
$л_T$ = TR - TC = $600 - $500= $100

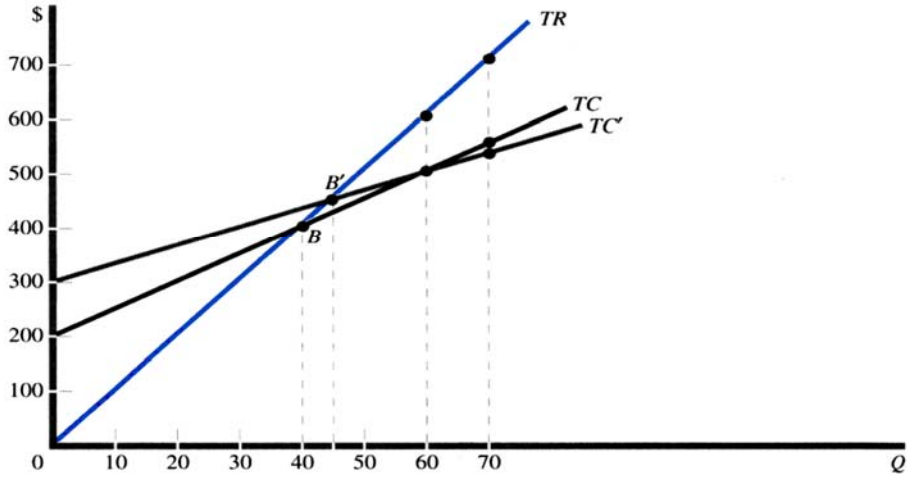
While linear cost-volume-profit charts and analysts are frequently used by business executives, government agencies, and not-for-profit organizations, care must be exercised to apply them only when the assumption of constant prices and average variable costs holds. Cost-volume-profit analysis also assumes that the firm produces a single product or a constant mix of products. Over time, the product mix changes, and it may be difficult to allocate the fixed costs among the various products. Despite these shortcomings, cost-volume-profit analysis can be very useful in managerial decision making.

DEGREE OF OPERATING LEVERAGE

Operating leverage refers to the ratio of the firm's total fixed costs to total variable costs. The higher is this ratio, the more leveraged the firm is said to be. As the firm becomes more automated or more leveraged (i.e., substitutes fixed for variable costs), its total fixed costs rise but its average variable costs fall. Because of higher overhead costs, the breakeven output of the firm increases. This is shown in Figure 3. Cost-volume-profit analysis is also a useful tool for analyzing the financial characteristics of alternative production systems. This analysis focuses on how total costs and profits vary with operating leverage or the extent to which fixed production facilities versus variable production facilities are employed. The absolute slope of TC' > TC that is 5 > 3.33.

Figure 3

TC' has a higher DOL than TC
And therefore a higher Q_{BE}



The degree of operating leverage is the percentage change in profit from a 1 unit change in units sold:

(9) $$DOL = \frac{\%\Delta л}{\%\Delta Q} = \frac{\Delta л / л}{\Delta Q / Q} = \frac{\Delta л}{\Delta Q} \cdot \frac{Q}{л}$$

But л = Q(P - AVC) - TFC and Δл = ΔQ(P - AVC). Substituting these values into Equation (9), we get

(10) $$DOL = \frac{\Delta Q (P - A VC) Q}{\Delta Q[Q(P - A VC) - TFC]} = \frac{Q(P - A VC)}{Q(P - A VC) - TFC}$$

The numerator in Equation (10) is the total contribution to fixed costs, and profits of all units sold by the firm, and the denominator is total (economic) profit.

In terms of calculus:

$$DOL = \frac{\delta л}{\delta Q} \cdot \frac{Q}{л}$$

For example, for an increase in output from 60 to 70 units, the degree of operating leverage with TC is:

$$DOL = \frac{60(\$10 - \$5)}{60(\$10 - \$5) - \$200} = \frac{\$300}{\$100} = 3$$

With TC' (i.e., when the firm becomes more leveraged), the degree of operating leverage becomes

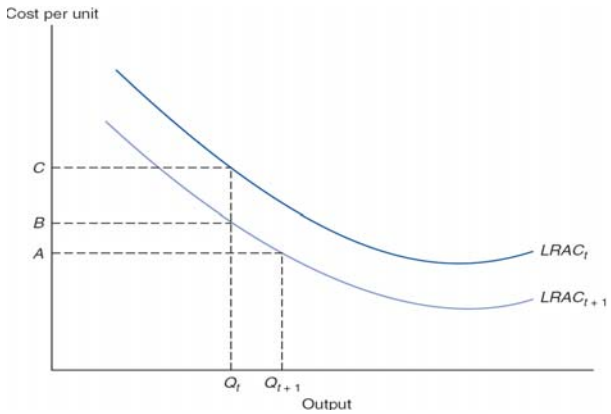
DOL' = 60($10 - $3.33) = $400 = 4
60($10 $3.33)-$300 $100

Thus, the degree of operating leverage (DOL) increases as the firm becomes more leveraged or capital intensive. It is also higher the closer we are to the breakeven point because the base in measuring the percentage change in profits (the denominator in Equation 9) is close to zero near the breakeven point. The denominator becomes smaller and smaller so that DOL is higher. Note that when the firm's sales and output, are high (greater than 60 units in Figure 3), the firm makes larger profits when it is more leveraged (i.e., with TC'). But it also incurs losses sooner and these losses rise more rapidly than when the firm is less highly leveraged (i.e., with TC). The larger profits of the more highly leveraged firm when output is high (greater than 60 units in Figure 3) cant thus be regarded as the return for its greater risk.

LEARNING CURVES
For many manufacturing processes, average costs decline substantially as *cumulative* total output increases. Improvements in the use of production equipment and procedures are important in this process, as are reduced waste from defects and decreased labor requirements as workers become more proficient in their jobs.

LEARNING CURVE CONCEPT
When knowledge gained from manufacturing experience is used to improve production methods, the resulting decline in average costs is said to reflect the effects of the firm's learning curve. The learning curve or experience curve phenomenon affects average costs in a way similar to that for any technical advance that improves productive efficiency. Both involve a downward shift in the long-run average cost curve at all levels of output. Learning through production experience permits the firm to produce output more efficiently at each and every output level. To illustrate, consider Figure 4, which shows hypothetical long-run average cost curves for period's t and t + 1. With increased knowledge about production methods gained through the experience of producing Qt units in period t, long-run average costs have declined for every output level in period t + 1, which means that Qt units could be produced during period t + 1 at an average cost of B rather than the earlier cost of C. The learning curve cost savings is BC. If output were expanded from Qt to Qt+1 between these periods, average costs would fall from C to A. This decline in average costs reflects both the learning curve effect, BC, and the effect of economies of scale, AB.

Figure 4



To isolate the effect of learning or experience on average cost, it is necessary to identify carefully that portion of average-cost changes over time that is due to other factors. One of the most important of these changes is the effect of economies of scale. As seen before, the change in average costs experienced between periods t and t + 1 can reflect the effects of both learning and economies of scale. Similarly, the effects of important technical breakthroughs, causing a downward shift in LRAC curves, and input-cost inflation, causing an upward shift in LRAC curves, must be constrained to examine learning curve characteristics. Only when output scale, technology, and input prices are all held constant can the learning curve phenomenon be identified.

Because the **learning curve** concept is often improperly described as a cause of economies of scale, it is worth repeating that the two are distinct concepts. **Scale economies** relate to cost differences associated with different output levels along a single LRAC curve. Learning curves relate cost differences to total cumulative output. They are measured by shifts in LRAC curves over time. These shifts result from improved production efficiencies stemming from knowledge gained through production experience. Care must be exercised to separate learning and scale effects in cost analysis.

Research in a number of industries, ranging from aircraft manufacturing to semiconductor memory-chip production, has shown that learning or experience can be very important in some production systems. Learning or experience rates of 20 percent to 30 percent are sometimes reported. These high learning rates imply rapidly declining manufacturing costs as cumulative total output increases. It should be noted, however, that many learning curve studies fail to account adequately for the expansion of production. Therefore, reported learning or experience rates sometimes include the effects of both learning and economies of scale. Nevertheless, managers in a wide variety of industries have found that the learning curve concept has considerable strategic implications.

STRATEGIC IMPLICATIONS OF THE LEARNING CURVE CONCEPT

The rule used for representing the learning curve effect states that the more times a task has been performed, the less time will be required on each subsequent iteration. This relationship was probably first quantified in 1936 at Wright-Patterson Air Force Base in the United States, where it was determined that every time total aircraft production doubled, the required labor time decreased by 10 to 15 percent. Subsequent empirical studies from other industries have yielded different values ranging from only a couple of percent up to 30 percent, but in most cases it is a

constant percentage: It did not vary at different scales of operation. Learning curve theory states that as the quantity of items produced doubles, costs decrease at a predictable rate. Suppose, for example, the AC/unit for a new product were $100 during 2004 but fell to $90 during 2005. The Learning Rate is given as:

Learning Rate = [1 – AC2/AC1] * 100

= [1 – 90/100] * 100

= 10%

As cumulative total output doubles, average cost is expected to fall by 10 percent. A classic example illustrating the successful use of the learning curve concept is Dallas– based Texas Instruments (TI). TI's main business is producing semiconductor chips, which are key components used to store information in computers and a wide array of electronic products. With growing applications for computers and "intelligent" electronics, the demand for semiconductors is expanding rapidly. Some years ago, TI was one of a number of leading semiconductor manufacturers. At this early stage in the development of the industry, TI made the decision to price its semiconductors well below then-current production costs, given expected learning curve advantages in the 20 percent range. TI's learning curve strategy proved spectacularly successful. With low prices, volume increased dramatically. Because TI was making so many chips, average costs were even lower than anticipated; it could price below the competition; and dozens of competitors were knocked out of the world market. Given a relative cost advantage and strict quality controls, TI rapidly achieved a position of dominant leadership in a market that became a source of large and rapidly growing profits.

The Japanese have been frequently cited in academic studies and in the popular press for their use of the learning curve in driving down costs. This is more dramatically shown in their production of computer chips and consumer electronics.

As firms gain experience in the production of commodity or service, their average cost of production usually declines. That is, for a given level of output per time period, the increasing cumulative total output over many time periods often provides the manufacturing experience that enables firms to lower their average cost of production. The learning curve shows the decline in the average input cost of production with rising cumulative total outputs over time. For example, it might take 1,000 hours to assemble the 100th aircraft, but only 700 hours to assemble the 200th, aircraft because managers and workers become more efficient as they gain production experience. Contrast this to economies of scale, which refer instead to declining average cost as the firm's output per time period increases.

Figure 5

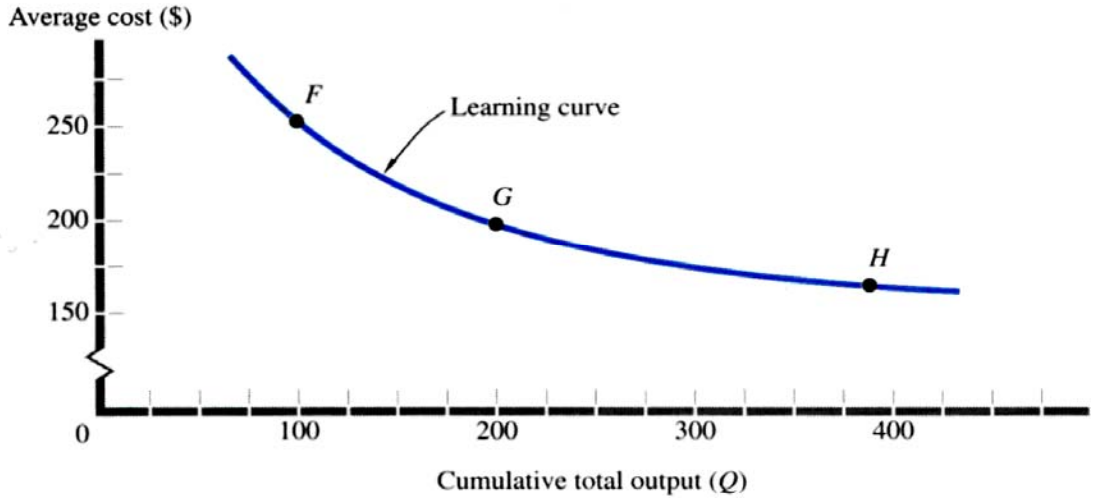


Figure 5 shows a learning curve, which indicates that the average cost decline from about $250 for producing the 100th unit of the product (point F), to about $200 for producing the 200th unit (point G), and to about $165 for the 400th unit (point H). Note that average cost declines at a decreasing rate so that the learning curve is convex to the origin. This is the usual shape of learning curves that is, firms usually achieve the largest decline in average input costs when the production process is relatively new and less as the firm matures.

The learning curve can be expressed algebraically as follows:

$$C = aQ^b \qquad (1)$$

Where C is the average input cost of the Qth unit of output, a is the average cost of the first unit of output, and b will be negative because the average input cost declines with increases in cumulative total output. The greater the absolute value of b, the faster average input cost declines. Taking the logarithm of both sides of Equation 1 give:

$$\log C = \log a + b \log Q \qquad (2)$$

In the above logarithmic form, b is the slope of the learning curve.

The parameter of the learning curve in the double-log form of Equation 2 (i.e., log a and b) can be estimated by regression analysis with historical data on average cost and cumulative output. Suppose that doing this gives the following result:

$$\log C = 3 - 0.3 \log Q \qquad (3)$$

In Equation 3, C is expressed in dollars, log a = 3 and b = -0.3. Thus, the average input cost of the 100th unit is:

$$\log C = 3 - 0.3 \log 100$$

Since the log of 100 is 2, we have

$$\log C = 3 - 0.3(2)$$
$$= 3 - 0.6$$

= 2.4.
Since the antilog of 2.4 is 251.19, the average input cost (C) of the 100th unit of output is $251.19. The average input cost for the 200th unit is

log C = 3 – 0.3 log 200
= 3 - 0.3 (2.30103)
= 3 - 0.690309
= 2.309691

Therefore, C = $204.03. While for the 400th unit, C= $165.72. These are the values shown by the learning curve in Figure 5.

Learning curves have been documented in many manufacturing and service sectors, ranging from the manufacturing of airplanes, appliances; shipbuilding, refined petroleum products, to the operation of power plants. They have also been used to forecast the needs for personnel, machinery, and raw materials, and for scheduling production, determining the price at which to sell output, and even for evaluating suppliers price quotations. For example, in its early days as a computer-chip producer, Texas Instruments adopted an aggressive price strategy based on the learning curve. Believing that the learning curve in chip production was steep, it kept unit prices low in order to increase its cumulative total output rapidly and thereby benefit from learning by doing. The strategy was successful, and Texas Instruments became one of the world's major players in this market.

How rapidly the learning curve (i.e., average input costs) declines can differ widely among firms and is greater the smaller the rate of employee turnover, the fewer the production interruptions (which would lead to "forgetting"), and the greater the ability of the firm to transfer knowledge from the production of other similar products. The average cost typically declines by 20, to 30 percent for each doubling of cumulative output for many firms. Firms however do not rely only on their production experience to lower costs and are looking farther and farther away from their industry to gain insights on how to increase productivity. Finally, the beneficial effects of learning are realized only when management systems tightly control costs and monitor potential sources of increased efficiency. Continuous feedback of information between production and management personnel is essential.

Lesson 21

COST ANALYSIS AND ESTIMATION (CONTINUED 3)

EMPIRICAL ESTIMATION OF COST FUNCTIONS

The study of cost curves has its origin with Joel Dean, who wrote the first textbook on managerial economics, and conducted many of the studies dating back to the 1930s. As in the case of production functions, we are interested in estimating cost functions both in the short run and in the long run. The short-run function helps define short-run marginal costs and thus assist the manager in determining output and prices. In the long run, the decision that a firm faces involves building the most efficient size of plant. That determination will depend on the existence of scale economies and diseconomies.

In investigating short-run cost functions, using regression analysis, the researchers have most frequently employed the time series technique with data for a specific plant or firm over time. The size of the plant or firm, as well as technology, should not change significantly during the time interval used in short-run cost function.

Long-run cost function---the planning functions ---allow for changes in all factors including plant size. Most studies of long-run cost functions have employed cross-sectional analysis.
Empirical estimates of cost functions are essential for many managerial decision purposes. Knowledge of short-run cost functions is necessary for the firm in determining the optimal level of output and the price to charge. Knowledge of long-run cost functions is essential in planning for the optimal scale of plant for the firm to build in the long run. We will examine the most important techniques for estimating the firm's short-run and long-run cost curves, and consider some of the data and measurement problems encountered in estimation, and summarize the results of some empirical studies of short-run and long-run cost functions.

THE ESTIMATION OF SHORT-RUN COST FUNCTIONS

Data and Measurement Problem

The most common method of estimation the firm's short-run cost functions is regression analysis, whereby total variable costs are regressed against output and a few other variables, such as input prices and operating conditions, during the time period when the size of the plant is fixed. The total variable cost rather than total cost function is estimated because of the difficulty of allocating fixed cost to the various products produced by the firms. The firm's total cost function can then be obtained by simply adding the best estimate possible of the fixed costs to the total variable costs. The firm's average variable and marginal cost functions can be easily obtained from total variable cost function.

EMPIRICAL ESTIMATION DATA COLLECTION ISSUES

- Opportunity Costs Must Be Extracted from Accounting Cost Data
- Costs Must Be Divided Among Products
- Costs Must Be Matched to Output Over Time
- Costs Must Be Corrected for Inflation

The firm's cost functions are based on the assumption of constant input prices. If input prices increase, they will cause an upward shift of the entire cost function. Therefore, input prices will have to be included as additional explanatory variables in the regression analysis in order to identify their independent effect on costs. Other independent variables that may have, to be included in the regression analysis are fuel and material costs, the quality of inputs, the technology used by the firm, weather conditions, and changes in the product mix and product quality. The actual independent or explanatory variables included in the regression (besides

output) depend on the particular situation under examination.
Thus, we can postulate that

(1) $C = f(Q, X_1, X_2, \ldots, X_n)$

Where C refers to total variable costs, Q is output, and the X's refer to the other determinants of the firm's costs. Using multiple regression analysis allows us to isolate the effect on costs of changes in each of the independent or explanatory variables. By concentrating on the relationship between costs and output, we can then identify the firm's total variable cost curve.

Economic versus Accounting Costs: Most empirical studies of cost functions have used Accounting data that record the actual costs and expenses on a historical basis. However decision-making data---economic data should also include opportunity costs. The problem that arise in the estimation of short-run cost functions because of differences between accounting and economic costs are the most difficult to solve.

One fundamental problem that arises in the empirical estimation of cost functions is that opportunity costs must be extracted from the available accounting cost data. That is, each input used in production must be valued at its opportunity cost based on what the input could earn in its best alternative use rather than the actual expenditures for the input. For example, if the firm owns the building in which it operates, the cost of using the building is not zero but is equal to the rent that the firm would obtain by renting the building to the highest bidder. Similarly inventories used in current production must be valued at current market prices rather than at historical cost. Finally, the part of the depreciation of fixed assets, such as machinery, that is based on the, actual usage of the assets (as contrasted to the depreciation of the assets based on the passage of time alone) should be estimated and included in current production costs for each product. These data are often difficult to obtain from the available accounting data.

Not only must costs be correctly apportioned to the various products produced by the firm but care must also be exercised to match costs to output over time (i.e., allocate costs to the period in which the output is produced rather than to the period when the costs were incurred) Specifically, the leads and lags in costs from the corresponding output must be adjusted so as to achieve a correct correspondence between costs and output. For example, while a firm may postpone all but emergency maintenance until a period of slack production, these maintenance costs must be allocated to the earlier production periods.

The manager must also determine the length of time over which to, estimate cost functions while daily, weekly, monthly, quarterly, or yearly data can be used monthly data over a period of two or three years are usually used. The period of time must be long-enough to allow for sufficient variation in output and costs but not long enough for the firm to change plant size (since the firm would then no longer be operating in the short run). Since output is usually measured in physical units (e.g., number of automobiles of a particular type produced per time period) while costs are measured in monetary units, the various costs must be deflated by the, appropriate price index to correct for inflation. That is, with input prices usually rising at different rates, the price index for each category of inputs will have to be used to obtain their deflated values to use in the regression analysis.

THE FUNCTIONAL FORM OF SHORT-RUN COST FUNCTIONS

Economic theory postulates an S-shaped (cubic) TVC curve as indicated in the left panel of Figure 1, with corresponding U-shaped AVC and MC curves. The general equations for these functions are, respectively,

Cubic TVC Function

(2) $TVC = a(Q) + bQ^2 + cQ^3$

(3) $AVC = \frac{TVC}{Q} = a + Q + cQ^2$

(4) $MC = a + 2bQ + 3cQ^2$

Where conditions for U-shape cost functions are: a > 0, b < 0 and c > 0

The right panel of Figure 1 shows a linear approximation to the cubic TVC curve, which often gives a good empirical fit of the data points over the observed range of outputs. The estimated equations of the linear approximation to the S-shaped or cubic TVC curve and of its corresponding AVC and MC curves are

Linear TVC Function

(5) $TVC = a + bQ$

(6) $AVC = a/Q + b$

(7) $MC = b$

Quadratic TC Function

$TC = a + bQ + cQ^2$ (8)
$AC = a/Q + b + cQ$ (9)
$MC = b + 2cQ$ (10)
Where $a > 0$, $c > 0$ and $b < 0$

SHORT-RUN COST FUNCTION

Alternative specifications of the Total Cost function (relating total cost and output)

- **Cubic relationship**

As output increases, total cost first increases at a decreasing rate, then increases at an increasing rate

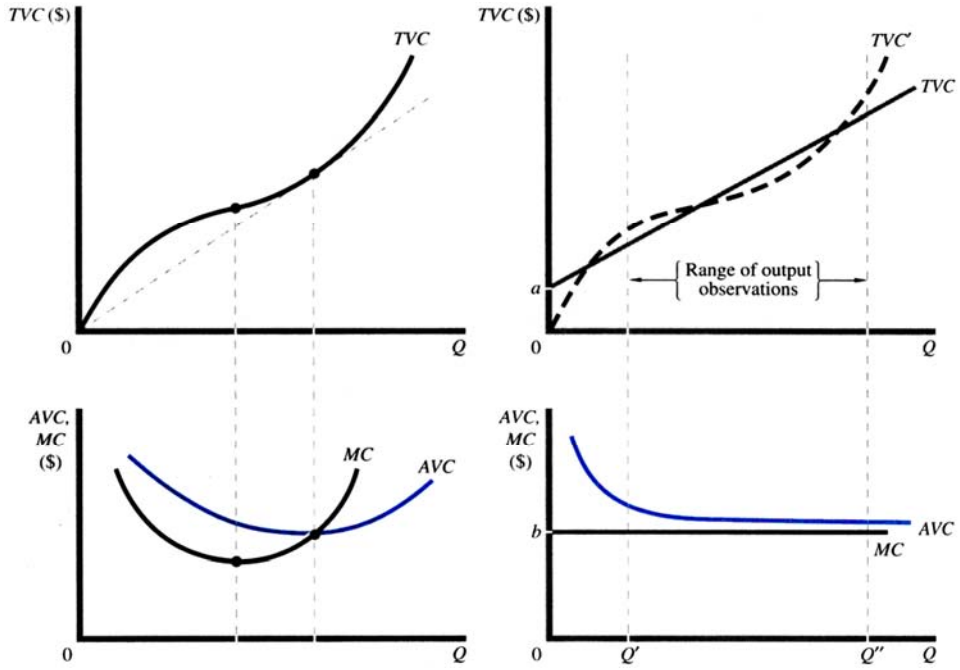
- **Quadratic relationship**

As output increases, total cost increases at an increasing rate

- **Linear relationship**

As output increases, total cost increases at a constant rate

Figure 1



Having estimated the parameters of the ***TVC*** curves (i.e; the value of a and b in Equation 5), we can use these estimated parameters to derive the corresponding AVC and MC functions of the firm, as indicated in Equations (6) and (7). We might note that estimated parameter ***a*** (the constant in estimated Regression 5) cannot be interpreted as the fixed cost of the firm since we are estimating the ***TVC*** function. Since ***Q = 0*** is usually far removed from the actual observed data points on the ***TVC*** curve (from ***Q' to Q"*** in the right panel of Figure1, no economic significance can be attached to the estimated parameter ***a***. We might also note that the ***AVC*** curve in the right panel becomes quite flat, approaching the value of (the horizontal MC curve). This i ***b*** s often observed in the actual empirical estimation. Another nonlinear theoretical from of the TVC curve that is often closely approximated by a linear ***TVC*** is the quadratic form. The quadratic ***TVC*** curve rises at an increasing rate (i.e., faces diminishing returns) throughout (and so do the corresponding ***AVC*** and ***MC*** curves). One possible explanation for this is that while the amount of capital (say the number of machines) that the firm has is fixed in the short run, the may keep some machines idle when output is low and bring them into operation by hiring more labor when it wants to increase output. Since the ratio of machines to output as well as machines to labor tends to remain constant in the face of changes in output, the firm's ***AVC*** and ***MC*** to remain approximately constant.

THE SHAPES OF SHORT-RUN COST FUNCTION

Three different specifications of cost functions are: cubic, quadratic and Linear. Each represents a possible shape of the cost curves. The economist, after collecting and adjusting the data, will use one of these specifications to measure the relationship between cost and output. Other statistical functions could be employed (e.g; the Cobb-Douglas power function), but the three shapes are the ones most frequently encountered in statistical studies, that is, cubic, quadratic or Linear.

EMPIRICAL ESTIMATION LONG-RUN COST CURVES

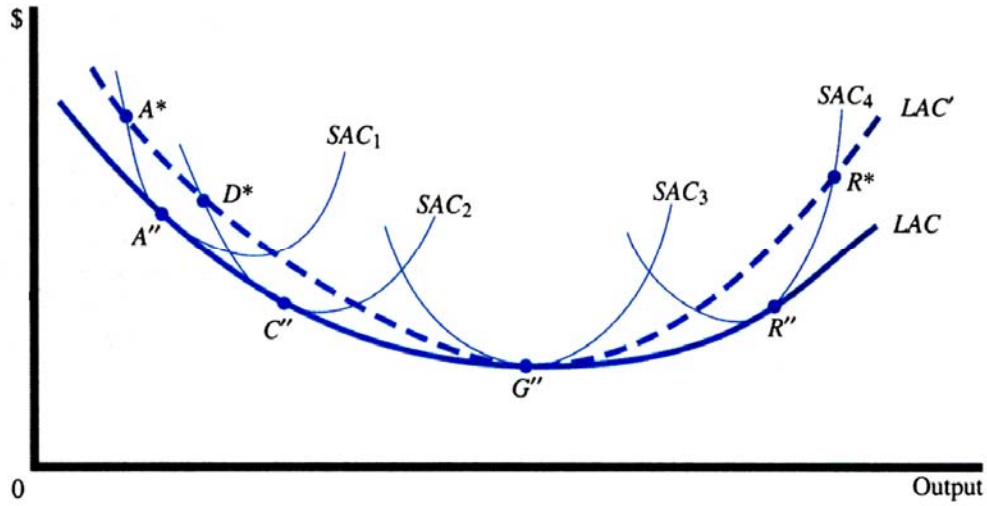
- Cross-Sectional Regression Analysis
- Engineering Method
- Survival Technique

ESTIMATING LONG-RUN COST FUNCTIONS WITH CROSS-SECTIONAL DATA

The empirical estimation of long-run cost curves is even more difficult than the estimation of short-run cost curves. The objective of estimating the long-run cost curves is to determine the best scale of plant for the firm to build in order to minimize the cost of producing the anticipated level of output in the long run. Theoretically, long-run cost curves can be estimated with regression analysis utilizing most of the time Cross-Sectional Data. Regression analysis using Cross-Sectional Data to estimate the long-run cost curves also presents some difficulties, however. For one thing, firms in different geographical regions are likely to pay different prices for their inputs, and so input prices must be included together with the levels of output as independent explanatory variables in the regression.

It may also be very difficult to determine if each firm is operating the optimal scale of plant at the optimal level of output (i.e; at the point on its ***SAC*** curve which forms part of its ***LAC*** curve). Specifically, in order to be able to estimate ***LAC*** curve ***A'' C'' G'' R''*** in Figure 2, the firms represented by ***SAC_1, SAC_2, SAC_3 and SAC_4*** must operate at points ***A'', C'', G'', and R'',*** respectively. If in fact the four firms are producing at points A****, C*, G*, and R*,*** respectively, we would be estimating the dashed ***LAC'*** curve, which overestimates the degree of both the economies and diseconomies of scale. The estimated long-run average cost curves seem to indicate sharply increasing returns to scale (falling ***LAC*** curve) at low levels of output followed by near-constant returns to scale at higher levels of output (i.e; the ***LAC*** curve seems to be L-shaped or nearly so).

Figure 2
Actual LAC versus empirically estimated LAC'



ESTIMATING LONG-RUN COST FUNCTIONS WITH ENGINEERING AND SURVIVAL TECHNIQUES

When sufficient data are not available for cross-sectional regression estimation of the long-run cost curves, the engineering or the survival techniques are used. **The engineering technique** uses knowledge of the physical relationship between inputs and output expressed by the

production function to determine the optimal input combination needed to produce, various levels of output by multiplying the optimal quantity of each input by the price of the input, we obtain the long-run cost function of the firm. Knowledgeable professionals calculate the quantity of inputs required to produce any quantity of outputs. The engineering technique is particularly useful in estimating the cost functions of new products or improved products resulting from the application of new technologies, where historical data are not available.

The advantage of the engineering technique over cross-sectional regression analysis is that it is based on the present technology, thus avoiding mixing the old and current technology used by different firms in cross-sectional analysis. Neither does the problem of different input prices in different geographical regions arise. Many of the difficult cost-allocation and input-valuation accounting problems that plague regression estimation are also avoided.

The engineering technique is not without problems, however. These arise because it deals only with the technical aspects of production without considering administrative, financing, and marketing costs; it deals with production under ideal rather than actual real-world conditions; and it is based on current technology, which may soon become obsolete. The engineering technique has been successfully applied to examine the cost-to-output relationship in many industrial sectors, such as petroleum refining and chemical production. The results obtained seem to confirm those obtained with cross-sectional regression analysis. That is, the ***LAC'*** Curve seem to be L-shaped. Generally, engineering cost estimates shows declining unit costs up to a point and substantially flat unit costs at higher production quantities.

Lecture 22

LINEAR PROGRAMMING

Linear programming (LP) has proven to be a skillful tool for solving problems encountered in a number of business, engineering, financial, and scientific applications. Though LP is applicable in a wide variety of contexts, it has been more frequently applied in production decisions. Production analysis also represents an excellent point of departure for introducing some basic LP concepts. We begin by defining the meaning of a production process and deriving isoquants. By then bringing in the production constraints, we show how the firm can determine the optimal mix of production processes to use in order to maximize output.

PRODUCTION PROCESSES ISOQUANTS IN LINEAR PROGRAMMING

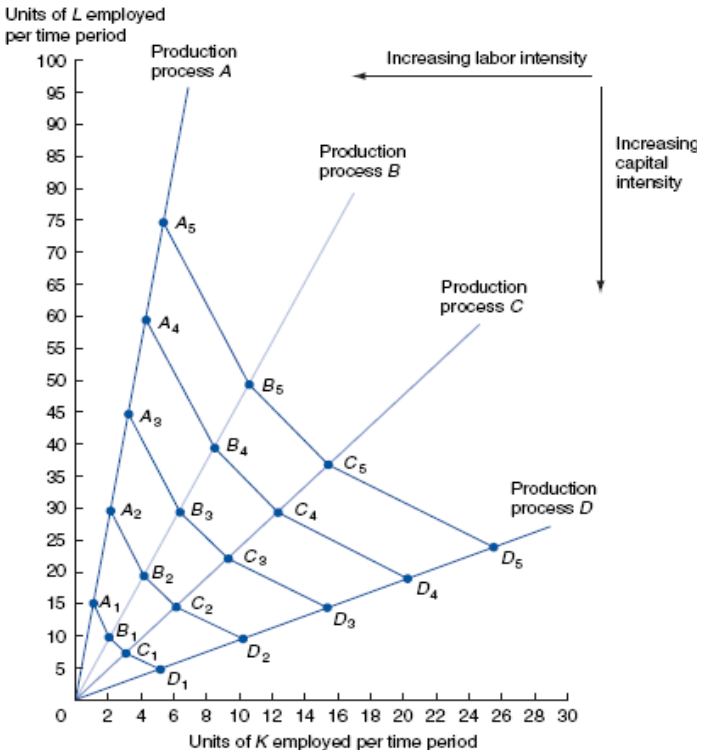
Assume that a firm produces a single product, *Q*, using two inputs, *L* and *K*, which might represent labor and capital. Instead of assuming continuous substitution between *L* and *K*, assume that *Q* can be produced using only four input combinations. In other words, four different production processes are available for making *Q*, each of which uses a different fixed combination of inputs *L* and *K*. The production processes might represent four different plants, each with its own fixed asset configuration and labor requirements. Alternatively, they could be four different assembly lines, each using a different combination of capital equipment and labor.

The four production processes are illustrated as rays in Figure 1. Process A requires the combination of 15 units of *L* and 1 unit of *K* for each unit of *Q* produced. Process *B* uses 10 units of *L* and 2 units of *K* for each unit of output. Processes *C* and *D* use 7.5 units of *L* and 3 units of *K*, and 5 units of *L* with 5 units of *K*, respectively, for each unit of *Q* produced. Each point along the production ray for process *A* combines *L* and *K* in the ratio 15 to 1; process rays *B*, *C*, and *D* are developed in the same way. Each point along a single production ray combines the two inputs in a fixed ratio, with the ratios differing from one production process to another. If *L* and *K* represent labor and capital inputs, the four production processes might be different plants employing different production techniques. Process *A* is very labor intensive in comparison with the other production systems, whereas *B*, *C*, and *D* are based on increasingly capital-intensive technologies.

Point *A*1 indicates the combination of *L* and *K* required to produce one unit of output using the *A* process. Doubling both *L* and *K* doubles the quantity of *Q* produced; this is indicated by the distance moved along ray *A* from *A*1 to *A*2. Line segment 0*A*2 is exactly twice the length of line segment 0A1 and thus represents twice as much output. Along production process ray *A*, the distance 0*A*1 = *A*1*A*2 = *A*2*A*3 = *A*3*A*4 = *A*4*A*5. Each of these line segments indicates the addition of one unit of output using increased quantities of *L* and *K* in the fixed ratio of 15 to 1.

Figure 1. Point *C*1 indicates the combination of *L* and *K* required producing 1 unit of *Q* using process *C*. The production of 2 units of *Q* using that process requires the combination of *L* and *K* indicated at point *C*2; the same is true for points *C*3, *C*4, and *C*5. Although production of additional units using process *C* is indicated by line segments of equal length, just as for process *A*, these line segments are of different lengths between the various production systems. Whereas each production process exhibits constant returns to scale, equal distances along *different* process rays do not ordinarily indicate equal output quantities.

Figure 1

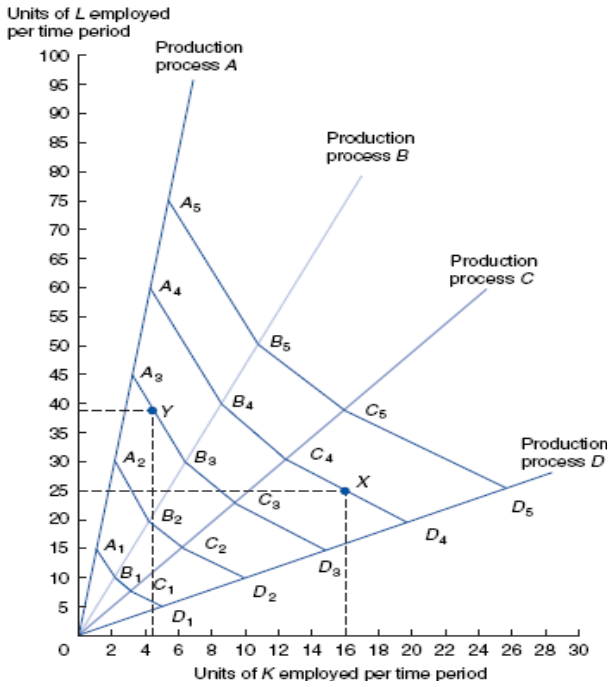


PRODUCTION ISOQUANTS

Joining points of equal output on the four production process rays creates a set of Isoquant curves. Figure 2 illustrates Isoquants for *Q* = 1, 2, 3, 4, and 5. Each Isoquant represents combinations of input factors *L* and *K* that can be used to produce a given quantity of output. Production Isoquants in linear programming are composed of linear segments connecting the various production process rays. Each of these Isoquant segments is parallel to one another. For example, line segment *A*1*B*1 is parallel to segment *A*2*B*2; Isoquant segment *B*3*C*3 is parallel to *B*2*C*2.

Points along each segment of an Isoquant between two process rays represent a combination of output from each of the two adjoining production processes. Consider point *X* in Figure 2, which represents production of 4 units of *Q* using 25 units of *L* and 16 units of *K*. That combination is possible by producing part of the output with process *C* and part with process *D*. In this case, 2 units of *Q* can be produced using process *C* and 2 units using process *D*.

Figure 2

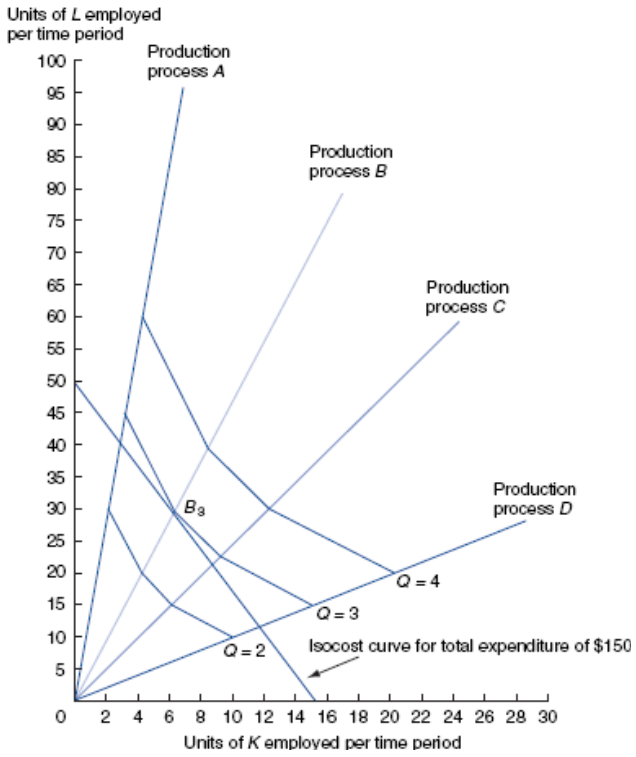


LEAST-COST INPUT COMBINATIONS

Adding Isocost curves to a set of Isoquants permits one to determine least-cost input combinations for the production of product *Q*. This is shown in Figure 3 under the assumption that each unit of *L* costs $3 and each unit of *K* costs $10. The Isocost curve illustrated indicates a total expenditure of $150.

The tangency between the Isocost curve and the Isoquant curve for *Q* = 3 at point *B*3 indicates that process *B*, which combines inputs *L* and *K* in the ratio 5 to 1, is the least-cost method of producing *Q*. For any expenditure level, production is maximized by using process *B*. Alternatively, process *B* is the least-cost method for producing any quantity of *Q*, given the assumed prices for *L* and *K*.

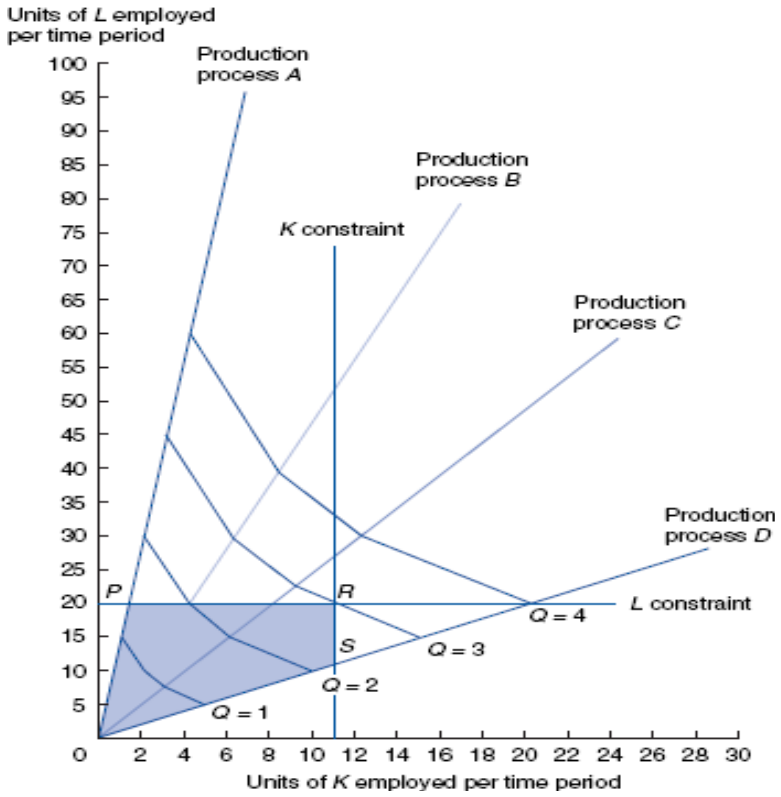
Figure 3



OPTIMAL INPUT COMBINATIONS WITH LIMITED RESOURCES

Frequently, firms faced with limited inputs during a production period find it optimal to use inputs in proportions other than the least-cost combination. To illustrate, consider the effect of limits on the quantities of *L* and K available in our example. Assume that only 20 units of *L* and 11 units of *K* are available during the current production period and that the firm seeks to maximize output of *Q*. These constraints are shown in Figure 4. The horizontal line drawn at *L* = 20 indicates the upper limit on the quantity of *L* that can be employed during the production period; the vertical line at *K* = 11 indicates a similar limit on the quantity of *K*.

Figure 4



Production possibilities for this problem are determined by noting that, in addition to limitations on inputs *L* and *K*, the firm must operate within the area bounded by production process rays *A* and *D*. Combining production possibilities with input constraints restricts the firm to operation within the shaded area on *0PRS* in Figure 4. This area is known as the **feasible space** in the programming problem.

HISTORICAL PERSPECTIVE OF LINEAR PROGRAMMING

The founders of the concept are Leonid Kantorovich, a Russian mathematician who developed linear programming problems in 1939, George B Dantzig, who published the simplex method in 1947, and John von Neumann, who developed the theory of the duality in the same year.

INTRODUCTION TO LINEAR PROGRAMMING

- A Linear Programming model seeks to maximize or minimize a linear function, subject to a set of linear constraints.
- The linear Programming model consists of the following components:
  - A set of decision variables.
  - An objective function.
  - A set of constraints.
  - A set of non-negativity constraint

- Constraints of production capacity, time, money, raw materials, budget, space, and other restrictions on choices. These constraints can be viewed as inequality constraints

- A "linear" programming problem assumes a linear objective function, and a series of linear inequality constraints

PROCEDURE USED IN FORMULATING AND SOLVING LP PROBLEMS

Linear programming is a useful method for analyzing and solving certain types of management decision problems. The most difficult aspect of solving a constrained optimization problem by LP is to formulate or state the problem in LP format. Simple LP problems with only a few variables are easily solved graphically or algebraically. More complex problems are almost always solved by the use of computers.

LP is a mathematical modeling technique used to determine a level of operational activity in order to achieve an objective, subject to restrictions called constraints.

- A linear function to be maximized
- Maximize: $\pi = c_1x_1 + c_2x_2$
- Subject to:

$$a_{11}x_1 + a_{12}x_2 \leq b_1$$
$$a_{21}x_1 + a_{22}x_2 \leq b_2$$
$$a_{31}x_1 + a_{32}x_2 \leq b_3$$

Non-negativity Constraints:

$$x_1 \geq 0 \; x_2 \geq 0.$$

- A linear function to be minimized
- Minimize: $C = c_1x_1 + c_2x_2$
- Subject to:

$$a_{11}x_1 + a_{12}x_2 \geq b_1$$
$$a_{21}x_1 + a_{22}x_2 \geq b_2$$
$$a_{31}x_1 + a_{32}x_2 \geq b_3$$

Non-negativity Constraints:

$$x_1 \geq 0 \; x_2 \geq 0.$$

COMPACT FORM MAXIMIZATION

- The problem is usually expressed in matrix *form*, and then becomes:
- Maximize: c^Tx
- Subject to: $Ax \leq b$,
- $x \geq 0$.

COMPACT FORM MINIMIZATION

- The problem is usually expressed in matrix *form*, and then becomes:
- Minimize: c^Tx
- Subject to: $Ax \geq b$,
- $x \geq 0$.

BASIC ASSUMPTIONS OF THE MODLE

- The parameter values are known with certainty.
- The objective function and constraints exhibit constant returns to scale.
- There are no interactions between the decision variables (the additivity assumption).
- The Continuity assumption: Variables can take on any value within a given feasible range

INEQUALITY CONSTRAINTS

Many production or resource constraints faced by managers are inequalities. Constraints often limit the resource employed to less than or equal to some fixed amount available. In other instances, constraints specify that the quantity or quality of output must be greater than or equal to some minimum requirement. Linear programming handles such constraint inequalities easily, making it a useful technique for finding the **optimal solution** to many management decision problems.

LINEARITY ASSUMPTION

(1) Constant prices for outputs (as in a perfectly competitive market).
(2) Constant returns to scale for production processes.
(3) Typically, each decision variable also has a non-negativity constraint. For example, the time spent using a machine cannot be negative.

As its name implies, linear programming can be applied only in situations in which the relevant objective function and constraint conditions are linear. Typical managerial decision problems that can be solved using the linear programming method involve revenue and cost functions and their composite, the profit function. Each must be linear; as output increases, revenues, costs, and profits must increase in a linear fashion. For revenues to be a linear function of output, product prices must be constant. For costs to be a linear function of output, both returns to scale and input prices must be constant. Constant input prices, when combined with constant returns to scale, result in a linear total cost function. If both output prices and unit costs are constant, then profit contribution and profits also rise in a linear fashion with output.

To illustrate, suppose that an oil company must choose the optimal output mix for a refinery with a capacity of 150,000 barrels of oil per day. The oil company is justified in basing its analysis on the $25-per-barrel prevailing market price for crude oil, regardless of how much is purchased or sold. This assumption might not be valid if the company was to quickly expand refinery output by a factor of 10, but within the 150,000 barrels per day range of feasible output, prices will be approximately constant. Up to capacity limits, it is also reasonable to expect that a doubling of crude oil input would lead to a doubling of refined output, and that returns to scale are constant.

PRODUCTION PLANNING FOR A MULTIPLE PRODUCTS

Many production decisions are quite complex. Consider the problem of finding the optimal output mix for a multi product firm facing restrictions on productive facilities and other inputs. This problem, which is faced by a host of companies producing consumer and producer goods alike, is readily solved with linear programming techniques.

- Linear programming problems can be solved using graphical techniques, algorithms using matrices, or using software, such as Fore Profit software, LINDO, MS Excel Solver.
- In the **graphical technique**, each inequality constraint is graphed as an equality constraint. The Feasible Solution Space is the area which satisfies all of the inequality constraints.
- The **Optimal Feasible Solution** occurs along the boundary of the Feasible Solution Space, at the extreme points or corner points.

OBJECTIVE FUNCTION SPECIFICATION

An equation that expresses the goal of a linear programming problem is called the **objective function**. Assume that the firm wishes to maximize total profits from the two products, *X* and *Y*,

during each period. If per-unit profit contribution (the excess of price over average variable costs) is $12 for product *X* and $9 for product *Y*, the objective function is:

$$\textbf{(1) Maximize } \Pi = \$12QX + \$9QY$$

QX and *QY* represent the quantities of each product produced. The total profit contribution, Π, earned by the firm equals the per-unit profit contribution of *X* times the units of *X* produced and sold, plus the profit contribution of *Y* times *QY*.

CONSTRAINT EQUATION SPECIFICATION

Table 1 specifies the available quantities of each input and their usage in the production of *X* and *Y*. This information is all that is needed to form the constraint equations. The table shows that 32 units of input *A* are available in each period. Four units of *A* are required to produce each unit of *X*, whereas 2 units of *A* are necessary to produce 1 unit of *Y*.

Because 4 units of *A* are required to produce a single unit of *X*, the total amount of *A* used to manufacture *X* can be written as 4*QX*. Similarly, 2 units of *A* are required to produce each unit of *Y*, so 2*QY* represents the total quantity of *A* used to produce product *Y*. Summing the quantities of *A* used to produce *X* and *Y* provides an expression for the total usage of *A*. Because this total cannot exceed the 32 units available, the constraint condition for input *A* is

$$\textbf{(2) } 4QX + 2QY \leq 32$$

The constraint for input *B* is determined in a similar manner. One unit of input *B* is necessary to produce each unit of either *X* or *Y*, so the total amount of *B* employed is 1*QX* + 1*QY*. The maximum quantity of *B* available in each period is 10 units; thus, the constraint requirement associated with input *B* is

$$\textbf{(3) } 1QX + 1QY \leq 10$$

Finally, the constraint relation for input *C* affects only the production of *Y*. Each unit of *Y* requires an input of 3 units of *C*, and 21 units of input *C* are available. Usage of *C* is given by the expression 3*QY*, and the relevant constraint equation is:

$$\textbf{(4) } 3QY \leq 21$$

Constraint equations play a major role in solving linear programming problems.

NON NEGATIVITY REQUIREMENT

Because linear programming is merely a mathematical tool for solving constrained optimization problems, nothing in the technique itself ensures that an answer makes economic sense. In a production problem for a relatively unprofitable product, the mathematically optimal output level might be a *negative* quantity, clearly an impossible solution. In a distribution problem, an optimal solution might indicate negative shipments from one point to another, which again is impossible.

To prevent economically meaningless results, a non negativity requirement must be introduced. This is merely a statement that all variables in the problem must be equal to or greater than zero. For the present production problem, the following expressions must be added:

$$QX \geq 0$$
and
$$QY \geq 0$$

Table 1
INPUTS AVAILABLE FOR PRODUCTION OF *X* AND *Y*

Input	Quantity Available per Time Period	Quantity Required per Unit of Output	
		X	Y
A	32	4	2
B	10	1	1
C	21	0	3

GRAPHIC SPECIFICATION AND SOLUTION
Having specified all the component parts of the firm's linear programming problem, the problem can now be illustrated graphically and analyzed algebraically.

ANALYTIC EXPRESSION
The decision problem is to maximize total profit contribution, Π, subject to resource constraints. This is expressed as

(1) Maximize $\Pi = \$12QX + \$9QY$
Subject to the following constraints:

(2) Input *A*: $4QX + 2QY \leq 32$

(3) Input *B*: $1QX + 1QY \leq 10$

(4) Input *C*: $3QY \leq 21$

Where

$QX \geq 0$ and $QY \geq 0$

Each variable and coefficient is exactly as specified previously.

GRAPHICAL SOLUTION OF LINEAR PROGRAMMING MODELS

- Graphical solution is limited to linear programming models containing only two decision variables. (Can be used with three variables but only with great difficulty.)
- Graphical methods provide visualization of how a solution for a linear programming problem is obtained.

THE STEPS FOLLOWED IN SOLVING A LP PROBLEM ARE:

1. Express the objective function as equations and the constraints as inequalities.
2. Graph the inequality constraints as equations, Plot model constraint on a set of coordinates in a plane
3. Identify the feasible solution space on the graph where all constraints are satisfied simultaneously
4. Graph the objective function as a series of isoprofit or Isocost lines. Plot objective function to find the point on boundary of this space (extreme points or corner points of the feasible region) that maximizes (or minimizes) value of objective function.
5. A solution is called feasible when it satisfies all the constraints. The Non-negativity Constraints are shown graphically by the positive quadrant. The complete determination

of the region of feasible solutions requires in addition the determination of the boundaries or limits set by the technical (functional) constraints, that is, the availability of the factors of production and the given stat of technology.

Lecture 23

LINEAR PROGRAMMING (CONTINUED 1)

GRAPHIC SPECIFICATION AND SOLUTION

Having specified all the component parts of the firm's linear programming problem, the problem can now be illustrated graphically and analyzed algebraically.

ANALYTIC EXPRESSION

The decision problem is to maximize total profit contribution, Π, subject to resource constraints. This is expressed as

(1) Maximize $\Pi = \$12QX + \$9QY$

Subject to the following constraints:

(2) Input *A*: $4QX + 2QY \leq 32$
(3) Input *B*: $1QX + 1QY \leq 10$
(4) Input *C*: $3QY \leq 21$

Where

$QX \geq 0$ and $QY \geq 0$

Each variable and coefficient is exactly as specified previously.

GRAPHICAL SOLUTION OF LINEAR PROGRAMMING MODELS

- Graphical solution is limited to linear programming models containing only two decision variables. (Can be used with three variables but only with great difficulty.)
- Graphical methods provide visualization of how a solution for a linear programming problem is obtained.

THE STEPS FOLLOWED IN SOLVING A LP PROBLEM ARE:

- Express the objective function as equations and the constraints as inequalities.
- Graph the inequality constraints as equations, Plot model constraint on a set of coordinates in a plane
- Identify the feasible solution space on the graph where all constraints are satisfied simultaneously
- Graph the objective function as a series of Isoprofit or Isocost lines. Plot objective function to find the point on boundary of this space (extreme points or corner points of the feasible region) that maximizes (or minimizes) value of objective function.
- Find the optimal solution at the extreme point or corner of the feasible region that touches the highest isoprofit line or the lowest Isocost line. This represents the optimal solution to the problem subject to the constraints faced.

GRAPHING THE FEASIBLE SPACE

A solution is called **feasible** when it satisfies all the constraints. The Non-negativity Constraints are shown graphically by the positive quadrant. The complete determination of the region of feasible solutions requires in addition the determination of the boundaries or limits set by the

technical (functional) constraints, that is, the availability of the factors of production and the given state of technology.

Here basically we have to determine the valid side for inequality constraints. That is we treat the inequality constraints of the problem as equations, graph them, and define the feasible region. The inequality sign indicates that the firm can use up to, but no more than, the 32 units of input A available to it to produce products X and Y. The firm can use less than 32 units of input A, but it cannot use more.

In Figure 1, the graph of the constraint equation for input A, 4QX + 2QY = 32, indicates the maximum quantities of X and Y that can be produced given the limitation on the availability of input A. A maximum of 16 units of Y can be produced if no X is manufactured; 8 units of X can be produced if no Y is manufactured. Any point along the line connecting these two outputs represents the maximum combination of X and Y that can be produced with no more than 32 units of A.

This constraint equation divides the XY plane into two half spaces. Every point lying on the line or to the left of the line satisfies the constraint expressed by the equation:

$4QX + 2QY \leq 32$

Every point to the right of the line violates that expression. Only points on the constraint line or to the left of it are in the feasible space. The shaded area of Figure 1 represents the feasible area limited by the constraint on input A.

In Figure 2, the feasible space is limited further by adding constraints for inputs *B* and *C*. The constraint on input *B* is expressed as $QX + QY = 10$. All combinations of *X* and *Y* lying on or to the left of the line connecting these two points are feasible with respect to utilization of input *B*.

The horizontal line at $QY = 7$ in Figure 2 represents the constraint imposed by input *C*. Because *C* is used only in the production of *Y*, it does not constrain the production of *X*. Seven units are the maximum quantities of *Y* that can be produced with 21 units of *C* available. These three input constraints, together with the non negativity requirement, completely define the feasible space shown as the shaded area of Figure 2. Only points within this area meet all constraints.

Figure 1

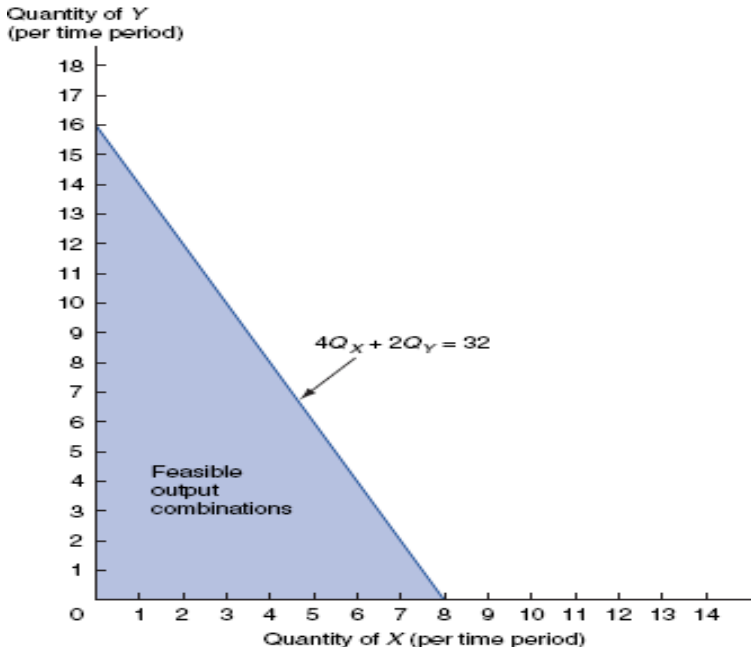
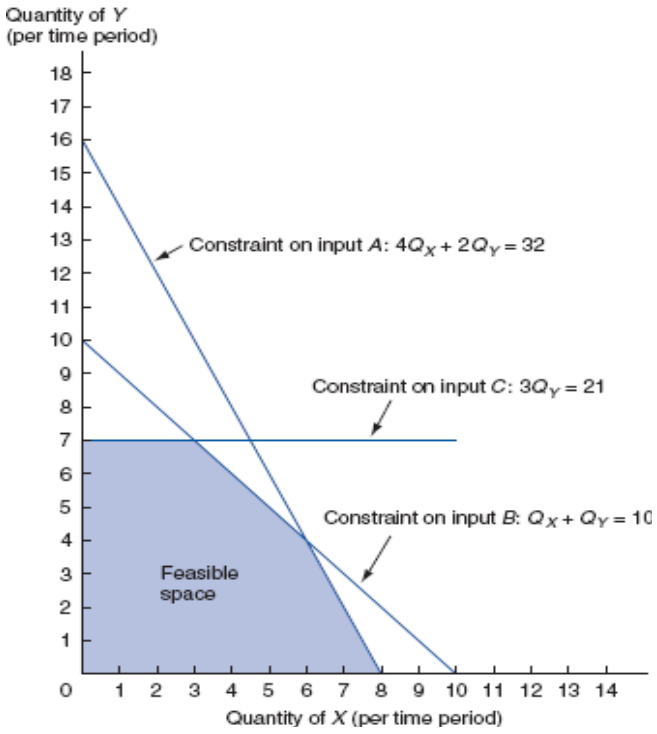


Figure 2



GRAPHING THE OBJECTIVE FUNCTION

The next step in solving the linear programming problem is to graph the objective function of the firm as a series of isoprofit lines. The slope of the objective function is influenced by the values of profit or cost. It is completely determined by co-efficient of the two variables of the objective function.

In our example, the objective function, Π= \$12*QX* + \$9*QY*, can be graphed in *QXQY* space as a series of isoprofit curves. This is illustrated in Figure 3, where isoprofit curves for \$36, \$72, \$108, and \$144 are shown. The isoprofit curve labeled Π= \$36 identifies each combination of *X* and *Y* that results in a total profit contribution of \$36; all output combinations along the Π = \$72 curve provide a total profit contribution of \$72; and so on. It is clear from Figure 3 that isoprofit curves are a series of parallel lines that take on higher values as one moves upward and to the right.

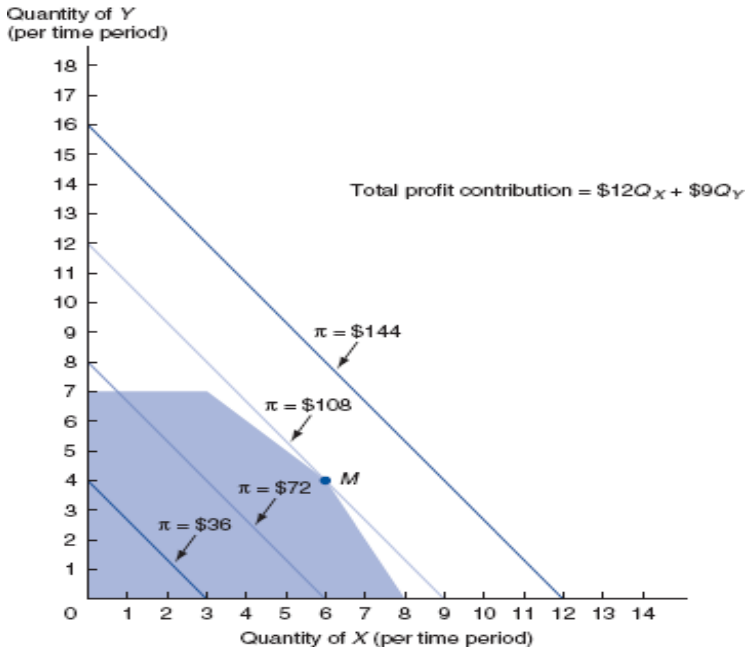
The general formula for isoprofit curves can be developed by considering the profit function Π = *aQX* + *bQY*, where *a* and *b* are the profit contributions of products *X* and *Y*, respectively. Solving the isoprofit function for *QY* creates an equation of the following form:

$QY = \Pi / b - a/b\ QX$

Given the individual profit contributions, *a* and *b*, the *QY* intercept equals the profit level of the isoprofit curve divided by the profit per unit earned on *QY*, Π/b. Slope of the objective function is given by the relative profitability of the two products, $-a/b$. Because the relative profitability of the products is not affected by the output level, the isoprofit curves consist of a series of parallel lines. In our example, all isoprofit curves have a slope of –4 / 3, or –1.33. That is:

π = 12X + 9Y **Solving for Y:**
Y = π/9 – 4/3X

Figure 3



GRAPHIC SOLUTION OF A MAXIMIZATION PROBLEM

Because the firm's objective is to maximize total profit, it should operate on the highest isoprofit curve obtainable. To see this point graphically, Figure 3 combines the feasible space limitations shown in Figure 2 with the family of isoprofit curves. Using this approach, point M in the figure is indicated as the optimal solution. At point M, the firm produces 6 units of *X* and 4 units of *Y*, and the total profit is $108 which is the maximum available under the conditions stated in the problem. No other point within the feasible spaces touches so high an isoprofit curve.

M can be identified as the point where *QX* = 6 and *QY* = 4. At *M*, constraints on inputs *A* and *B* are binding. At *M*, 32 units of input *A* and 10 units of input *B* are being completely used to produce *X* and *Y*. Thus, Equations 2 and 3 can be written as equalities and solved simultaneously for *QX* and *QY*. Subtracting two times Equation 3 from Equation 2 gives

4*QX* + 2*QY* = 32

Minus <u>2*QX* + 2*QY* = 20</u>
2*QX* = 12

***QX* = 6**

Substituting 6 for *QX* in Equation 3 results in

6 + *QY* = 10

***QY* = 4**

Notice that the optimal solution to the linear programming problem occurs at a **corner point** of the feasible space. A corner point is a spot in the feasible space where the *X*-axis, *Y*-axis, or

constraint conditions intersect. The optimal solution to any linear programming problem always lies at a corner point. Because all of the relations in a linear programming problem must be linear by definition, every boundary of the feasible space is linear. Since, the objective function is linear; the constrained optimization of the objective function takes place at a corner of the feasible space.

The final step in solving the linear programming problem graphically is to determine the mix of products *X* and *Y* that the firm should produce in order to reach the highest isoprofit line. This is obtained by superimposing the isoprofit lines on the feasible region shown in Figure 3, which shows that the highest isoprofit line that the firm can reach subject to the constraints it faces is Π = $108.

This is reached at point *M* where the firm produces 6*X* and 4*Y* and the total contribution to profit $108 is maximum at $12(6) - $9(4) = $108. We might notice that point *M* is at the intersection of the constraint lines for inputs *A* and *B* but below the constraint line for input *C.* This means that inputs *A* and *B* are fully utilized, while input *C* is not. We then say that inputs *A* and *B* are **binding constraints,** while input *C* is nonbinding or is a **slack variable.**

The firm's manager is interested in knowing not only the quantities of products *X* and *Y* that the firm must produce in order to maximize profits, but he is also interested in knowing which inputs are binding and which are non-binding (or slack) at the profit-maximizing point. This information is normally provided by the computer solution to the linear programming problem.

ALGEBRAIC SOLUTION OF THE PROFIT MAXIMIZATION PROBLEM

Algebraic Specification and Solution

- Dummy Variable (Slack and Surplus collectively called are called Dummy Variables
- Slack variables convert ≤ constraints into equalities.
- Surplus variables convert ≥ constraints into equalities.
- Zero slack implies full utilization.
- Positive slack implies excess capacity.
- Standard form requires that all constraints be in the form of equations.
- A slack variable represents unused resources.
- A slack variable contributes nothing to the objective function value.

- Algebraic Solution
  - Corner point with highest value is maximum.
  - Corner point with lowest value is minimum.
- Slack Variables at the Solution Point
  - Binding constraints imply no slack.
  - Nonbinding constraints imply slack.
- Computer-based solution methods work best for complex LP problems.

CORNER POINT PROPERTY is a very important property of Linear Programming problems:
This property states that optimal solution to LP problem will **always** occur at a corner point. The solution point will be on the boundary of the feasible solution area and at one of the corners of the boundary where two constraint lines intersect. This is one of the basic theorems of linear programming. This is that in searching for the optimal solution, we need to examine and compare the levels of Π at only the extreme points (corners) of the feasible region and can ignore all other points inside or on the borders of the feasible region. That is, with a linear objective function and linear input constraints, the optimal solution will always occur at one of

the corners. Since each corner is formed by the intersection of two constraint lines, the coordinates of the intersection point *QX* and *QY* at the corner can be found by solving simultaneously the equations of the two intersecting lines.

In the case of less-than-or-equal-to constraints, **slack variables** are used to increase the left side to equal the right side limits of the constraint conditions. In our LP problem, one slack variable is added to each constrain. Since this firm is faced with capacity constraints on input factors A, B, and C, we have to add three slack variables: S_A, S_B and S_C. With slack variables, each constraint equation becomes equality rather than an inequality.

- The slack for "less than or equal to" constraints is the difference between the right hand side of an equation and the value of the left hand side after substituting the optimal values of the decision variables.
- The slack represents the amount of unused units of the right hand side resources.

After adding the relevant slack variable, the complete specification of the illustrative programming problem is as follows:

(1) Maximize π= \$12*QX* + \$9*QY*

Subject to these constraints:

(2) 4*QX* + 2*QY* + *SA* = 32

(3) 1*QX* + 1*QY* + *SB* = 10

(4) 3*QY* + *SC* = 21

Where

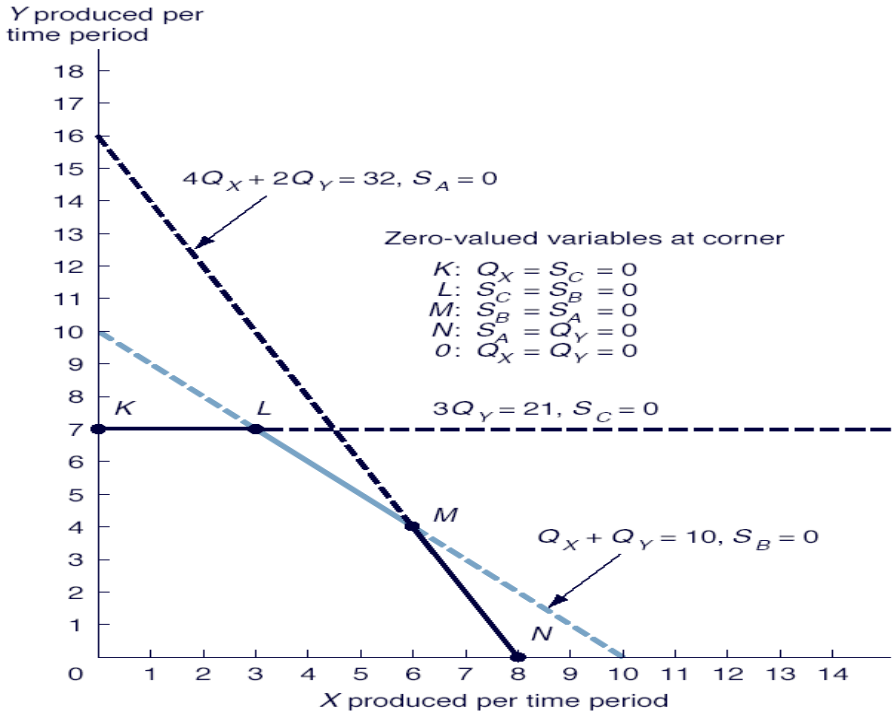
***QX* ≥0, *QY* ≥0, *SA* ≥0, *SB* ≥0, *SC* ≥0**

The problem is to find the set of values for variables *QX*, *QY*, *SA*, *SB*, and *SC* that maximizes Equation 1 and at the same time satisfies the constraints imposed by Equations 2, 3, and 4. By substituting the values of QX and QY at each corner of the feasible region into the objective function, we can then determine the firm's total profit contribution at each corner. These are shown in Table 1and Figure 4. The optimal or profit-maximizing point is at corner *M* at which **π= \$108.**

Table 1
Corner Point Solution

Corner point	X	Y	12X + 9Y	Profit
0	0	0	12(0) + 9(0)	$ 0
N	8	0	12(8) + 9(0)	$ 96
M	6	4	12(6) + 9(4)	$108
L	3	7	12(3) + 9(7)	$ 99
K	0	7	12(0) + 9(7)	$ 63

Figure 4



Graphic Solution of the Cost Minimization Problem

In order to solve graphically the cost minimization linear programming problem the first step is to treat each inequality constraint as an equation and plot it. Since each inequality constraint is expressed as "equal to or greater than," all points on or *above* the constraint line satisfy the particular inequality constraint. The feasible region is then given by the shaded area in Figure 6. All points in the shaded area simultaneously satisfy all the inequality and non negativity constraints of the problem. In the second step, we incorporate the objective function that is to be minimized. We superimpose the lowest Isocost line on the feasible region in Figure 7, where at point A the cost is minimized Z = $24.

A Minimization Model Example

Minimize Z = 6x_1$ + 3x_2
Subject to
2x_1 + 4x_2 ≥ 16 lb nitrogen
4x_1 + 3x_2 ≥ 24 lb phosphate
x_1, x_2 ≥ 0

Figure 5

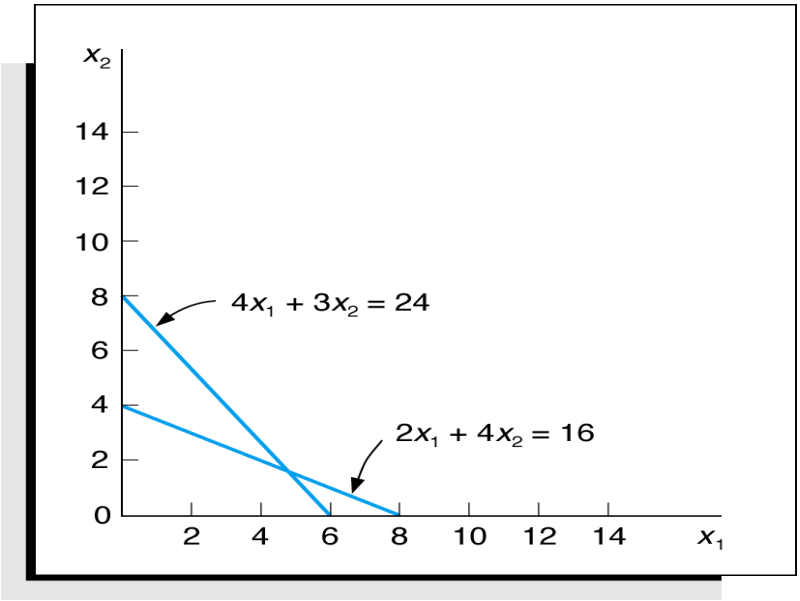


Figure 6

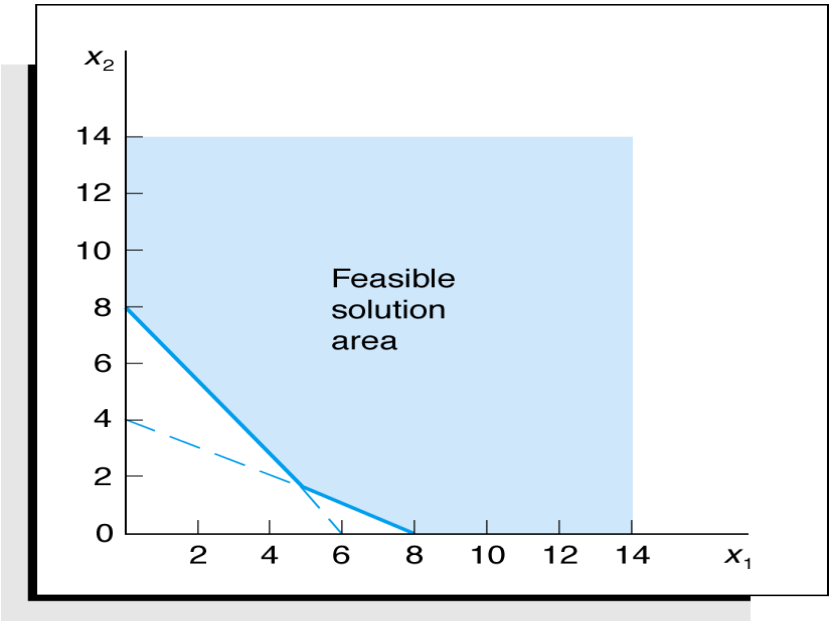


Figure 7

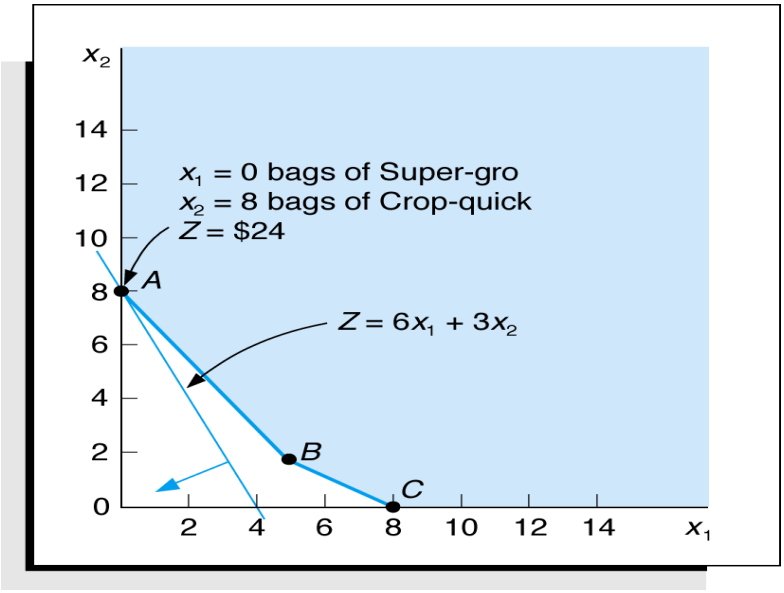
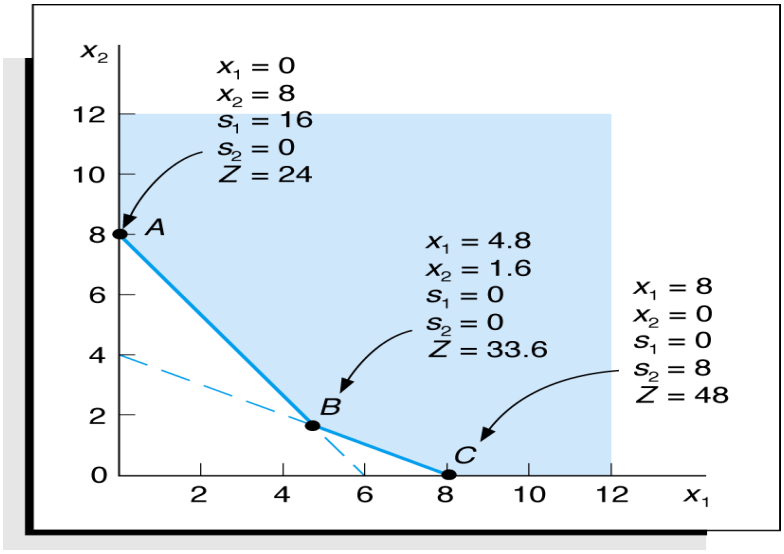


Figure 8



ALGEBRAIC SOLUTION OF THE COST MINIMIZATION PROBLEM

The cost minimization linear programming problem solved graphically above can also be solved algebraically by identifying (algebraically) the corners of the feasible region and then comparing the costs at each corner.

The **surplus** for "greater than or equal to" constraints is the difference between the right hand side of an equation and the value of the left hand side after substituting the optimal values of the decision variables. The surplus represents the number of units in which the optimal solution causes the constraint to exceed the right hand side lower limit.

- A surplus variable is subtracted from ≥ constraint to convert it to an equation (=).
- A surplus variable represents an excess above a constraint requirement level.
- Surplus variables contribute nothing to the calculated value of the objective function.

Adding surplus variables to our minimization problem:

Minimize Z = \$6x_1 + 3x_2 + 0s_1 + 0s_2
subject to
2x_1 + 4x_2 - s_1 = 16
4x_1 + 3x_2 - s_2 = 24
x_1, x_2, s_1, s_2 >= 0

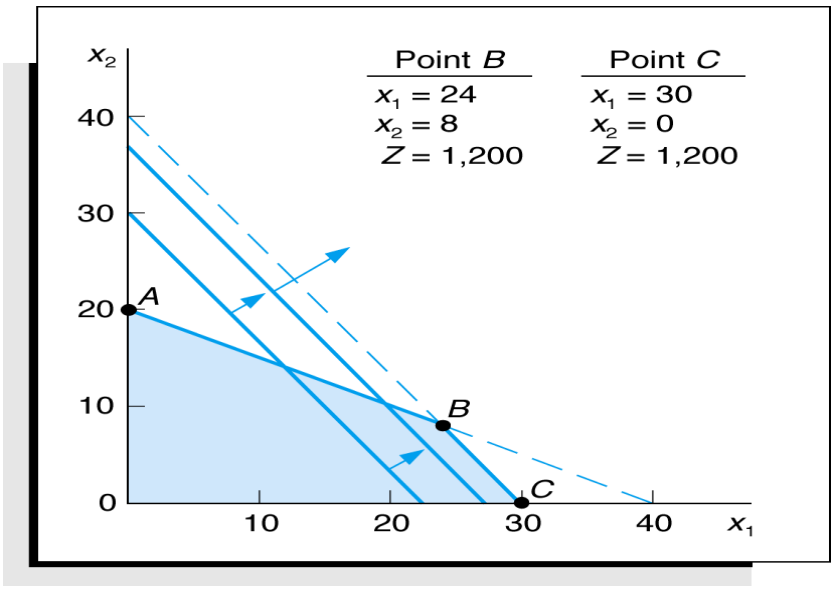
Corner Point Solution
Table 2

Point	Ordered Pair	$Z = 6x_1 + 3x_2$	Cost
A	(0,8)	6(0) + 3(8)	$24
B	(4.8, 1.6)	6(4.8) + 3(!.6)	$33.6
C	(8,0)	6(8) + 3(0)	$48

SPECIAL CASES OR IRREGULAR TYPES OF LINEAR PROGRAMMING PROBLEMS
The general rules do not apply for some linear programming models. Special types of problems include those with:

1. Multiple optimal solutions
2. Infeasible solutions
3. Unbounded solutions

MULTIPLE OPTIMAL SOLUTIONS
Figure 9



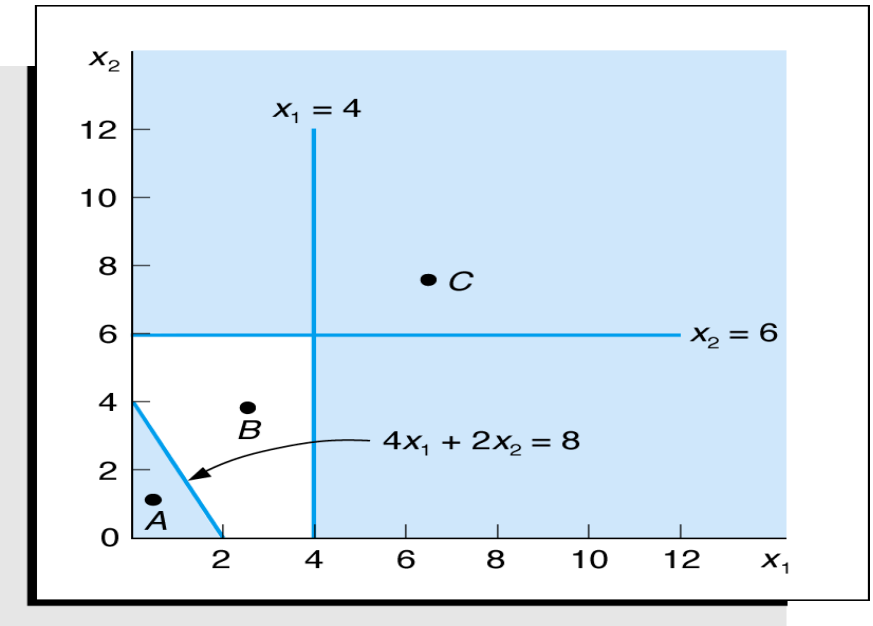
Objective function is parallel to one of the constraint line: that is they have the same slope Line segment BC in Figure 9. An LP problem may have **more than one** *optimal solution*. This is called a case of multiple optimal solutions or Alternate Optimal Solutions.

- Graphically, when the **isoprofit (or Isocost) line** runs *parallel* to a *constraint* in problem which lies in direction in which **isoprofit (or Isocost)** line is located.
- In other words, when they have same slope.

AN INFEASIBLE PROBLEM

Every possible solution violates at least one constraint. The direction of inequality is opposite, that is a problem with no feasible region because of the presence of conflicting constraints. As a result the feasible region is not a convex set.

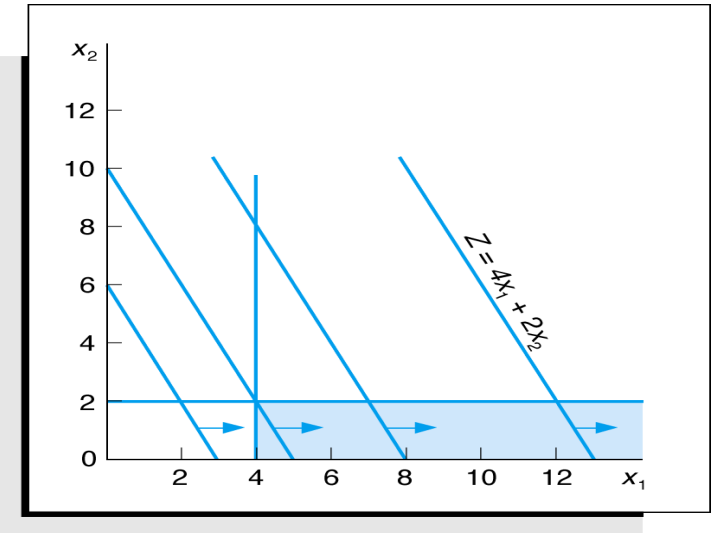
An Infeasible Solution
Figure 10



AN UNBOUNDED PROBLEM

Value of objective function increases indefinitely: This usually applies to Maximization problems. In this case LP model does not have a finite solution as shown in Figure 11.

Figure 11



Lecture 24

LINEAR PROGRAMMING (CONTINUED 2)

DUAL IN LINEAR PROGRAMMING (LP)

DUALITY CONCEPT

- Pairs of symmetrical LP problems are called the primal and the dual.
- Every primal has a dual and *vice versa*.
- Primal and dual solutions are related.

SHADOW PRICES: Assuming there are no other changes to the input parameters, the change to the objective function value per unit increase to a right hand side of a constraint is called the "Shadow Price". That is Change in objective value = [Shadow price][Change in the right hand side value]. The managers are interested to know if it is worthwhile to increase its production by purchasing additional units of raw materials and by either expanding its production facilities or working overtime.

THE MEANING OF DUAL AND SHADOW PRICES

Every linear programming problem, called the **primal problem,** has a corresponding or symmetrical problem called the **dual problem.** A profit maximization primal problem has a cost minimization dual problem, while a cost minimization primal problem has a profit maximization dual problem. The solutions of a dual problem are the **shadow prices.** They give the change in the value of the objective function per unit change in each constraint in the primal problem. For example, the shadow prices in a profit maximization problem indicate how much total profits would rise per unit increase in the use of each input. Shadow prices thus provide the imputed value or marginal valuation or worth of each input to the firm. If a particular input is not fully employed, its shadow price is zero because increasing the input would leave profits unchanged and its slack variable > 0, that is, un-used resource in case of a maximization problem. A firm should increase the use of the input as long as the marginal value or shadow price of the input to the firm exceeds the cost of hiring the input.

DUAL SPECIFICATION AND SOLUTION

- **Dual Objective Function**
  - Dual of profit maximization problem seeks minimum cost solution given constraints.
  - Dual of minimum cost problem seeks highest production value given resource constraints.
- **Dual Constraints**
  - Binding constraints imply no slack.
  - Nonbinding constraints imply slack.
- **Dual Slack Variables**
  - Binding constraints imply zero slack variable values.
  - Nonbinding constraints imply nonzero slack variables.

RULES OF TRANSFORMING PRIMAL TO OBTAIN THE DUAL

1. The direction of optimization is reversed.
2. The inequality sign of the constraints are reversed, but the non-negativity restraint on decision variable is always maintained
3. The rows of the coefficient matrix of the constraints in the primal are transposed into Columns for the coefficient matrix of constraints in the dual.

4. The row vector of coefficients in the objective function in the primal is transposed into a column vector of constants in the dual. This mean that the no of choice variables in a Primal = the no of constraints in the Dual.
5. The column vector of constants from the primal constraints is transposed into a row vector of coefficients for the objective function in the dual.
6. Primal decision variables (xj) are replaced by dual decision variables (yi).

Primal	**Dual**
Maximize:	**Minimize:**
$\pi = \mathbf{c}_1\mathbf{x}_1 + \mathbf{c}_2\mathbf{x}_2 + \mathbf{c}_3\mathbf{x}_3$	$C = \mathbf{r}_1\mathbf{y}_1 + \mathbf{r}_2\mathbf{y}_2$
Subject to:	**Subject to:**
$a_{11}\mathbf{x}_1 + a_{12}\mathbf{x}_2 + a_{13}\mathbf{x}_3 \leq \mathbf{r}_1$	$a11\mathbf{y}1 + a21\mathbf{y}2 \geq \mathbf{c}1$
$a_{21}\mathbf{x}_1 + a_{22}\mathbf{x}_2 + a_{23}\mathbf{x}_3 \leq \mathbf{r}_2$	$a_{12}\mathbf{y}_1 + a_{22}\mathbf{y}_2 \geq \mathbf{c}_2$
	$a_{13}\mathbf{y}_1 + a_{23}\mathbf{y}_3 \geq \mathbf{c}_3$
and	and
$\mathbf{x}_1 \geq ,0\ \mathbf{x}_2 \geq 0,\ \mathbf{x}_3 \geq 0$	$\mathbf{y}_1 \geq ,0\ \mathbf{y}_2 \geq 0$

DUALITY THEOREMS

DUALITY THEOREM I

a) The optimal values of the primal and the dual objective functions are always identical, provided that optimal feasible solutions do exist.
b) The maximum value of the primal (profit max problem) equals the minimum value of the dual (cost minimization) problem.

DUALITY THEOREM II (COMPLEMENTARY SLACKNESS)

a) If a certain *choice* variable in a LP is optimally *nonzero*, then the corresponding *dummy* variable in the counterpart program must be optimally *zero*.

$Yi > 0 \Rightarrow Si = 0$ and $Xj > 0 \Rightarrow tj = 0$

b) If a certain *dummy* variable in a LP is optimally *nonzero*, then the corresponding *choice* variable in the counterpart program must be optimally *zero*.

$Si > 0 \Rightarrow Yi = 0$ and $tj > 0 \Rightarrow Xj = 0$

ADVANTAGES OF THE DUAL

From the Duality theorems, it is clear that the solution of one program provides full solution to the other. This is beneficial because:

1. It enables us to solve Minimization problem in terms of Maximization, which is frequently easier to solve.
2. For primal with 3 decision variables, the dual reduces the program to 2 decision variables, which can be solved graphically.

THE DUAL OF PROFIT MAXIMIZATION

We formulate and solve the dual problem for the constrained profit maximization problem examined in the previous Lesson. In the dual problem we seek to minimize the imputed values, or shadow prices, of inputs *A*, *B*, and *C* used by the firm. Defining *VA*, *VB*, and *VC* as the shadow prices of inputs *A*, *B*, and *C*, respectively, and *C* as the total imputed value of the fixed quantities of inputs *A*, *B*, and *C* available to the firm, we can write the dual objective function as

Minimize $C = 32VA + 10VB + 21VC$

Where the coefficients 32, 10, and 21 represent, respectively, the fixed quantities of inputs *A*, *B*, and *C* available to the firm. The constraints of the dual problem assumes that the sum of the

shadow price of each input times the amount of that input used to produce 1 unit of a particular product must be equal to or larger than the profit contribution of a unit of the product.

Primal	**Dual**
Maximize: $\pi = 12X + 9Y$	***Minimize*** $C = 32VA + 10VB + 21VC$
Subject to:	***Subject to:***
A: $4X + 2Y \leq 32$	*A:* $4VA + VB \geq 12$
B: $X + Y \leq 10$	*B:* $2V_A + V_B + 3V_C \geq 9$
C: $3Y \leq 21$	
Where	**Where**
$X \geq 0$ and $Y \geq 0$.	$V_A \geq 0$, $V_B \geq 0$, $V_C \geq 0$

We find the values of the decision variables (*VA*, *VB*, and *VC*) at each corner and choose the corner with the lowest value of *C.* Since we have three decision variables and this requires a three-dimensional figure, which is difficult to draw and interpret, therefore, we solve the above dual problem algebraically. The algebraic solution is simplified because in this case we know from the solution of the primal problem (Lesson # 23) that input *C* is a slack variable so that *VC* equals zero. Setting *VC* = 0 and then subtracting the first constraint from the second constraint, treated as equations, we get:

$4V_A + V_B = 12$
Minus $2V_A + V_B = 9$

So that: $2V_A = 3$ or $V_A = \$1.5$, $V_B = \$6$ and $V_C = \$0$:
C = $32V_A + 10V_B + 21V_C$
C = 32(1.5) + 10(6) +21(o) = $108

This is the minimum cost that the firm would incur in producing 6 unit of X and 4 units of Y (the solution of the primal profit maximization problem in Lesson 23). We might notice that the maximum profits found in the solution of the primal problem (that is, $108) equals the minimum cost in the solution of the corresponding dual problem (that is, *C* = $108) as required by the duality theorem I.

Substituting $V_A = 1.5$ and $V_B = 6$ and $V_C = 0$ into constraints, we get

$4(1.5) + 6 + \mathbf{S}_A = 12$
$6 + 6 + \mathbf{S}_A = 12$
$2(1.5) + 6 + 3(0) + \mathbf{S}_B = 9$
$3 + 6 + 0 + \mathbf{S}_B = 9$

$\mathbf{S}_{A\,=\,}\mathbf{S}_{B\,=\,0}$

Which implies that both constraints are binding and they are fully utilized.

CORNER POINT SOLUTION OF THE DUAL PROBLEM

Table 1

Point	V_A	V_B	V_C	t_1	t_2	Imputed Value
1	0	0	0	-12	-9	---
2	0	12	0	0	3	$120
3	4.5	0	0	6	0	$144
4	3	0	1	0	0	$117
5	1.5	6	0	0	0	$108

USING THE DUAL SOLUTION TO SOLVE THE PRIMAL

Maximize: $\pi = 12X + 9Y + 0SA + 0SB + 0SC$

Subject to:

Input A: $4X + 2Y + S_A = 32$

Input B: $X + Y + S_B = 10$

Input C: $3Y + S_C = 21$

Where $X, Y, S_A, S_B, S_A >= 0$

$4X + 2Y = 32$

Minus $2 (X + Y = 10)$

So that $2X = 12$

and : $X = 6$ & $Y = 4$

Since $\pi = 12X + 9Y$ $\pi = 12(6) + 9(4) = \$108$

$3(4) + S_C = 21$

$S_C = 21 – 12 = 9$

As a result, we get the same optimal values for the decision variables, slack variable and the objective function.

DUALITY PROPERTIES

Some relationships between the primal and dual problems:

1. If one problem has feasible solutions and a bounded objective function (and so has an optimal solution), then so does the other problem, so both the weak and the strong duality properties are applicable
2. If the optimal value of the primal is unbounded then the dual is infeasible.
3. If the optimal value of the dual is unbounded then the primal is infeasible.

Lesson 25

COMPETITIVE MARKETS

A MARKET consists of all firms and individuals willing and able to buy or sell a particular product.

MARKET STRUCTURE describes the competitive environment in the market for any good or service. Market structure refers to the competitive environment in which the buyers and sellers of the product operate. Market structure is typically characterized on the basis of four important industry characteristics:

1. The number and size distribution of active buyers and sellers and potential entrants,
2. The degree of product differentiation,
3. The amount and cost of information about product price and quality,
4. Conditions of entry and exit.

Effects of market structure are measured in terms of the prices paid by consumers, availability and quality of output. In general, the greater the number of market participants, the more vigorous is price and product quality competition. The more even the balance of power between sellers and buyers. As a result, the more likely it is that the competitive process will yield maximum benefits.

POTENTIAL ENTRANT

A **potential entrant** is an individual or firm posing a sufficiently credible threat of market entry to affect the price/output decisions of incumbent firms. Potential entrants play extremely important roles in many industries. Some industries with only a few active participants might at first appear to hold the potential for substantial economic profits. However, a number of potential entrants can have a substantial effect on the price/output decisions of incumbent firms. For example, Dell, Gateway, Hewlett-Packard, IBM, and other leading computer manufacturers are viable potential entrants into the computer component manufacturing industry. These companies use their threat of potential entry to obtain favorable prices from suppliers of microprocessors, monitors, and peripheral equipment.

FACTORS THAT SHAPE THE COMPETITIVE ENVIRONMENT

Effect of Product Characteristics on Market Structure

Transportation service is available from several sources; railroads compete with bus lines, truck companies, airlines, and private autos. The substitutability of these other modes of transportation for rail service increases the degree of competition in the transportation service market. Good substitutes always increase competition.

EFFECT OF ENTRY AND EXIT CONDITIONS ON COMPETITION

In order to maintain above-normal profits over the long run requires barriers to entry, mobility, or exit. A barrier to entry is any factor that creates an advantage for existing firms over new arrivals. Legal rights such as patents and provincial or federal licenses can present substantial barriers to entry in beverages, pharmaceuticals, cable television, television and radio broadcasting, and other industries.

A barrier to mobility in any factor that creates an advantage for large leading firms over smaller non leading rivals. Factors that sometimes create barriers to entry and/or mobility include substantial economies of scale, scope economies, large capital or skilled-labor requirements, and ties of customer loyalty.

Competitive forces can also be diminished through barriers to exit just as barriers to entry could. A barrier to exit is any restriction on the ability of incumbents to redeploy assets from one industry or line of business to another.

EFFECT OF PRODUCT DIFFERENTIATION ON COMPETITION

Product differentiation includes any real or perceived differences in the quality of goods and services offered to consumers. Sources of product differentiation include all of the various forms of advertising promotion, plus new products and processes made possible by effective programs of research and development that is innovation.

In short, market structure is broadly determined by entry and exit conditions. Low regulatory barriers, modest capital requirements, and nominal standards for skilled labor and other inputs all increase the likelihood that competition will be vigorous. Because all of these elements of market structure have important consequences for the price/output decisions made by firms, the study of market structure is an important ingredient of managerial economics.

MARKET STRUCTURE AND DEGREE OF COMPETITION

Market structure refers to the competitive environment in which the buyers and sellers of the product operate. Four types of market structure are usually identified. These are perfect competition at one extreme, pure monopoly at the opposite extreme, and monopolistic competition and oligopoly in between.

PERFECT COMPETITION (NO MARKET POWER)

- Many buyers and sellers
- Buyers and sellers are price takers
- Product is homogeneous
- Very easy market entry and exit
- Non price competition not possible
- Perfect mobility of resources
- Economic agents have perfect knowledge

Examples: Stock Market, agricultural products, financial instruments, precious metals, petroleum products, prominent markets for intermediate goods and services, e.g., discount retailing, unskilled labor market.

MONOPOLY (absolute market power s.t. Govt regulation)

- One firm, firm is the industry
- No close substitutes for product
- Significant barriers to resource mobility
- Market entry and exit difficult or legally impossible
- Non price competition not necessary

Examples: pharmaceuticals, Microsoft, Government franchise: Post office, Water Supply, Energy, National Airlines

Monopolistic Competition (market power based on Differentiated products)

- Many sellers and buyers (large no of relatively small firms acting independently)
- Differentiated product
- Market entry and exit relatively easy
- Non price competition is very important
- Perfect mobility of resources

Examples: Fast-food outlets, boutiques, restaurants

OLIGOPOLY (market power based on product differentiation and/or the firm's dominance in the market)

- Few sellers and many buyers (small no of relatively large firms)
- Product may be homogeneous or differentiated
- Market entry and exit difficult
- Non price competition is very important among firms selling Differentiated products
- Barriers to resource mobility

Examples: Automobile manufacturers, oil refining, processed foods, airlines

CHARACTERISTICS OF PERFECTLY COMPETITIVE MARKETS

Perfect competition exists when individual producers have no influence on market prices; they are price takers as opposed to price makers. This lack of influence on price typically requires.

- Large numbers of buyers and sellers. Each firm produces a small portion of industry output, and each customer buys only a small part of the total.
- Product homogeneity. The output of each firm is essentially the same as the output of any other firm in the industry.
- Free entry and exit. Firms are not restricted from entering or leaving the industry.
- Perfect dissemination of information. Cost, price, and product quality information is known by all buyers and all sellers.
- Opportunity for normal profit in long run equilibrium. Fierce price competition keeps P = MC and P =AR = AC.
- Non price competition not possible.

There is a great number of buyers and sellers of the product, and each seller and buyer is too small in relation to the market to be able to affect the price of the product. This means that a change in the output of a single firm will not affect the market price of the product. Similarly, each buyer of the product is too small to be able to extract from the seller such things as quantity discounts and special credit terms.

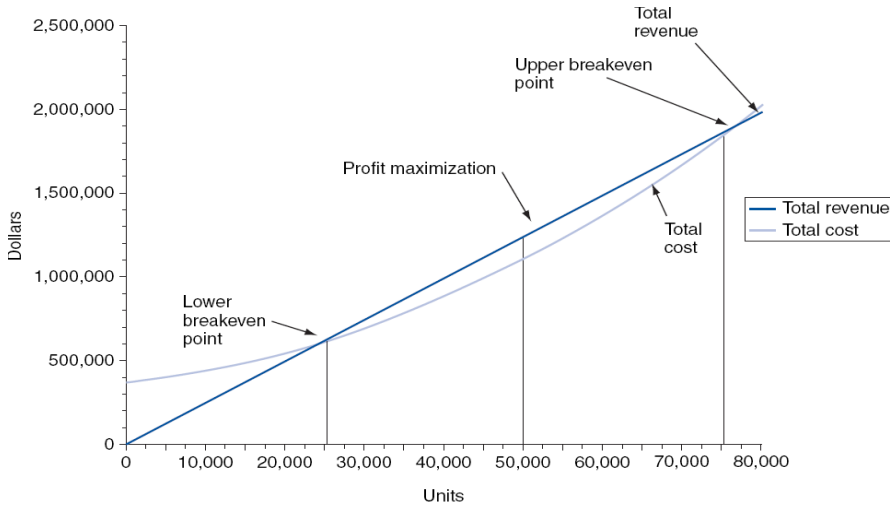
The product of each competitive firm is homogeneous, identical, or perfectly standardized. An example of this might be grading of wheat and cotton crops. As a result buyers cannot distinguish between the output of one firm and the output of another, so they are indifferent from which firm they buy the product.
Under perfect competition, there is perfect mobility of resources. That is, workers and other inputs can easily move geographically from one job to another and can respond quickly to monetary incentives. That is, there are no patents or copyrights, "vast amounts" of capital are not necessary to enter the market, and already established firms do not have any lasting cost advantage over new entrants because of experience or size..

Finally, under perfect competition, consumers, resource owner, and firms in the market have perfect knowledge as to present and future prices, costs, and economic opportunities in general. Thus, consumers will not pay a higher price than necessary for the product. Price differences are quickly eliminated, and a single price will prevail throughout the market for the product.

Perfect competition, as defined above, has never really existed. Perhaps the closest we might come today to a perfectly competitive market is the stock market. Another example is the market for such agricultural commodities as wheat, cotton and corn. The natural gas industry and the trucking industries also approach perfect competition. In the milk market, each dairy farmer produces the milk that is essentially identical to that offered by other dairy farmers.

Similarly each milk buyer purchases a small portion of aggregate production, so that he does not receive a cut-rate or volume discount. Because both buyers and sellers can trade as much milk as they want at the going price, both are price-takers and the milk market is said to be perfectly competitive.
The fact that perfect competition in its pure form has never really existed in the real world, does not reduce the usefulness of the perfectly competitive model. A theory must be accepted, or rejected on the basis of its ability to explain and to predict correctly and not on the realism of its assumptions. And the perfectly competitive model does give us some useful explanations and predictions of many real world economic phenomena when the assumptions of the perfectly competitive model are only approximately satisfied.

PRICE DETERMINATION UNDER PERFECT COMPETITION
In Figure 1, the Total cost and total revenue curves of a perfectly competitive firm are shown. The TR curve is a straight line through the origin showing that price is constant at all levels of output. The firm is a price taker and can sell any amount of output at the given market price, with its TR increasing proportionately with its sales. The slope of the TR curve is the MR. It is constant and equal to the prevailing market price, since all units are sold at the same price. Thus P = MR = AR. The shape of the TC curve reflects the U shape of the AC and MC curves. The firm maximizes its profit at Q = $50,000 units, where the distance between the TR and TC is the greatest.
Figure 1

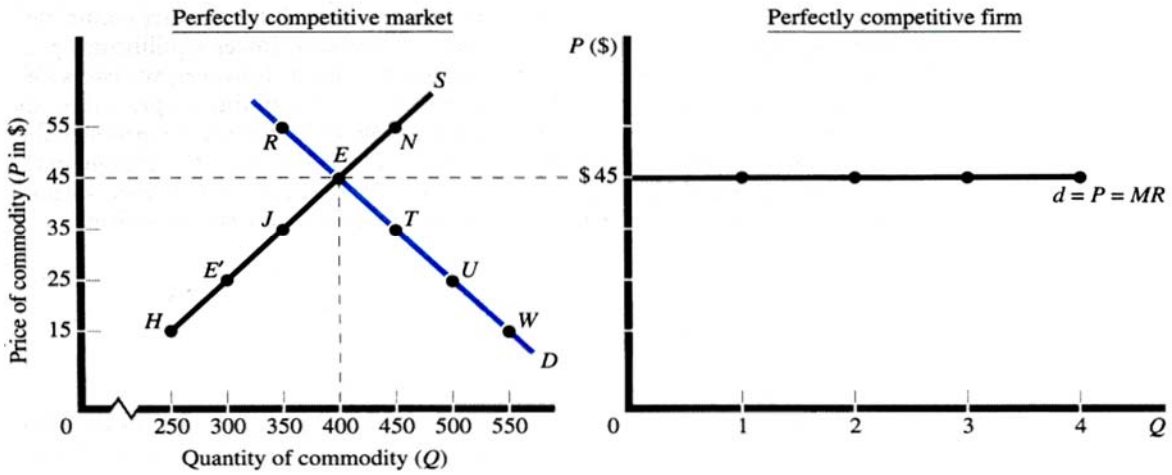


The TR-TC approach is awkward to use when firms are combined together in study of the industry. The alternative approach which is based on marginal cost- marginal revenue, uses price as an explicit variable, and shows clearly the behavioral rule leads to profit maximization.
Under perfect competition, the price of a product is determined at the intersection of the market demand curve and the market supply curve of the product. The market demand curve for product is simply the horizontal summation of the demand curves of all the consumers in the market. The market supply curve of a product is similarly obtained from the horizontal summation of the supply curve of the individual producers of the product.

Given that the market price of a product is determined at the intersection, of the market demand and supply curves of the product, the perfectly competitive firm is a price taker. That is, the perfectly competitive firm takes the price of the product as given and has no perceptible effect on that price by varying its own level of output and sales of the product. Since the products of all firms are homogeneous, a firm cannot sell at a price higher than the market price of the product;

otherwise the firm would lose all its customers. On the other hand, there is no reason for the firm to sell at a price below the market price, since it can sell any quantity of the product at the given market price. As a result, the firm faces a horizontal or infinitely elastic demand curve for the product at the market price determined at the intersection of the market demand and supply curves of the product. For example, a small wheat farmer can sell any amount of wheat at the given market price of wheat. This is shown in Figure 2.

Figure 2
Perfect Competition: Price Determination



Role of Marginal Analysis

- Set Mπ = MR – MC = 0 to maximize profits.
- MR= MC when profits are maximized.

Normal Profit Equilibrium

- There are no economic profits in competitive equilibrium; firms earn a normal rate of return.
- With a horizontal market demand curve, MR = P, so **P = MR = MC = ATC**.

Given the equilibrium price of P = $45, a perfectly competitive firm producing, the product faces the horizontal or infinitely elastic demand curve shown by d at P = $45 in Figure 2. The perfectly competitive firm only determines what quantity of the product to produce at P = $45 in order to maximize its total profits. When the product price is constant, the change in the total revenue per unit change in output or marginal revenue (MR) is also constant and is equal to the product price. That is, for a perfectly competitive firm, **P = MR**

The equilibrium price and quantity can be determined algebraically by setting the market demand and supply functions equal to each other and solving for the equilibrium price. Substituting the equilibrium price into the demand or supply functions and solving for Q, we the get the equilibrium quantity. For example, the equations for the market demand and supply curves for the product in Figure 2 are:

$Q_D = 625 – 5P$
$Q_S = 175 + 5P$
$Q_D = Q_S$
$625 – 5P = 175 + 5P$

Solving for P, we have
450 = 10P
P = $45
Substituting P = $45 into the demand function and solving for Q, we have
Q_D = 625 – 5P = 625 – 5(45) = 400 units

PROFIT MAXIMIZATION WITH CALCULUS
π = TR – TC
dπ/dQ = dTR/dQ - dTC/dQ = 0
so that dTR/dQ = dTC/dQ
Since dTR/dQ = MR and dTC/dQ = MC
The above condition becomes MR = MC. But under perfect competition, the price is given to the firm and is constant.

Therefore,
dTR/dQ = d(PQ)/dQ = P = MR
so that the **FOC** for maximization under perfect competition becomes P = MR = MC.
(The product Rule of differentiation will not apply as under perfect competition P (price) is constant so d(PQ)/dQ = P)
The second order condition for profit maximization requires that the second derivative of Π with respect to Q be negative. That is,
SOC
$d^2 \pi / Dq^2 = d^2 TR / dq^2 - d^2 TC / dq^2 < 0$
$d^2 TR / dq^2 < d^2 TC / dq^2$
So verbally, Slope of MR curve < Slope of MC curve, thus MC must have a steeper slope than the MR curve or the MC curve must cut the MR curve from below.
Under perfect competition, t he slope of MR curve is zero, hence the SOC is simplified as follows :
$0 < d^2 TC / dq^2$
MC must have a +tive slope, or the MC must be rising.

SHORT-RUN ANALYSIS OF A PERFECTLY COMPETITIVE FIRM
In the short run, some inputs are fixed, and these give rise to fixed costs, which go on whether the firm produces or not. Thus, it pays for the firm to stay in business in the short run even if it incurs losses, as long as these losses are smaller than its fixed costs. Thus, the best level of output of the firm in the short run is the one at which the firm maximizes profits or minimizes losses.

Figure 3

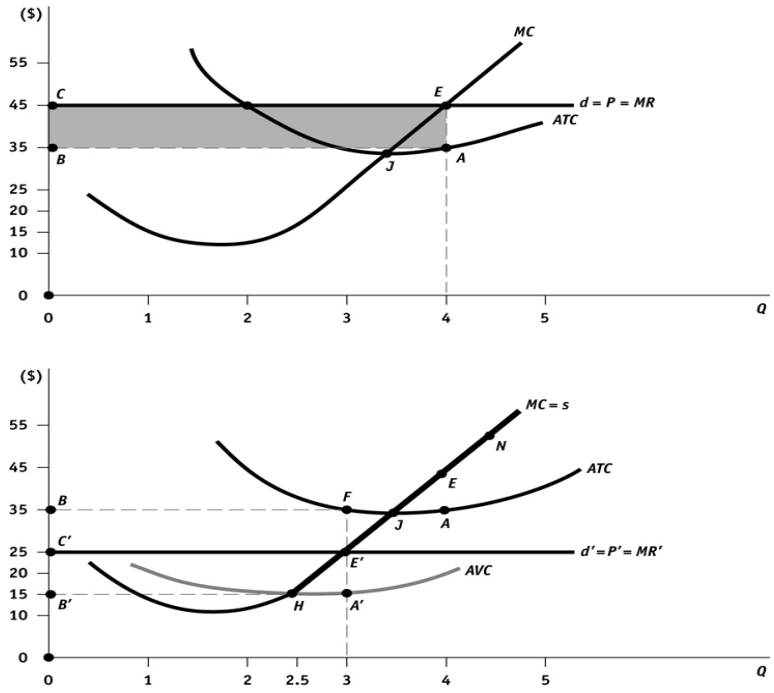


FIGURE 8-2 Short-Run Analysis of a Perfectly Competitive Firm With *d* the best level of output is 4 units and is shown in the top panel by point *E*, at which $P = MR = MC$, and the firm earns profit EA = \$10 per unit, and $EABC$ = \$40 in total. With *d* in the bottom panel, the best level of output is 3 units and is given by point *E′* at which the firm incurs a loss of FE' = \$10 per unit, and $FE'C'B$ = \$30 in total. At point *E′* the firm minimizes losses. The shut-down point is at point *H*. The rising portion of the *MC* curve above the *AVC* curve (shut-down point) is the firm's short-run supply curve (the heavy portion of the *MC* curve in the bottom panel).

The best level of output of the firm in the short run is the one at which the marginal revenue (MR) of the firm equals its short run marginal cost (MC). As long as MR exceeds MC, it pays for the firm to expand output because by doing so the firm would add more to its total revenue than to its total costs (so that its total profits increase or its total losses decrease). On the other hand, as long as MC exceeds MR, it pays for the firm to reduce output because by doing so the firm will reduce its total costs more than its total revenue (so that, once again, its total profits increase or its total losses decrease). Thus, the best level of output of any firm (not just a perfectly competitive firm) is the one at which MR = MC. Since a perfectly competitive firm faces a horizontal or infinitely elastic demand curve, P = MR, so that the condition for the best level of output can be restated as the one at which P = MR = MC. This can be seen in Figure 3.

In the top panel of Figure 3, d is the demand curve for the output of the perfectly competitive firm and the marginal and average total cost (i.e., MC and ATC) curves are also drawn. The best level of output of the firm is given at point E, where the MC curve intersects the firm's d or MR curve. At point E, the firm produces 4 units of output at P = MR = MC = \$45. Since at point E, P = \$45 and ATC = \$35, the firm earns a profit of EA = \$10 per unit and EABC = \$40 in total (the shaded area). This is the largest total profit that the firm can earn. Thus, the best level of output for the firm is $\mathbf{Q_X = 4}$, at which **MR = P = MC** and the total profits of the firm are maximized.

The bottom panel of Figure 3 shows that if the market price of the product is \$25 instead of \$45, so that the demand curve faced by the perfectly competitive firm is d', the best level of output of the firm is 3 units, as indicated by point **E'**, where **P'= MR' = MC** At $\mathbf{Q_X}$ = 3, **P** = \$25 and **ATC** = \$35, so that the firm incurs the loss of FE' = \$10 per unit and FE'C'B = \$30 in total. If the firm

stopped producing the product and left the market, however, it would incur the greater loss of FA' =' $20 per unit and FA'B'B = $60 (its total fixed costs). Another way of looking at this is to say that at the best level of output of Q = 3, the excess of P = $25 over the firm's average variable cost (AVC) of $15 can be applied to cover part of the firm's fixed costs (FA' per unit and FA'B'B in total). Thus, the firm minimizes its losses by continuing to produce its best level of output. If the market price of the product declined to slightly below $15, so that the demand curve facing the firm crossed the MC curve at point H (see the bottom panel of Figure 3), the firm would be indifferent whether to produce or not. The reason is that at point H, P = AVC and the total losses of the firm would be equal to its total fixed costs whether it produce or not. Thus, point H is the shut down point of the firm. Below point H, the firm would not even cover its variable costs, and so by going out of business, the firm would limit its losses to be equal to its total fixed costs.

Lesson 26

COMPETITIVE MARKETS (CONTINUED)

SHORT-RUN SUPPLY CURVE OF THE COMPETITIVE FIRM

The rising portion of the firm's MC curve above the AVC curve or shut-down point is or represents the short run supply curve of the perfectly competitive firm (the heavier portion of the MC curve labeled **s** in the bottom panel of Figure 1. The reason for this is that the perfectly competitive firm always produces where P = MR = MC, as long as P > AVC. Thus, at P = $55, the firm produces 4.5 units (point N); at P = $45, Q = 4; at P = $25, Q = 3; and at P = $15, Q = 2.5. That is, given P, we can determine the output supplied by the perfectly competitive firm by the point where P = MC. Thus, the rising portion of the competitive firm's MC curve above AVC shows a unique relationship between P and Q, which is the definition of the supply curve.

Figure 1

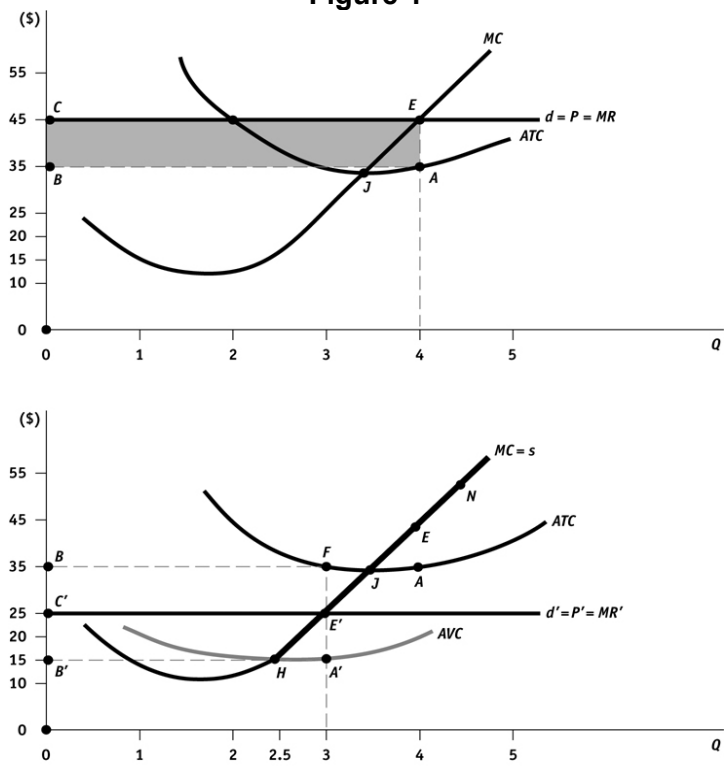


FIGURE 8-2 Short-Run Analysis of a Perfectly Competitive Firm With *d* the best level of output is 4 units and is shown in the top panel by point *E*, at which $P = MR = MC$, and the firm earns profit EA = $10 per unit, and $EABC$ = $40 in total. With *d* in the bottom panel, the best level of output is 3 units and is given by point *E′* at which the firm incurs a loss of FE' = $10 per unit, and $FE'C'B$ = $30 in total. At point *E′* the firm minimizes losses. The shut-down point is at point *H*. The rising portion of the *MC* curve above the *AVC* curve (shut-down point) is the firm's short-run supply curve (the heavy portion of the *MC* curve in the bottom panel).

SHORT-RUN FIRM SUPPLY CURVE

Example

A competitive firm **short-run supply curve** is the marginal cost curve, so long as **P > AVC.**

Given:

$TC = 361{,}250 + 5Q + 0.0002Q^2$

TR = 25Q

MR = MC

25 = 5 + 0.0004Q

Q = 50,000

Since the firm is price taker: P = MR = $25

π = TR – TC

$= 25(50{,}000) – 361{,}250 – 5(50{,}000) – 0.0002(50{,}000)^2$

Π = $138,750

MC = ATC

$$5 + 0.0004Q = \frac{361{,}250 + 5Q + 0.0002Q^2}{Q}$$

Q = 42,500

It is clear that the point of minimum AVC is reached in the output range between 15,000 to 20,000, where AVC = 50,000 + 5Q + $0.0002Q^2$. The minimum point on the AVC occurs at Q = 15,811 and AVC = $11.32, and is found by setting MC = AVC and solving for Q:

P = $25

MC = AVC

$$5 + 0.0004Q = \frac{50{,}000 + 5Q + 0.0002Q^2}{Q}$$

Q = 15,811

LONG-RUN ANALYSIS OF A PERFECTLY COMPETITIVE FIRM

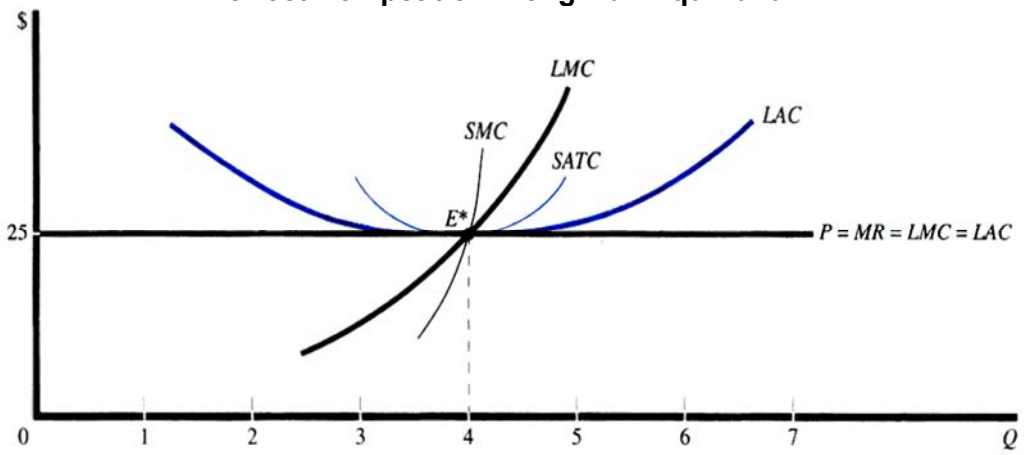
In the long run all inputs and costs of production are variable, and the firm can construct the optimum or most appropriate scale of plant to produce the best level of output. The best level of output is the one at which price equals the long-run marginal cost (LMC) of the firm. The optimum scale of plant is the one with the short run average total cost (SATC) curve tangent to the long-run average cost of the firm at the best level of output.

On one hand, if existing firms earn profits, more firms enter the market in the long run. This increases (i.e., shifts to the right) the market supply of the product and results in a lower product price until all profits are squeezed out. On the other hand, if firms in the market incur losses, some firms will leave the market in, the long run. This reduces the market supply of the product until all firms remaining in the market just break even. Thus, when a competitive market is in long run equilibrium all firms produce at the lowest point on their long-run average cost (LAC) curve and break even. This is shown by point E* in Figure 2.

Figure 2 shows that at p = $25, the best level of output of the perfectly competitive firm is 4 units and is given by point E*, at which P = LAC. Because of free or easy entry into the market, all profits and losses have been eliminated, so that P = LMC = lowest LAC. Thus, for a competitive market to be in long run equilibrium, all firms in the industry must produce where P = MR = LMC = lowest LAC so that all firms break even. The perfectly competitive firm operates the scale of plant represented by SATC at its lowest point (point E*), so that its short run marginal cost (SMC) equals LMC also.

Figure 2

Perfect Competition: Long-Run Equilibrium



MARGINAL COST AND FIRM SUPPLY

Short-run Firm Supply

- Competitive market price (P) is shown as a horizontal line because P=MR.
- Marginal cost schedule is the short-run supply curve so long as P > AVC.

Long-run Firm Supply

- Profit. P=MR = MC
- Marginal cost curve is the long-run supply curve so long as P > ATC.

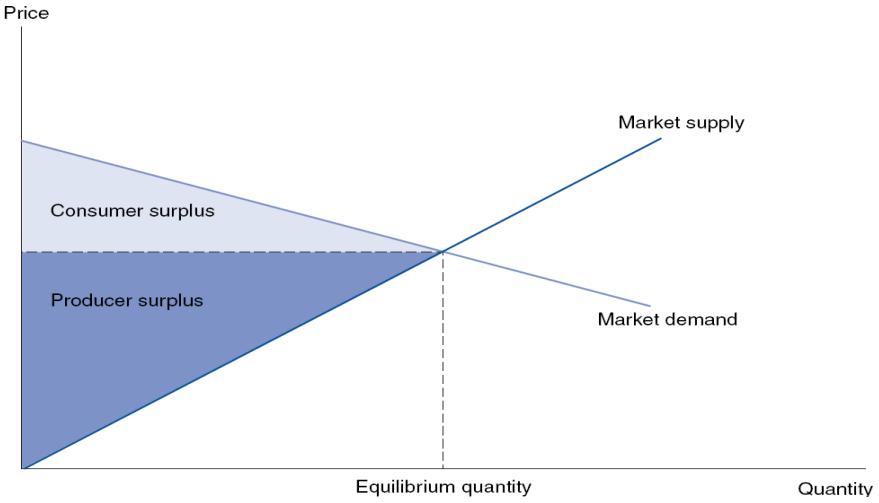
Perfect Competition and Efficiency

The model of perfect competition achieves efficiency in two ways:

- Allocative efficiency: Price = MC and therefore consumer and producer surpluses are maximised
- Productive efficiency: In the long run in perfect competition equilibrium output is produced where average costs are at their lowest point

Welfare economics is the study of how the allocation of economic resources affects the material well-being of consumers and producers. Competitive markets creates balance between supply and demand, also maximize the total social welfare of the society derived from such activity. The measurement of social welfare is closely related to consumer and producer surpluses. Consumer surplus is the amount that consumers are willing to pay for a given good or service minus the amount that they are required to pay. Producer surplus is the net benefit derived by producers from production. Figure 3 shows consumer and producer surpluses.In long run, firm must cover all necessary costs of production and earn a normal.

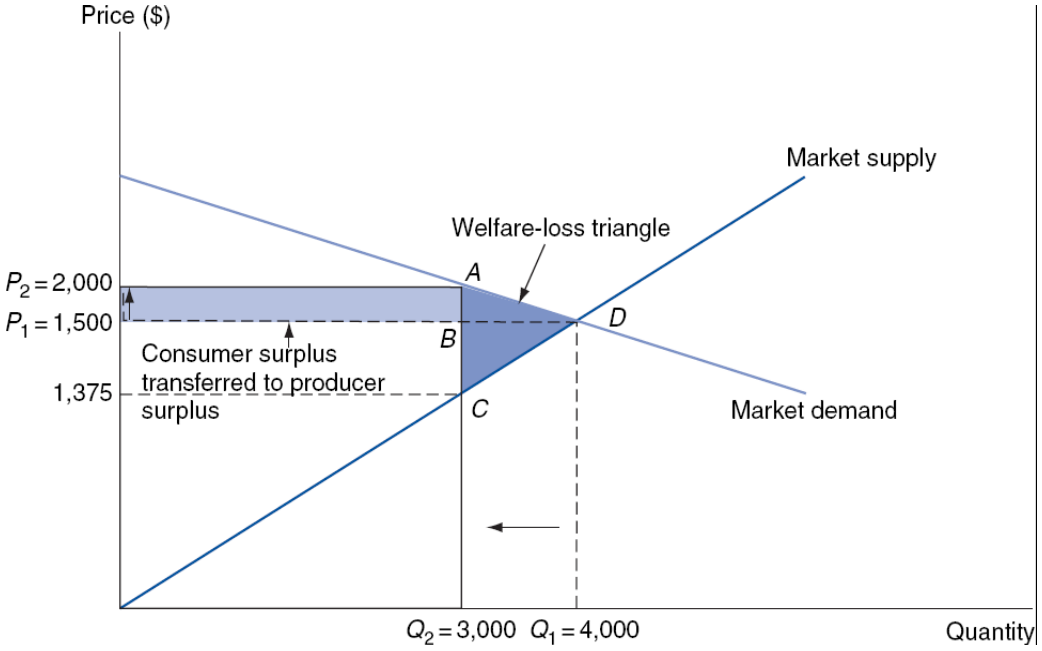
Figure 3



DEADWEIGHT LOSS PROBLEM

There is deadweight loss problem associated with deviations from competitive market equilibrium. A deadweight loss is any loss suffered by consumers or producers that is not transferred but is instead simply lost due to market imperfections or government policies. Because deadweight losses are shown as triangles (shown in Figure 4) when linear demand and supply curves are used, deadweight losses are often described as welfare loss triangles. Deadweight losses occur when market imperfections reduce transaction volume.

Figure 4



DEADWEIGHT LOSS ILLUSTRATION

Supply	**Q = -8000 + 8P**
or	P = 1000 + 0.125Q
Demand	**Q = 7000 – 2P**
or	P = 3500 – 0.5Q

Supply = Demand
-8000 + 8P = 7000 – 2P
10P = 15000
P = $1,500 Per ton
Supply = Demand
1000 + 0.125Q = 3500 – 0.5Q
0.625Q = 2500
Q = 4,000 (000) tons

Total Deadweight Loss

Consumer Deadweight Loss
= ½ [(4000 – 3000) * (2000 – 1500)]
= $250,000 (000)
Producer Deadweight Loss
= ½ [(4000 – 3000) * (1500 – 1375)]
= $62,500 (000)
Total Deadweight Loss = Consumer Loss + Producer Loss
= $250,000 (000) + $62,500 (000)
= $312,500 (000)

Lesson 27

MONOPOLY

Perfect monopoly is the opposite extreme of perfect competition. Monopoly exists when a single firm is the sole producer of a good that has no close substitutes. In other words, there is a single firm in the industry. Perfect monopoly, like perfect competition, is quite rare in the real world.

CHARACTERISTICS OF MONOPOLY MARKETS

Monopoly exists when an individual producer has the ability to set market prices. Monopoly firms are price makers, not price takers.

- A single seller.
- Unique product.
- Blockaded entry and/or exit.
- Imperfect dissemination of information.
- Opportunity for long-run economic profits.

Classic examples include electricity utilities, gas, sanitary services, transportation and telecommunication services.

SOURCES OF MONOPOLY

Monopoly is the form of market organization in which a single firm sells a product for which there are no close substitutes. Thus, the monopolist represents the market and faces the market negatively sloped demand curve for the product. As opposed to a perfectly competitive firm, a monopolist can earn profits in the long run because entry into the industry is essentially blocked. Thus monopoly is at the opposite extreme from perfect competition in the range of market organizations.

There are four basic reasons that can give rise to monopoly. First, the firm may control the entire supply of raw materials required to produce the product. Second, the firm may own a patent or copyright that precludes other firms from using a particular production process or producing the same product. For example, Xerox had a monopoly on copying machines and Polaroid on instant cameras. Patents are granted by the government for a period of 17 years as an incentive to inventors in the US while in Pakistan it is granted for a period of 20 years.

Third, in some industries, economies of scale may operate over a sufficiently large range of outputs as to leave only one firm supplying the entire market. Such a firm is called a natural monopoly. Examples of these are public utilities (electrical, gas, water, and local transportation and telecommunication companies). To have more than one such firm in a given market would lead to duplication of supply lines and to much higher costs per unit. To avoid this, governments usually allow a single firm to operate in the market but regulate the price of the services provided, so as to allow the firm only a normal return on investment.

Fourth, a monopoly may be established by a government franchise. In this case, the firm is set up as the sole producer and distributor of a product or service but is subjected to governmental regulation. The best example of a monopoly established by government franchise is the post office, railways, airlines. Local governments also require a license to operate many types of businesses, such as taxis, broadcasting, TV channels, medical units, and private health care clinics.

MONOPOLY PRICE/OUTPUT DECISIONS

Under monopoly, the industry demand curve is identical to the firm demand curve. Because industry demand curves slope downward, monopolists also face a downward-sloping demand curve. The monopolist can set either price or quantity, but not both. Given one, the value of the other is determined along the demand curve.

MONOPOLY OUTPUT RULE

A monopoly uses the same profit-maximization rule as does any other firm: It operates at the output level at which marginal revenue equals marginal cost. A profit-maximizing monopolist should produce the output, Q, such that marginal revenue equals marginal cost:

MR = MC

Because the demand (average revenue) curve is negatively sloped and hence declining, the marginal revenue curve must lie below it. When a monopoly equates marginal revenue and marginal cost, it simultaneously determines the output level and the market price for its product. Profit maximization always require that firms operate at the output level at which MR = MC.

MONOPOLY PRICING RULE

Given the level of output, Q, that maximizes profits, the monopoly price is the price on the demand curve corresponding to the Q units produced:

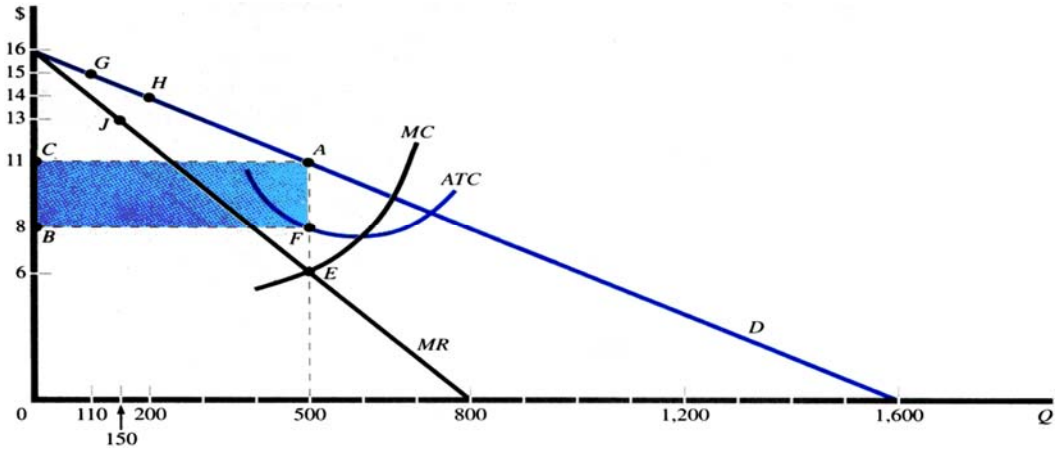
p = p (Q)

Given a downward sloping monopoly demand curve, price always exceeds marginal revenue under monopoly. This is due to the fact that price is average revenue, and a downward-sloping demand curve requires that marginal revenue be less than average revenue. In a competitive market, P = MR = MC = AC in long-run equilibrium. In monopoly markets, profit-maximization requires MR = MC, but barrier to entry make above-normal profits possible, and P > AC in long-run equilibrium.

SHORT-RUN PRICE AND OUTPUT DETERMINATION UNDER MONOPOLY

In Figure 1, D is the market demand curve faced by the monopolist, and MR is the corresponding marginal revenue curve. Profit maximizing output level is Q = 500 units, at this level of output AR = P = $11; AC = $8. At this level the monopolist earns a profit of AF = $3 per unit and AFBC = $1,500 in total (the shaded area in Figure 1)

Figure 1



Since MR = MC =$6
So that P = AR > MR, since P > MR, P is also > MC.
As MC cuts MR from below i.e. beyond the point E, MC > MR and since P = AR > MR, P > MC. While the monopolist of Figure 1 is earning short run profit a monopolist (just like a perfect competitor) could also break even or incur losses in the short run. It all depends on the height of the ATC at the best level of output. If ATC = P at best level of output, the monopolist breaks even, and if ATC > P at the best level of output, the monopolist incur a loss. Again, as in the case of perfect competition, it pays for a monopolist to remain in business in the short run even if it incurs losses, as long as P > AVC.

Assume that the monopoly firm total revenue and total cost functions are:

π = TR – TC
We can re-write this as:
π = (TR/Q – TC/Q) * Q
π = (P – AC) * Q
π = ($37.50 - $22.23) * 50,000
= $15.27* 50,000
= $763,750

Finding the marginal revenue and marginal cost functions from the given TR and TC functions, we get:
$M\pi$ = MR – MC = 0

MR = MC

P = AR > MR

$TR = 50Q – 0.00025Q^2$

MR = 50 – 0.0005Q

$TC = 361,250 + 5Q + 0.0002Q^2$

MC = 5 + 0.0004Q

The optimal price/output combination can be determined by setting marginal revenue equal to marginal cost and solving for Q:

MR = MC

50 – 0.0005Q = 5 + 0.0004Q

Q = 50,000

At this output level, maximum economic profits are:

π = TR – TC

$= 50Q – 0.00025Q^2 – (361,250 + 5Q + 0.0002Q^2)$

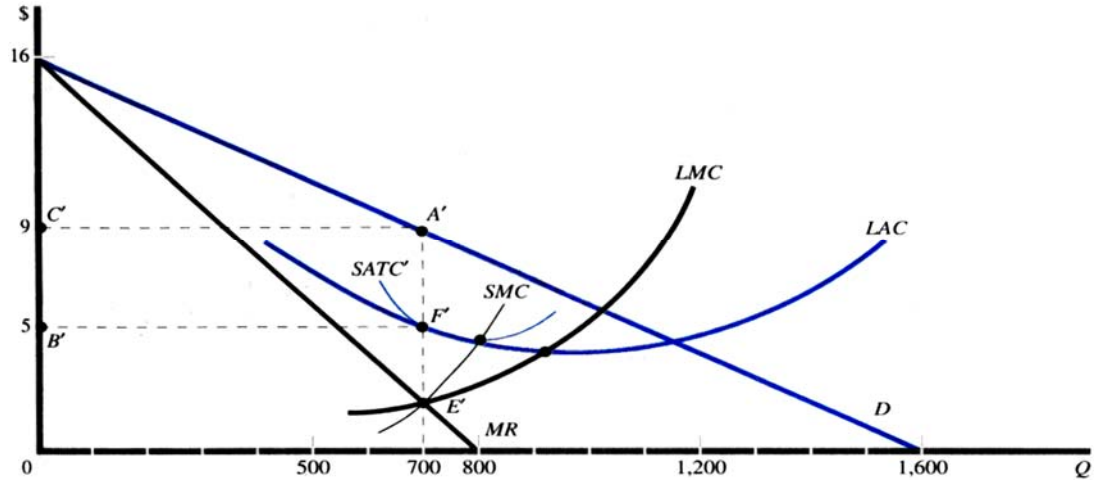
= $763,750

LONG-RUN PRICE AND OUTPUT DETERMINATION UNDER MONOPOLY

In the long run, all inputs and costs of production are variable, and the monopolist can construct the optimal scale of plant to produce the best level of output. As in the case of perfect competition, the best level of output of the monopolist is given at the point at which P = LMC, and the optimum scale of plant is the one with the SATC curve tangent to the LAC curve at the best level of output. As contrasted with perfect competition, however, entrance into the market is blocked under monopoly, and so the monopolist can earn economic profits in the long run. Because of blocked entry, the monopolist is also not likely to produce at the lowest point on its LAC curve. This is shown in Figure 2.

Figure 2 shows that the best level of output for the monopolist in the long run is 700 units and is given by point E', at which MR = LMC. At Q = 700, P = $9 (point A' on the D curve). The

Figure 2



monopolist has had time in the long run to build the optimum scale of plant given by the SATC curve tangent to the LAC curve at Q = 700 (point F' in Figure 2). Operating the optimum scale of plant at F' at the best level of output of Q = 700, the monopolist has SATC = LAC = $5 (point F'). Thus, the monopolist is earning a long run profit of A'F' = $4 per unit and A'F'B'C' = $2,800 in total (as compared to $1,500 in the short run). Because entry into the market is blocked, the monopolist will continue to earn these profits in the long run as long as his demand and cost conditions remain unchanged.

SOCIAL COSTS OF MONOPOLY

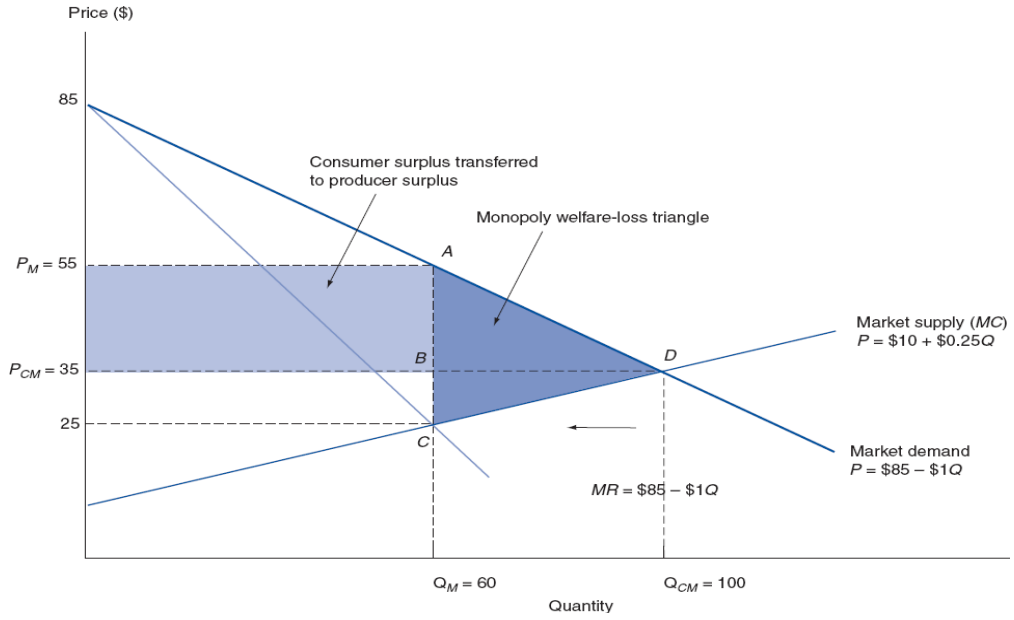
Monopolies have an incentive to under produce and earn economic profits. Underproduction results when a monopoly curtails output to a level at which the value of resources employed, as measured by the marginal cost of production, is less than the social benefit derived, where social benefit is measured by the price that customers are willing to pay for additional output. Under monopoly, marginal cost is less than price at the profit-maximizing output level. Although resulting economic profits serve the useful functions of providing incentives and helping allocate resources, it is difficult to justify above-normal profits that result from market power rather than from exceptional performance.

DEADWEIGHT LOSS FROM MONOPOLY

- Monopoly markets create a loss in social welfare due to the decline in mutually beneficial trade activity.

- There is also a wealth transfer problem associated with monopoly; consumer surplus is transferred to producer surplus as shown in Figure 3.

Figure 3



DEADWEIGHT LOSS FROM MONOPOLY (EXAMPLE)

Given Q_D and Q_S:

$Q_S = -40 + 4p$

or, solving for price,

$4p = 40 + Q_S$

$p = 10 + 0.25\ Q_S$

$Q_D = 170 - 2p$

or, solving for price,

$2p = 170 - Q_D$

$p = 85 – 0.5\ Q_D$

Competitive Market

Supply = Demand

$-40 + 4p = 170 - 2p$

$6p = 210$

p = $35/month

Supply = Demand

$10 + 0.25\ Q = 85 – 0.5\ Q$

$0.75Q = 75$

Q = 100 (millions) Customers

Monopoly

MR = MC

$85 – Q = 10 + 0.25Q$

$1.25Q = 75$

Q = 60 (million)

At Q = 60

P = 85 – 0.5 Q
P = 85 – 0.5(60)
P = $55/ month

The Area of ABD = Consumer Deadweight Loss
= ½[(100 – 60) * (55 – 35)]
= $400(million) per month
The Area of BCD = Producer Deadweight Loss
= ½[(100 – 60) * (25 – 35)]
= $200(million) per month
Total Deadweight Loss = C's + P's = $600(million)/ month
Area of Rectangle = Transfer to P'S = 60 * (55 – 35)
= $1,200(million) per month

Lesson 28

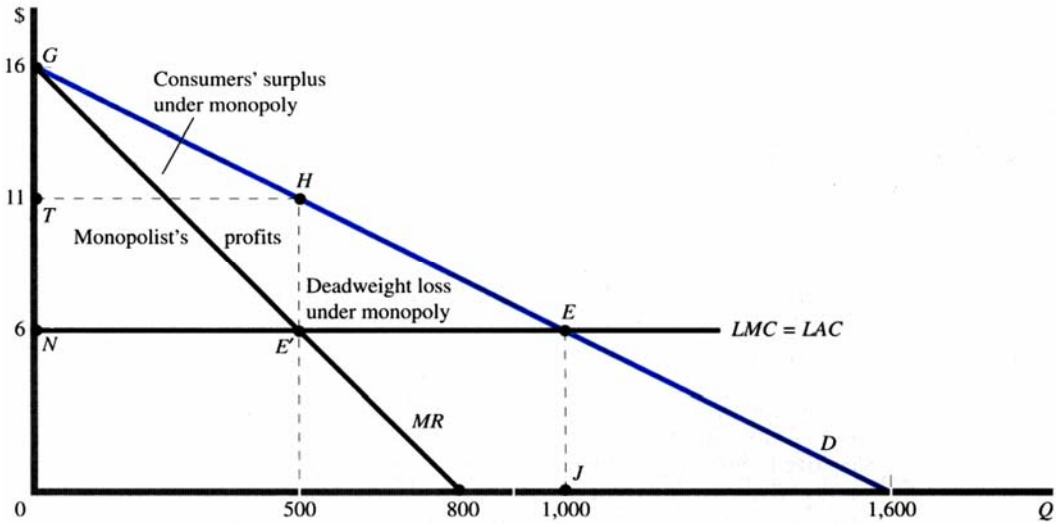
MONOPOLY / MONOPOLISTIC COMPETITION

SOCIAL COST OF MONOPOLY

Patents

One of the major sources of monopoly is that the firm might own a patent. The ownership of a patent or copyright that precludes other firms from using a particular production process or producing the same product. Intellectual Property (IP) is critical for competitive economy in the back drop of ongoing globalization. Sustainable economic growth now depends largely on Hi-tech R&D base and efficient knowledge input. The new concept of IP based nation is gaining ground because it is Intellectual Property which enables technology creation and technology transfer by providing the necessary enabling environment. For these considerations Intellectual Property was mainstreamed in Pakistan in 2005. A patent for an invention is grant of exclusive rights to make, use and sell the invention for a limited period of 20 years. The patent grant excludes others from making, using, or selling the invention. Patent protection does not start until the actual grant of a patent. Patent is granted when the application for it is submitted to IPO (Intellectual Property Rights Organization of Pakistan).

Figure 1



In case of competitive markets, we have observed that price and quantity that creates balance between supply and demand also maximize the total social welfare derived from such activity. We have measured the social welfare of the society by consumers' surplus and producers' surplus. For example in case of the markets for drugs when a patent gives a firm a monopoly over the sale of a drug, the firm charges the monopoly price, (point T in Figure 1, price during patent life) which is well above the marginal cost of making the drug. When the patent on a drug runs out, (Point N, price after patent expires) new firms enter the market, making it more competitive. As a result, price falls from the monopoly price to marginal cost.

SOCIAL BENEFITS OF MONOPOLY

Economies of Scale

- In natural monopoly, LRAC declines continuously and one firm is most efficient.
- Some real-world monopolies are government-created or government-maintained

Dilemma of Natural Monopoly

- Monopoly has the potential for efficiency.
- Unregulated monopoly can lead to economic profits and underproduction.

Natural monopoly presents something of a dilemma. On the one hand, economic efficiency could be enhanced by restricting the number of producers to a single firm. On the other hand, monopolies have an incentive to under produce and can generate unwarranted economic profits. A very large scale of operation is often required to produce most products efficiently, and this is possible when only a few firms are operating. For example, economies of scale operate over such a large range of outputs that steel, aluminum, automobiles, mainframe computers, aircraft, and many other products and services can be produced efficiently only by very large firms, so that a handful of such firms can meet the entire market demand for the product or service. Perfect competition under such conditions would either be impossible or lead to extremely high production costs. One could only imagine how high the cost per unit would be if automobiles were produced by 100 or more firms instead of by three or four very large firms.

Figure 2

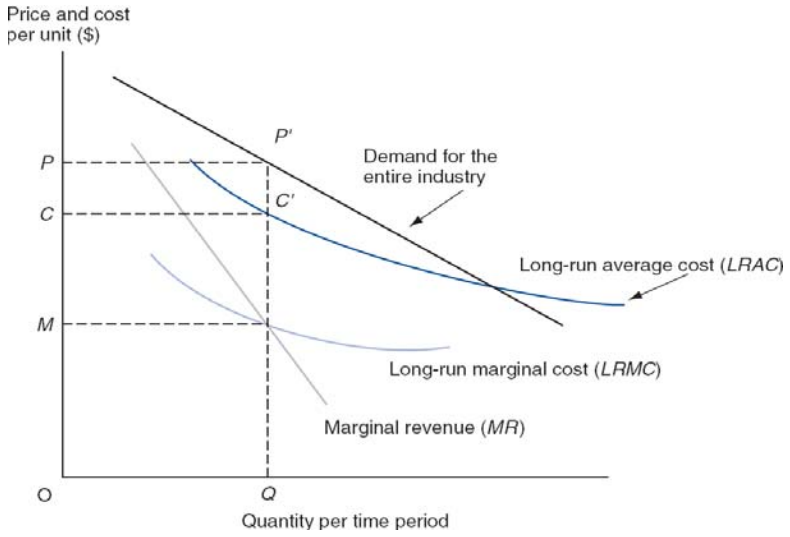


Figure 2 shows that without regulation, natural monopolies would charge quite high prices (P') and produce too little output (OQ).

MONOPOLISTIC COMPETITION

A market structure that lies between the extreme of monopoly and perfect competition is *monopolistic competition*. The partly competitive, partly monopolistic market structure faced by the firms in the clothing, food, hotel, retailing, and consumer products industries is called monopolistic competition. Given the lack of perfect substitutes, monopolistically competitive firms have some discretion in setting prices—they are not price takers. However, given fierce competition from imitators offering close but not identical substitutes, such firms enjoy only a normal rate of return on investment in long-run equilibrium.

Monopolistic competition is similar to perfect competition in that there is large number of sellers this two market structure model. The major difference between these two market structure models is that consumers perceive important differences among the products offered by monopolistically competitive firms, whereas the output of perfectly competitive firms is

homogeneous. This gives monopolistically competitive firms at least some discretion in setting prices.

MONOPOLISTIC COMPETITION CHARACTERISTICS

- Many buyers and sellers.
- Product heterogeneity.
- Free entry and exit.
- Perfect information.
- Opportunity for normal profits in long-run equilibrium.

MEANING AND IMPORTANCE OF MONOPOLIST COMPETITION

- Monopolistic Competition Characteristics
  - Many buyers and sellers.
  - Product heterogeneity.
  - Free entry and exit.
  - Perfect information.
  - Opportunity for normal profits in long-run equilibrium.

Monopolistic competition is defined as the form of market organization in which there are many sellers of a differentiated product and entry into and exit from the industry are rather easy in the long run. Differentiated products are those that are similar but not identical and satisfy the same basic need. Examples are the numerous brands of breakfast cereals, toothpaste, cigarettes, detergents, cold medicines and cosmetics. The differentiation may be real (for example, the various breakfast cereals may have greatly different nutritional and sugar contents) or imaginary (for example, all brands of aspirin contain the same basic ingredients).

As the name implies, monopolistic competition is a, blend of competition and monopoly. The competitive element results from the fact that in a monopolistically competitive market (as in a perfectly competitive market), there are many sellers of the differentiated product, each too small to affect others. The monopoly element arises from product differentiation (i.e., from the fact that the product sold by each seller is somewhat different from the product sold by any other seller). The resulting monopoly power is severely limited, however, by the availability of many close substitutes. Thus, if the seller of a particular brand of aspirin increased its price even moderately, it would expect to lose a great deal of its sales.

Monopolistic competition is most common in the retail and service sectors of our economy. Clothing, cotton textiles, bakers and food processing are the industries that come close to monopolistic competition at the national level. At the local level, the best examples of monopolistic competition are fast-food outlets, shoe stores, gasoline stations, beauty salons, drugstores, video rental stores, and pizza parlors.

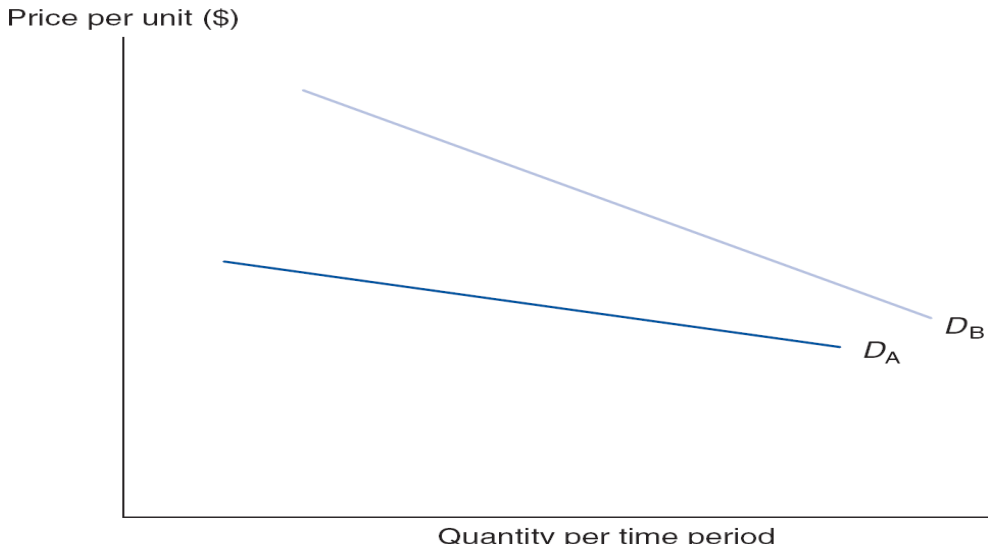
Since each firm sells a somewhat different product under monopolistic competition, we cannot derive the market demand curve and the market supply curve of the product as we did under perfect competition and we do not have a single equilibrium price for the differentiated products but a cluster of prices. Our analysis must, therefore, necessarily be restricted to that of the "typical" firm: The graphical analysis will also be greatly simplified by assuming (with Edward Chamberlin, the originator of the monopolistically competitive model) that all firms selling similar products face identical demand and cost curves. This is unrealistic because the production of differentiated products is likely to lead to somewhat different demand and cost curves. Making such an assumption, however, will greatly simplify the analysis.

As in perfectly competitive markets, a large number of competitors make independent decisions in monopolistically competitive markets. A price change by any one firm does not cause other firms to change prices. If price reactions did occur, then an oligopoly market structure would be present. The most distinctive characteristic of monopolistic competition is that each competitor offers a unique product that is an imperfect substitute for those offered by rivals.

RELATION BETWEEN PRODUCT DIFFERENTIATION AND ELASTICITY OF DEMAND

The effect of product differentiation is to create downward-sloping firm demand curves in monopolistically competitive markets. Unlike a price taker facing a perfectly horizontal demand curve, the firm is able to independently determine an optimal price/output combination. The degree of price flexibility enjoyed depends on the strength of product differentiation. The more differentiated a firm's product, the lower the substitutability of other products for it. Strong differentiation results in greater consumer loyalty and greater control over price. This is illustrated in Figure 3, which shows the demand curves of firms A and B. Consumers view firm A's product as being only slightly differentiated from the entire industry's output. Because many other firms offer acceptable substitutes, firm A is close to being a price taker. On the other hand, firm B has successfully differentiated its product, and consumers are therefore less willing to accept substitutes for B's output. Firm B's demand is relatively less sensitive to price changes. Weaker product differentiation implies (Case of firm A's demand) high sensitivity to price changes.

Figure 3



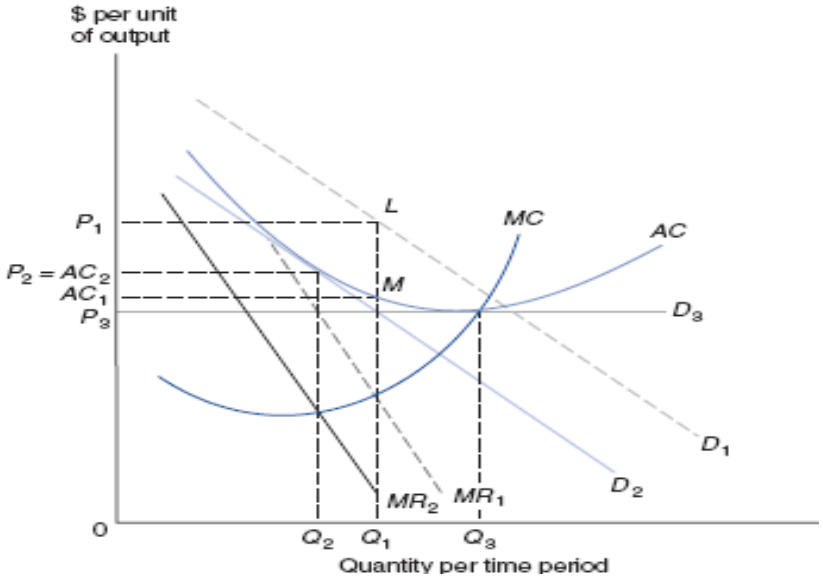
MONOPOLIST COMPETITION PRICE/OUTPUT DECISIONS

As its name suggests, monopolistic competition combines elements of both monopoly and perfect competition. The monopoly aspect is most forcefully observed in the short run. For example, consider Figure 4. With the demand curve, D1, and its related marginal revenue curve, MR1, the optimum output, Q1, is found at the point where MR1 = MC. Short-run monopoly profits equal to the area P1LMAC1 are earned. Such profits can be derived from new product introductions, product and process improvements, creative packaging and marketing, or other factors such as an unexpected rise in demand.

Over time, short-run monopoly profits attract competition, and other firms enter the industry. This competitive aspect of monopolistic competition is seen most forcefully in the long run. As

competitors emerge to offer close but imperfect substitutes, the market share and profits of the initial innovating firm diminish. Firm demand and marginal revenue curves shift to the left as, for example, from D1 to D2 and from MR1 to MR2 in Figure 4. Optimal long-run output occurs at Q2, the point where MR2 = MC. Because the optimal price P2 equals ATC2, where cost includes a normal profit just sufficient to maintain capital investment, economic profits are zero. The price/output combination (P2Q2) describes a monopolistically competitive market equilibrium characterized by a high degree of product differentiation. If new entrants offered perfect rather than close substitutes, each firm's long-run demand curve would become more nearly horizontal, and the perfectly competitive equilibrium, D3 with P3 and Q3, would be approached. Like the (P2Q2) high-differentiation equilibrium, the (P3Q3) no-differentiation equilibrium is something of an extreme case. In most instances, competitor entry reduces but does not eliminate product differentiation. An intermediate price/output solution, one between (P2Q2) and (P3Q3), is often achieved in long-run equilibrium. Indeed, it is the retention of at least some degree of product differentiation that distinguishes the monopolistically competitive equilibrium from that achieved in perfectly competitive markets.

Figure 4



A firm will never operate at the minimum point on its average cost curve in monopolistically competitive equilibrium. Each firm's demand curve is downward sloping and is tangent to the ATC curve at some point above minimum ATC. However, this does not mean that a monopolistically competitive industry is inefficient. The very existence of a downward-sloping demand curve implies that consumers value an individual firm's products more highly than they do products of other producers. The higher prices and costs of monopolistically competitive industries, as opposed to perfectly competitive industries, reflect the economic cost of product variety. If consumers are willing to bear such costs, then such costs must not be excessive. The success of branded products in the face of generic competition, for example, is powerful evidence of consumer preferences for product variety.

MONOPOLISTIC COMPETITION PROCESS

Short-run Monopoly Equilibrium

- Monopolistically competitive firms take full advantage of short-run monopoly.
- In short run, MR = MC, P > AC, and $\pi > 0$.

Long-run High-price/Low-output Equilibrium

- With differentiated products, MR = MC and P = AR = AC at a point above minimum LRAC.
- No excess profits exist, so π = 0.

Long-run Low-price/High-output Equilibrium

- With homogenous products, MR = MC and P=AC at minimum LRAC.
- No excess profits exist, so π = 0. This is competitive market equilibrium.

SHORT-RUN MONOPOLY EQUILIBRIUM (EXAMPLE)

Given $TR = 20{,}000Q - 15.6Q^2$

$TC = 400{,}000 + 4640Q + 10Q^2$

$MR = 20{,}000 - 31.2Q$

$MC = 4640 + 20Q$

$P = AR = 20{,}000 - 15.6Q$

$P = 20{,}000 - 15.6(300)$

P = $15,320

$\pi = TR - TC$

$= 20{,}000Q - 15.6(300)^2 - 400{,}000 - 4640(300)Q - 10(300)^2$

$= - 361{,}250 + 45Q - 0.00045Q^2$

π = $1,904,000 or

= $1.9 million

MR = MC

$20{,}000 - 31.2Q = 4640 + 20Q$

$51.2Q = 15{,}360$

Q = 300 units

$P = AR = 20{,}000 - 15.6Q$

$P = 20{,}000 - 15.6(300)$

P = $15,320

$\pi = TR - TC$

$= 20{,}000Q - 15.6(300)^2 - 400{,}000 - 4640(300)Q - 10(300)^2$

$= - 361{,}250 + 45Q - 0.00045Q^2$

π = $1,904,000 or

= $1.9 million

Therefore, the financial planning committee should recommend a $15,320 price and 300-unit output level to maximize short run profits.

LONG-RUN HIGH-PRICE/LOW-OUTPUT EQUILIBRIUM

$AC = TC/Q = 400{,}000/Q + 4640Q/Q + 10Q^2 /Q$

$= 400{,}000\ Q^{-1} + 4640 + 10Q$, the slope of this AC curve is given by the expression:

$AC = -400{,}000\ Q^{-2} + 10$

The slope of the new demand curve is given by:

= - 15.6 (same as the original D-curve)

Slope of AC curve = Slope of Demand curve

$-400{,}000\ Q^{-2} + 10 = - 15.6$

Q = 125 Units

= 400,000/125 + 4640 + 10(125) = $9,090

$\pi = P * Q - TC$
$= 9{,}090(125) - 400{,}000 - 4{,}640(125) - 10(125)^2$
$= \$0$

LONG-RUN LOW-PRICE/HIGH-OUTPUT EQUILIBRIUM

The low-price/high-output (perfectly competitive) equilibrium combination occurs at the point where *P* = *MR* = *MC* = *AC*. This reflects that the firm's demand curve is perfectly horizontal, and average costs are minimized. To find the output level of minimum average costs.

Set *MC* = *AC* and solve for *Q*:

$MC = AC$
$1{,}640 + 20Q = 400{,}000\ Q^{-1} + 4640 + 10Q$
$Q^2 = 40{,}000$
$Q = \sqrt{40{,}000}$
200 Units
$P = AC$
$400{,}000/200 + 4{,}640 + 10(200)$
$= \$8{,}640$

Under this low-price equilibrium scenario, ABC monopoly price falls in the long run from an original $15,320 to $8,640, and output falls from the monopoly level of 300 units to the competitive equilibrium level of 200 units per month. The company would earn only a risk adjusted normal rate of return, and economic profits would equal zero.

$\pi = P * Q - TC$
$= 8{,}640(200) - 400{,}000 - 4{,}640(200) - 10(200)^2$
$= \$0$

Following the onset of competition, the firm XYZ's will reduce its output from 300 units/month to a level between
Q = 125 and Q = 200 units/month.
S-Run profit-Max price = $15,320 will fall between
P = $ 9,090 (High-price/Low-output Equilibrium)
And P = $8,640 (Low-price/High-output Equilibrium)

Lesson 29

OLIGOPOLY

OLIGOPOLY: BASIC CHARACTERISTICS AND SOURCES

Oligopoly is the kind of market structure in which there are few sellers (usually less than 10) of a homogenous or differentiated product. If there are only two sellers, we have duopoly. If there are three sellers, we have Triopoly. If the product is homogenous, it is called pure oligopoly. If the product is differentiated, it is called differentiated oligopoly.

BASIC FEATURES OF OLIGOPOLY ARE:

- Few sellers.
- Homogenous or unique products.
- Blockaded entry and exit.
- Imperfect dissemination of information.
- Opportunity for above-normal (economic) profits in long-run equilibrium.

SOURCES OF OLIGOPOLY ARE:

The sources of oligopoly are generally the same as for monopoly.
(1) Economies of scale, may operate over a sufficiently large range of outputs as to leave only a few firms supplying the entire market
(2) Huge capital investments and specialized inputs are usually required to enter an oligopolistic industry (say, automobiles, aluminum and similar industries), and this acts as an important natural barrier to entry
(3) A few firms may own a patent for the exclusive right to produce a commodity or to use a particular production process
(4) Established firms may have a loyal following of customers based on product quality and service (brands) that new firms would find very difficult to match
(5) A few firms may own or control the entire supply of a raw material required in the production of the product
(6) The government may give a franchise to only a few firms to operate in the market.

The above points are not only the sources of oligopoly but also represent the barriers to other firms entering the market in the long run. If entry were not so restricted, the industry could not remain oligopolistic in the long run. A further barrier to entry is provided by limit pricing, whereby existing firms charge a price low enough to discourage entry into the industry. By doing so, they voluntarily sacrifice short-run profits in order to maximize long-run profits.

ROLE OF STRATEGIC INTERDEPENDENCE

Since there are a few firms selling a homogenous or differentiated product in oligopolistic markets, the action of each firm affects the other firms in the industry. That is, your actions affect the profits of your rivals and your rivals' actions affect your profits. Having said that, it is clear that the distinguishing characteristic of oligopoly is the interdependence or rivalry among firms in the industry. This is the natural result of fewness. Each firm must take into account the expected reaction of other firms.

Since an oligopolist knows that its own actions will have a significant impact on the other oligopolist in the industry, each oligopolist must consider the possible reaction of competitors in deciding its pricing policies, the degree of product differentiation to introduce, the level of advertising to undertake, the amount of service to provide, and so on. Since competitors can react in many ways, we do not have a single oligopoly model but many, each based on the particular behavioral response of competitors to the actions of the first. We must keep in mind,

however, that each model is, at, best, incomplete and more or less unrealistic. So we can say that no single general model of oligopoly behavior exists.

Because of this interdependence, managerial decision making is much more complex under oligopoly than under any other forms of market structure. For example, if firm A and firm B sell differentiated products. How does the quantity demanded for firm A's product change when firm A change its price? The effect of a price reduction on the quantity demanded of firm A's product depends upon whether its rival firm B responds by cutting its prices too. Similarly the effect of a price increase on the quantity demanded of firm A's product depends upon whether its rival firm B responds by raising its prices too. The gist of the above discussion is that:
Strategic interdependence: the rival firms aren't in complete control of their own destiny!

PROFIT MAXIMIZATION IN FOUR OLIGOPOLY SETTINGS

1. Cournot Model
2. Bertrand Model
3. Stackelberg Model
4. Sweezy (Kinked-Demand) Model

COURNOT OLIGOPOLY MODEL: OLIGOPOLY OUTPUT-SETTING MODELS

The earliest model of oligopoly was developed in 1838 by Augustin Cournot, a French economist. The basic features of Cournot model a Cournot model are:

- A few firms produce goods that are either perfect substitutes (homogeneous) or imperfect substitutes (differentiated).
- Firms set output, as opposed to price.
- Each firm believes their rivals will hold output constant if it changes its own output (The output of rivals is viewed as given or "fixed").
- Barriers to entry exist.

The Cournot model postulates that firms in oligopoly markets make simultaneous and independent output decisions. In other words Cournot model assumes that oligopoly demand curves are stable because each firm treats the *output* of the other firm as given, and then makes its own output decision. The relationship between an oligopoly firm's profit-maximizing output level and competitor output is called the oligopoly output- reaction curve because it shows how oligopoly firms react to competitor production decisions. A firm's reaction function shows how its optimal output varies with each possible action by its rival firm.

Cournot equilibrium output is found by simultaneously solving output-reaction curves for both competitors. Cournot equilibrium output exceeds monopoly output but is less than competitive output. Cournot equilibrium (Nash Equilibrium) is where the two reaction curves intersect. In Cournot model, a rival's action is its output choice. Cournot's model involves competition in quantities (or sales volume) and price is less explicit.

COURNOT MODEL: A NUMERICAL EXAMPLE

To illustrate the Cournot model and the concept of Cournot equilibrium, consider a two-firm **duopoly** facing a linear demand curve:

P = 1600 - Q

Where P is price and Q is total output in the market, thus Q= QA + QB
ie industry output constitutes firm A and firm B's output respectively

- Further, assume MCA = MCB = $100
- and average (AC) and marginal cost (MC)

AC = MC = $100

- To find the profit maximising output of Firm A given Firm B's output we need to find Firm A's marginal revenue (MR) and set it equal to MC. So, Firm A's Total Revenue is

$$TRA = (1600 - Q)\,QA$$
$$= [1600 - (QA + QB)]\,QA$$
$$= 1600QA - QA^2 - QAQB$$

Firm A's MR is thus

$$MRA = 1600 - 2QA - QB$$

If MC=100 then equating: MRA= MC

$$1600 - 2QA - QB = 100$$
$$QA = 750 - 0.5QB$$

This is Firm A's Reaction Curve. We can see that the profit-maximizing level of output for firm A depends upon the level of output produced by itself and firm B. Similarly, the profit-maximizing level of output for firm B depends upon the level of output produced by itself and firm A.
If we had begun by examining Firm B's profit maximising output we would find its reaction curve, i.e.

$$QB = 750 - 0.5QA$$

We can solve these 2 equations and find equilibrium quantity and price.
Solving for QA we find:

$$QA = 750 - 0.5QB$$
$$QA = 750 - 0.5(750 - 0.5QA)$$
$$0.75QA = 75$$
$$QA = 500 \text{ units}$$

Cournot equilibrium output = QA + QB
= 500 + 500 = 1000 units

Cournot equilibrium price = 1600 – Q
= 1600 – (1000)
= $600

π = TR – TC
= ($600 * 500) – (100* 500)
= $250,000 by each firm

Under perfect competition firms set prices equal to MC. So,

$$P = MC = \$100$$
$$P = 1600 - Q \Rightarrow Q = 1500$$

And equilibrium quantity

$$Q = 750 \text{ units each}$$

Under monopoly TR = P * Q
= (1600 – Q)Q
= 1600Q – Q2
MR = 1600 – 2 Q

Profit -Maximizing output level for monopoly firm:

MR = MC
1600 – 2 Q = 100
2Q = 1500
Q = 750

Profit –Maximizing monopoly price:

P = 1600 – Q

= 1600 – 750
= $850 &
π = $562,500

Figure 6

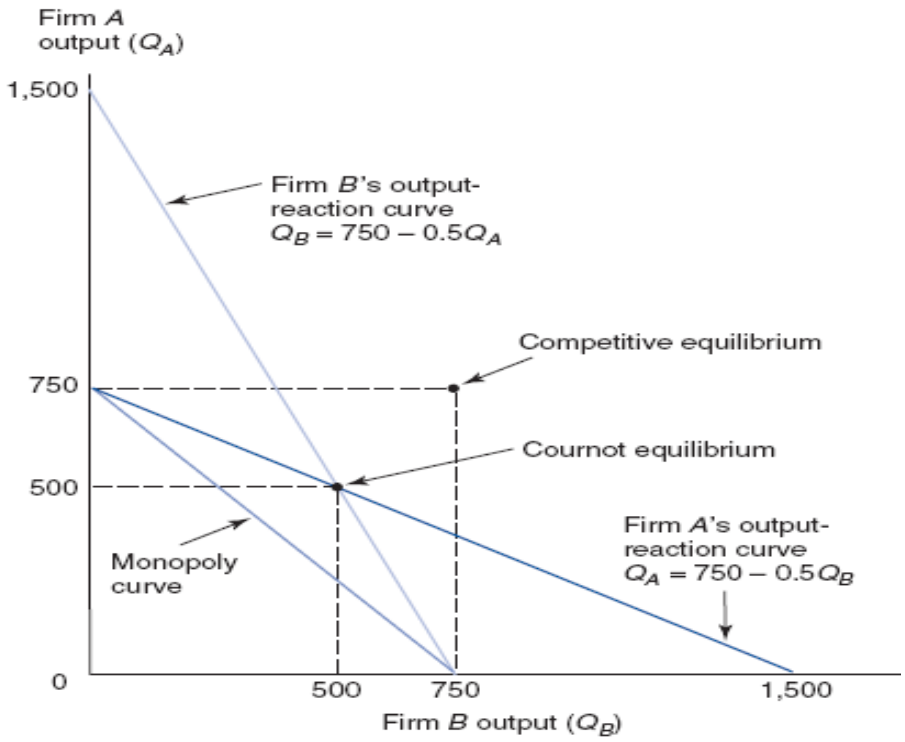


Figure 6 shows that in Cournot Equilibrium firms are maximizing profits simultaneously. The market is shared equally among the duopoly firm. Price is above the competitive equilibrium and below the monopoly equilibrium.

BERTRAND MODEL: OLIGOPOLY PRICE-SETTING MODELS

Joseph Bertrand, a French economist, presented his oligopoly model in 1883. Bertrand argued that a major problem with the Cournot model is that it failed to make price explicit. He showed that if firms compete on price when goods are homogenous, at least in consumer's eyes, then a price war will develop such that price approaches marginal cost. However, the introduction of differentiation leads to equilibrium closer in spirit to Cournot.

BERTRAND OLIGOPOLY: IDENTICAL PRODUCTS

- The Bertrand model focuses on price reactions.
- The Bertrand model predicts a competitive market price/output solution in oligopoly markets with identical products.

BERTRAND OLIGOPOLY: DIFFERENTIATED PRODUCTS

- The Bertrand model demonstrates how price-setting oligopolies profit with differentiated products.

BASIC FEATURES OF BERTRAND MODEL

• Few firms that sell to many consumers.

• Firms produce identical products at constant marginal cost.
• Each firm independently sets its price in order to maximize profits.
• Barriers to entry.
• Consumers enjoy
  - Perfect information.
  - Zero transaction costs.

BERTRAND EQUILIBRIUM
• Firms set P1 = P2 = MC! Why?
• Suppose MC < P1 < P2.
• Firm 1 earns (P1 - MC) on each unit sold, while firm 2 earns nothing.
• Firm 2 has an incentive to slightly undercut firm 1's price to capture the entire market. Firm 1 then has an incentive to undercut firm 2' price. This undercutting continues until each firm charges P1 = P2 = MC. in equilibrium.

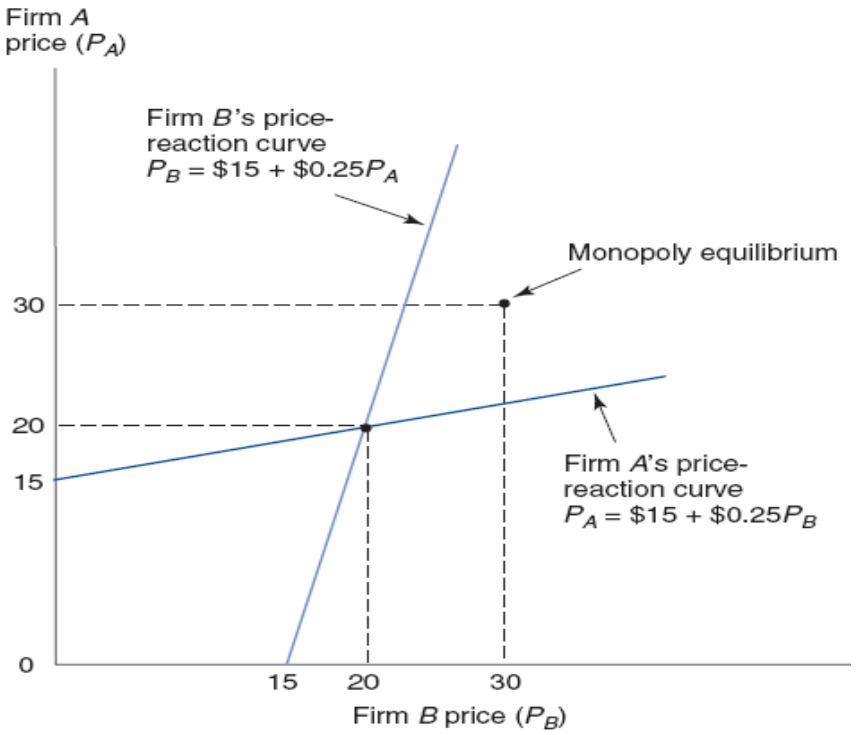
From the viewpoint of the manager, Bertrand model is undesirable. It leads to zero economic profit even if there are only two firms in the market. From the viewpoint of the consumers, Bertrand model is desirable: it leads to precisely the same outcome as a perfectly competitive market.
The Bertrand model shows that how price-setting oligopoly firms can profit by selling differentiated products. In the Bertrand model, the relationship between the profit-maximizing price level and competitor price is called the oligopoly price-reaction curve because it shows how the oligopoly firm reacts to competitor pricing decisions.

BERTRAND MODEL: A NUMERICAL EXAMPLE
Firm A demand: QA = 60 – 2 PA + PB
Firm B demand: QB = 60 – 2 PB + PA
MCA + MCB = 0
TRA = PA * QA
= PA * (60 – 2 PA + PB)
= 60 PA – 2 PA2 + PA PB
- π A = PA* QA - TCA Since cost = 0, therefore
  = 60 PA – 2 PA2 + PA PB

Since Price is the decision variable, firm A's profit-maximizing price is found by setting:
$\partial \pi$ A /∂PA = 0
60 – 4 PA + PB = 0
PA = 15 – 0.25 PB
Firm A price-reaction curve: PA = 15 – 0.25 PB
Firm B price-reaction curve: PB = 15 – 0.25 PA
PA = 15 – 0.25 (15 – 0.25 PA)
PA = 15 – 3.75 + 0.0625 PA
- monopoly market
- π = TR – TC where TR = P*Q

= P (60 -2P + p) – 0
= 60P – p2 since price is the decision variable:
$\partial \pi$ A /∂PA = 0
60 – 2P = 0 or P = $30
The profit-max monopoly output is:
Q = 60 – P or Q = 60 – 30 = 30 (000) units
Monopoly profits are: π = 60P - p2 = $900 (000)

Figure 7
Bertrand Equilibrium in a Two-Firm Duopoly



With 40 units sold at a market price of $20, firm A and firm B will each generate revenues of $800(000). Using the simplifying assumptions of no fixed costs and MCa = MCb = 0, each firm will also generate profits of $800(000). This is a stable equilibrium because given the competitor price, neither firm has any incentive to change price. Figure 7 illustrates this Bertrand market equilibrium price/output solution and price-reaction curves for each competitor.

In Bertrand, price competition is much "tougher." Hence, profits are lower. In Cournot, quantity commitments allow each firm to exclude buyers. Hence, they have some monopoly power.

THE CONTESTABLE MARKET MODEL

According to the ***contestable market model,*** barriers to entry and barriers to exit determine a firm's price and output decisions.

- Even if the industry contains only one firm, it could still be a competitive market if entry is open.
- In the contestable market model, an oligopoly with no barriers to entry sets a competitive price
- The stronger the ability of oligopolist to collude and prevent market entry, the closer it is to a monopolistic situation.
- The weaker the ability to collude is, the more competitive it is.
- Oligopoly markets lie between these two extremes.

In fact, according to the theory of contestable markets developed during the 1980s, even if an industry has a single firm (monopoly) or only a few firms (oligopoly), it would still operate as if it were perfectly competitive if entry is "absolutely free" (i.e., if other firms can enter the industry and face exactly the same costs as existing firms) and if exit is "entirely costless" (i.e., if there

are no sunk costs so that the firm can exit the industry without facing any loss of capital).

An example of this might be an airline that establishes a service between two cities already served by other airlines if the new entrant faces the same costs as existing airlines and could subsequently leave the market by simply reassigning its planes to other routes without incurring any loss of capital. When entry is absolutely free and exit is entirely costless, the market is contestable. Firms will then operate as if they were perfectly competitive and sell at a price which only covers their average costs (so that they earn, zero economic profit) even if there is only one firm or a few of them in the market.

KEY ASSUMPTIONS

- Producers have access to same technology.
- Consumers respond quickly to price changes.
- Existing firms cannot respond quickly to entry by lowering price.
- Absence of sunk costs.

KEY IMPLICATIONS

- Threat of entry disciplines firms already in the market.
- Incumbents have no market power, even if there is only a single incumbent (a monopolist).

Lesson 30

OLIGOPOLY MODELS

STACKELBERG OLIGOPOLY

Stackelberg model, developed by German economist H. Von Stackelberg in 1934 postulates a first-mover advantage for the oligopoly firm that initiates the process of determining market output. Up till now, we have analyzed oligopoly situations that are symmetric in that firm 2 is the "mirror image" of firm 1. In many oligopoly markets, however, firms differ from one another.

BASIC FEATURES OF THE MODEL

- Firms produce differentiated or homogeneous products.
- There are few firms selling serving many consumers.
- A single firm (the leader) chooses an output before all other firms choose their output. The leader commits to an output before all other firms.
- All other firms (the followers) take as given the output of the leader and choose outputs that maximize profits given the leader's output.
- Barriers to entry exist.

To illustrate Stackelberg first-mover advantages, we reconsider the Cournot model but now assume that firm A, as a leading firm, correctly anticipate the output–reaction of firm B, the following firm. With prior knowledge of firm B's output–reaction curve, **QB = 750 - 0.5QA**, firm A's total revenue curve becomes:

$P = 1600 - Q$

Where $Q = QA + QB$

Further, assume MCA = MCB = $100

And average (AC) and marginal cost (MC)

$AC = MC = \$100$

Firm A's Total Revenue is

$TRA = (1600 - Q)\,QA$

$= [1600 - (QA + QB)]\,QA$

$= 1600QA - QA^2 – QAQB$

Substituting Firm B's output-reaction curve, $QB = 750 - 0.5QA$

$= 1600QA - QA^2 – QA(750 - 0.5QA)$

$TRA = 850QA – 0.5QA^2$

Marginal revenue for firm A is:

$MRA = 850QA - QA$

$MRA = MCA$

$850QA – QA = 100$

$QA = 750$ units

After firm A determined its output level, the amount produced by firm B is calculated from Firm B's output-reaction curve:

$QB = 750 - 0.5QA$

$= 750 – 0.5(750) = 375$

Stackelberg market Equilibrium level of output is:

$Q = QA + QB$

$= 750 + 375$

$= 1{,}125$ units

Stackelberg market Equilibrium Price is:

= 1600 – Q
= 1600 – (1,125)
= $ 475
ΠA = $356,250 and ΠB = $178,125

Cournot output = 1000 units
Cournot price = $600

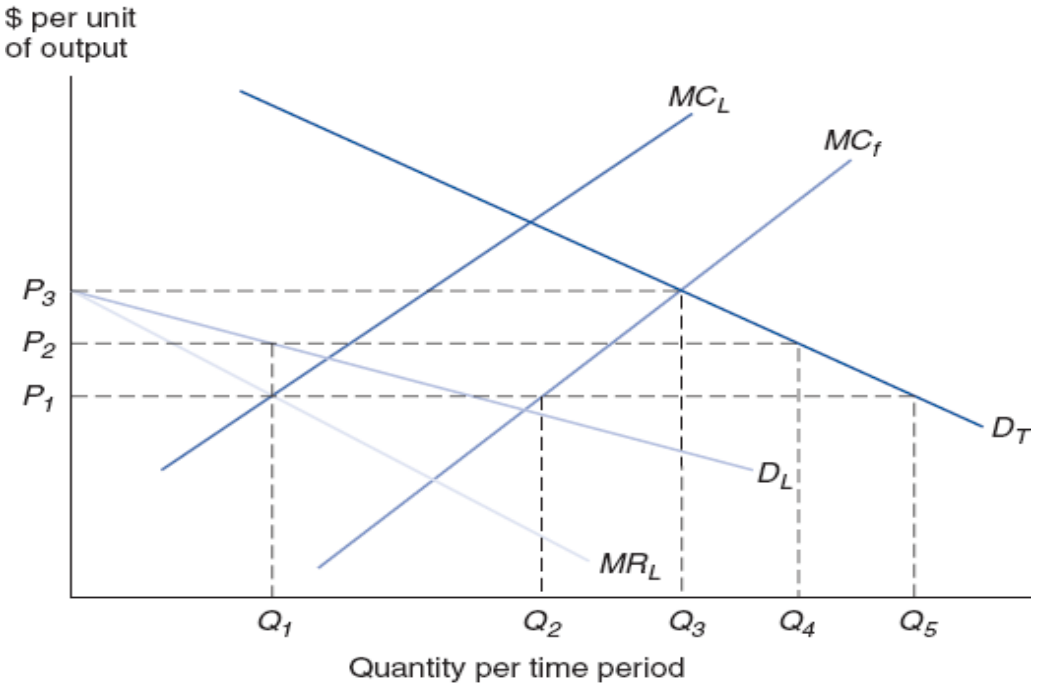
We might notice that market output is greater in Stackelberg equilibrium than in Cournot equilibrium because the first mover, firm A, produces more output while the follower, firm B, produces less output. Stackelberg equilibrium also results in a lower market price than that observed in Cournot equilibrium.

PRICE LEADER (BAROMETRIC FIRM)

- Largest, dominant, or lowest cost firm in the industry
- Demand curve is defined as the market demand curve less supply by the followers
- Followers
- Take market price as given and behave as perfect competitors

Price leadership occurs when one firm establishes itself as the industry leader and other firms follow its pricing policy. This leadership may result from the size and strength of the leading firm, from cost-efficiency, or as a result of the ability of the leader to establish prices that produce satisfactory profits throughout the industry. The leader faces a price / output problem similar to monopoly; other firms are price takers and face a competitive price / output problem. This is illustrated in Figure where the total market demand curve is DT, the marginal cost curve of the leader is MCL, and the horizontal summation of all of the follower firms is labeled MCf.

Figure 1

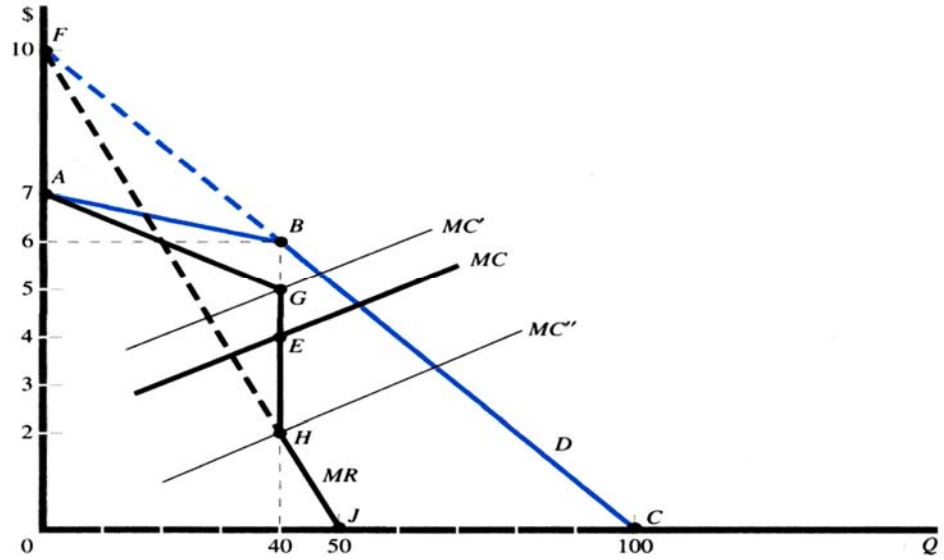


DL = DT – MCf
MCf = Sf found by setting P = MCf because DT and Sf are functions of price, DL is also a functions of price.

KINKED DEMAND CURVE MODEL (SWEEZY OLIGOPOLY)
BASIC FEATURES OF THE MODEL

- Few firms in the market serving many consumers.
- Firms produce differentiated products.
- Barriers to entry.
- Each firm believes rivals will match (or follow) price reductions, but won't match (or follow) price increases.
- If an oligopolist raises price, other firms will not follow, so demand will be elastic
- If an oligopolist lowers price, other firms will follow, so demand will be inelastic
- Implication is that demand curve will be kinked, MR will have a discontinuity, and oligopolist will not change price when marginal cost changes.

KEY FEATURE OF SWEEZY MODEL
PRICE-RIGIDITY
Figure 2



The Kinked demand curve model was purposed by Paul Sweezy in 1939. This model basically explains the price rigidity.
According to Sweezy, oligopolist faces a demand curve that has a kink at the prevailing price and is highly elastic for price increases but much less elastic for price cuts. In this model, oligopolist recognize their interdependence but act without collusion in keeping their prices constant, even in the face of changed cost and demand conditions--preferring instead to compete on the basis of quality, advertising, service, and other forms of non price competition. The Sweezy model is shown in Figure 2.

As in other forms of market organization, the firm under oligopoly can earn profits, break even, or incur losses in the short run, and it will continue to produce as long as P > AVC. From Figure 2 we can also see that the oligopolist's marginal cost curve can rise or fall anywhere within the

discontinuous portion of the MR curve (i.e., from MC' to MC") without inducing the oligopolist to change the prevailing price of $6 and sales of 40 units (as long as P> AVC). Only if the MC curve shifts above the MC' curve will the oligopolist be induced to increase its price and reduce quantity, or only if the MC curve shifts below MC" will the oligopolist lower price and increase quantity. With a rightward or a leftward shift in the demand curve, sales will increase or fall, respectively, but the oligopolist will keep the price constant as long as the kink on the demand curve will remain at the same price and the MC curve continues to intersect the discontinuous or vertical portion of the MR curve.

SWEEZY OLIGOPOLY: A NUMERICAL EXAMPLE

The demand functions for price increases and price decreases are:

QA = 280 - 40PA or PA = 7 – 0.025QA

QB = 100 - 10PB or PB = 10 – 0.1QB

$TC = 2Q + 0.25\ Q^2$ and MC = 2 + 0.05Q

TRA = PA *QA = (7 – 0.025QA) QA

TRB = PB *QB = (10 – 0.1QB) QB

MRA = 7 – 0.05QA and

MRB = 10 – 0.02QB

To find the Kink in Demand Curve, we set

Q= QA + QB, since the two Demand Functions are:

PA = 7 – 0.025 Q

PB = 10 – 0.1Q

PA = PB and solving for Q

7 – 0.025 Q = 10 – 0.1Q

Q = 40 and P= 7 – 0.025(40) = $6

The upper and lower limit of the MR gap is:

MRA = 7 – 0.5(40) = 7 – 2 = 5

MRB = 10 – 0.2(40) = 10 – 8 = 2

Since MC = 2 + 0.05(40) = 4

The MC curve intersects the vertical portion of the MR curve. The total profit (Π) of the firm are:

Π = TR – TC = PQ – 2Q – 0.025 Q^2

= 6(40) – 2(40) – 0.025$(40)^2$

= $120

CARTELS AND COLLUSIONS

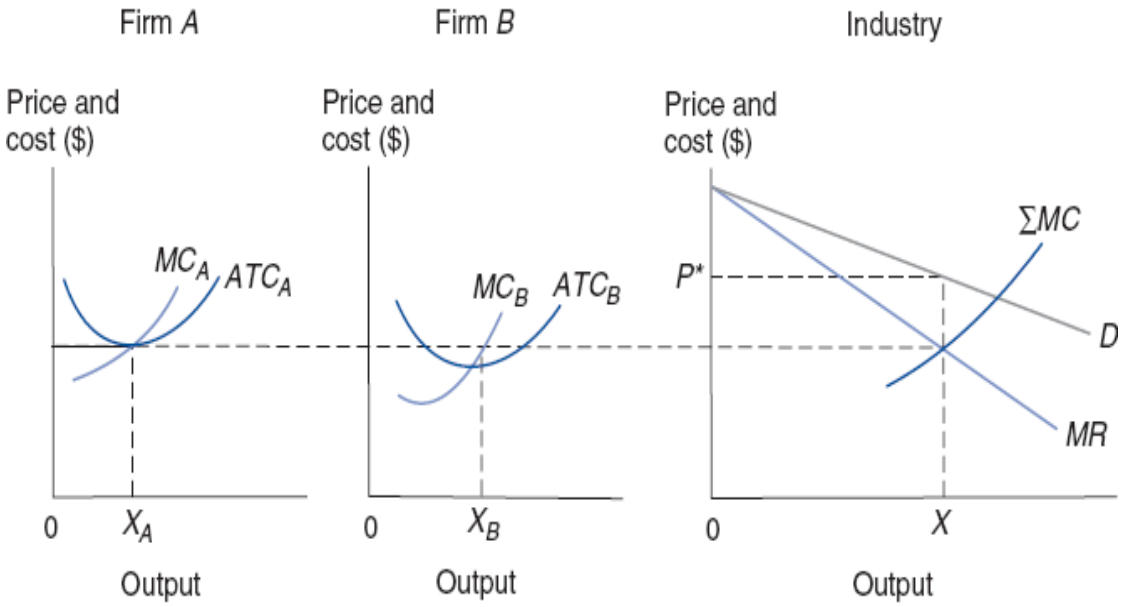
In the kinked demand curve model, oligopolist did not collude to restrict or eliminate competition in order to increase profits. Collusion can be overt or explicit, as in centralized and market sharing cartels, or tacit or implicit, as in price leadership models. There are two types of cartels: the centralized cartels and market sharing cartels. Centralized cartels: cartels directly (although secretly) reach agreements among competitors to reduce the uncertainty arising from their mutual interdependence.

All firms in an oligopoly market benefit if they get together and set prices to maximize industry profits. A group of competitors operating under such a formal over (explicit) agreement is called a **cartel**. If an informal covert or explicit agreement is reached, the firms are said to be operating in **collusion**. Both practices are illegal in the various countries. However, cartels are legal in some parts of the world. Several important domestic markets are also dominated by producer associations that operate like cartels and appear to flourish without interference from the government. Agricultural commodities such as milk are prime examples of products marketed under cartel-like arrangements.

A cartel that has absolute control over all firms in an industry can operate as a monopoly. To illustrate, consider the situation shown in Figure 3. The marginal cost curves of each firm are summed horizontally to arrive at an industry marginal cost curve. Equating the cartel's total marginal cost with the industry marginal revenue curve determines the profit-maximizing output and the price, *P**, to be charged. Once this profit-maximizing price/output level has been determined, each individual firm finds its optimal output by equating its own marginal cost curve to the previously determined profit-maximizing marginal cost level for the industry.

Profits are often divided among firms on the basis of their individual level of production, but other allocation techniques can be employed. Market share, production capacity, and a bargained solution based on economic power have all been used in the past. For a number of reasons, cartels are typically rather short-lived. In addition to the long-run problems of changing products and of entry into the market by new producers, cartels are subject to disagreements among members. Although firms usually agree that maximizing joint profits is mutually beneficial, they rarely agree on the equity of various profit-allocation schemes. This problem can lead to attempts to subvert the cartel agreement.

Figure 3
Price / Output Determination of a Cartel



Cartel subversion can be extremely profitable. Consider a two-firm cartel in which each member serves 50 percent of the market. Cheating by either firm is very difficult, because any loss in profits or market share is readily detected.

Lesson 31

OLIGOPOLY: GAME THEORETIC APPROACH

OLIGOPOLY: GAME THEORETIC APPROACH

When just a few large firms dominate a market so that actions of each one have an important impact on the others. In such a market, each firm recognizes its strategic interdependence with others. An ***oligopoly*** is a market dominated by a small number of ***strategically interdependent*** firms.

Oligopoly presents the greatest challenge to economists. The essence of oligopoly is strategic interdependence. The economists have had to modify the tools used to analyze other market structures and to develop entirely new tools as well. One approach—game theory—has yielded rich insights into oligopoly behavior. Economists have developed:

- Collusive Models
- Limit-Pricing Models
- Managerial Models
- Behavioral Models

But these models do not provide a general theory of oligopoly in the sense that none of these models could fully explain the **decision-making process** of oligopolists.

GAME THEORY: is concerned with the choice of the best or optimal strategy in conflict situation. The lessons drawn from homely games like chess and poker had nearly universal application to economic situations in which the participants had the power to anticipate and affect other participants' actions.

While John von Neumann and Oskar Morgenstern did pioneering work in this field as early as the late 1940s, the analytical breakthrough was made by John Harsanvi, John Nash and Reinhard Selton, who were awarded Nobel Prize (in 1994) "for their pioneering analysis of equilibrium in the theory of non-cooperative games." The Nobel prize (2005) was awarded to: Robert J. Aumann and Thomas C. Schelling "for having enhanced our understanding of conflict and cooperation through game-theory analysis".

John Nash introduced the distinction between cooperative games, in which binding agreements can be made, and non-cooperative games, where binding agreements are not feasible. Nash also developed an equilibrium concept for predicting the outcome of non-cooperative games that is called “Nash equilibrium”.

GAME THEORY BASICS

Types of Games

- Zero-sum game: offsetting gains/losses.
- Positive sum game: potential for mutual gain.
- Negative-sum game: potential for mutual loss.
- Cooperative games: joint action is favored.
- Non-Cooperative Games

Role of Interdependence

- Sequential games involve successive moves.
- Simultaneous-move games incorporate coincident moves.

Game theory approach

- An approach to modeling strategic interaction of oligopolists in terms of moves and countermoves

GAME THEORY: ELEMENTS

- **Strategies:** A Strategy is a specific course of action with clearly defined values for policy variables. A strategy is dominant if it is optimal regardless of what the other player does.
- **Payoff** of a strategy is the "net gain" it will bring to the firm for any given counter-strategy of the rival firm
- **Payoff matrix** is the table giving the payoffs from all the strategies open to the firm and the rivals' responses
- **Policy Variables**: Price, Quantity, Quality, advertising Research and Development Expenditure, Changes in the number of products.
- **Players:** players are interdependent **Players** are the decision-makers (the managers of oligopolist firms). other players' actions are not entirely predictable

GAME THEORY concepts are used to develop effective competitive strategies for setting prices, the level of product quality, research and development, advertising, and other forms of non price competition in oligopoly markets.

STRATEGIC BEHAVIOR refers to the plan of action or behavior of an oligopolist, after taking into consideration all possible reactions of its competitors, as they compete for profits or other advantages. Since there are only a few firms in the industry, the actions of each affects the others, and the reaction of the others must be kept in mind by the first in charting its best course of action. Thus, each oligopolist changes the product price, the quantity of the product that it sells, the level of advertising, and so on, so as to maximize its profits after having considered all possible reactions of its competitors to each of its courses of action.

Every game theory model includes players, strategies, and payoffs. The players are the decision makers (here, the managers of oligopolist firms) whose behavior we are trying to explain and predict. The strategies are the choices to change price, develop new products, undertake a new advertising campaign, build new capacity, and all other such actions that affect the sales and profitability of the firm and its rivals. The payoff is the outcome or consequence of each strategy. For each strategy adopted by a firm, there is usually a number of strategies (reactions) available to a rival firm. The payoff is the outcome or consequence of each combination of strategies by the two firms. The payoff is usually expressed in terms of the profits or losses of the firm that we are examining as a result of the firm's strategies and the rivals responses. The table giving the payoffs from all the strategies open to the firm and the rivals' responses is called the payoff matrix.

In a simultaneous-move game, each decision maker makes choices without specific knowledge of competitor counter moves. In a sequential-move game, decision makers make their move after observing competitor moves. If two firms set prices without knowledge of each other's decisions, it is a simultaneous-move game. If one firm sets its price only after observing its rival's price, the firm is said to be involved in a sequential-move game. In a one shot game, the underlying interaction between competitors occurs only once; in a repeat game, there is an ongoing interaction between competitors.

A game theory strategy is a decision rule that describes the action taken by a decision maker at any point in time. A simple introduction to game theory strategy is provided by perhaps the most famous of all simultaneous-move one-shot games: The so-called Prisoner's Dilemma.

PRISONERS' DILEMMA

The **prisoner's dilemma** is a fundamental problem in game theory that demonstrates why two people might not cooperate even if it is in both their best interests to do so. It was originally

framed by Merrill Flood and Melvin Dresher working at RAND in 1950. Albert W. Tucker formalized the game with prison sentence payoffs and gave it the "prisoner's dilemma" name (Poundstone, 1992).

A CLASSIC EXAMPLE OF THE PRISONER'S DILEMMA IS PRESENTED AS FOLLOWS:
Two suspects are arrested by the police. The police have insufficient evidence for a conviction, and, having separated the prisoners, visit each of them to offer the same deal. If one testifies for the prosecution against the other (*defects*) and the other remains silent (*cooperates*), the defector goes free and the silent accomplice receives the full 10-year sentence. If both remain silent, both prisoners are sentenced to only 1 year in jail for a minor charge. If each betrays the other, each receives a five-year sentence. Each prisoner must choose to betray the other or to remain silent. Each one is assured that the other would not know about the betrayal before the end of the investigation. How should the prisoners act?

TABLE 1
PAYOFF MATRIX

		Individual B	
		Confess	Don't Confess
Individual A	Confess	(5, 5)	(0, 10)
	Don't Confess	(10, 0)	(1, 1)

TABLE 2
DOMINANT STRATEGY
Both Individuals Confess
(Nash Equilibrium)

		Individual B	
		Confess	Don't Confess
Individual A	Confess	(5, 5)	(0, 10)
	Don't Confess	(10, 0)	(1, 1)

If we assume that each player cares only about minimizing his or her own time in jail, then the prisoner's dilemma forms a non-zero-sum game in which two players may each either cooperate with or defect from (betray) the other player. In this game, as in most game theory, the only concern of each individual player (prisoner) is maximizing his or her own payoff, without any concern for the other player's payoff. The unique equilibrium for this game is a Pareto sub-optimal solution, that is, rational choice leads the two players to both play *defects*, even though each player's individual reward would be greater if they both played cooperatively.

In the classic form of this game, cooperating is strictly dominated by defecting, so that the only possible equilibrium for the game is for all players to defect. No matter what the other player does, one player will always gain a greater payoff by playing defect. Since in any situation playing defect is more beneficial than cooperating, all rational players will play defect, all things being equal.

Oligopolistic firms often face a problem called the prisoners dilemma. This refers to a situation in which each firm adopts its dominant strategy but each could do better (i.e., earn larger profits) by cooperating.
From Table1, we see that confessing is the best or dominant strategy for suspect A no matter what suspect B does. The reason is that if suspect B confesses, suspect A receives a 5-year sentence if he confesses and a 10-year jail sentence if he does not: Similarly, if suspect B does not confess, suspect A goes free if he confesses and receives a 10-year jail sentence if he does not. Thus, the dominant strategy for suspect A is to confess. Confessing is also the best or dominant strategy for suspect B. The reason is that if suspect A confesses, suspect B gets a 5-year jail sentence if he also confesses and a 10-year jail sentence if he does not. Similarly, if suspect A does not confess, suspect B goes free if he confesses and gets 1 year if he does not. Thus, the dominant strategy for suspect B is also to confess. From Table 2, we see that with each suspect adopting his dominant strategy of confessing, each end up receiving a 5-year jail Sentence.

PRICE COMPETITION AND THE PRISONERS' DILEMMA

The concept of the prisoners' dilemma can be used to analyze price and non-price competition in oligopolistic markets, as well as the incentive to cheat (i.e., the tendency to secretly cut price or sell more than its allocated quota) in a cartel. Oligopolistic price competition in the presence of the prisoners' dilemma can be examined with the payoff matrix in Table 3.
The payoff matrix of Table 3 shows that if firm B charged a low price (say, $6), firm A would earn a profit of 2 if it also charged the low price ($6) and 1 if it charged a high price (say, $8). Similarly, if firm B charged a high price ($8), firm A, would earn a profit of 5 if it charged the low price and 3 if it charged the high price. Thus, firm A should adopt its dominant strategy of charging the low price. Turning to firm B, we see that if firm A charged the low price, firm B would earn a profit of 2 if it charged the low price and 1 if it charged the high price. Similarly, if firm A charged the high price, firm B would earn a profit of 5 if it charged the low price and 3 if it charged the high price. Thus, firm B should also adopt its dominant strategy of charging the low price. However, both firms could do better (i.e., earn the higher profit of 3) if they cooperated and both charged the high price (the bottom right cell).

Thus, the firms are in a prisoners' dilemma: Each firm will charge the low price and earn a smaller profit because if it charges the higher price, it cannot trust its rival to also charge the high price.

NON-PRICE COMPETITION, CARTEL CHEATING, AND THE PRISONERS' DILEMMA

By simply changing the heading of the columns and rows of the payoff matrix in Table 3, we can use the same payoff matrix to examine non price competition and cartel cheating. For example, if we changed the heading of "low price" to "advertise" and changed the heading of "high price" to "don't advertise" in the columns and rows of the payoff matrix of Table 3, we can use the same payoff matrix of Table 4 to analyze advertising as a form of non-price competition in the presence of the prisoners 'dilemma. We would then see that each firm would adopt its dominant strategy of advertising and (as in the case of charging a low price) would earn a profit of 2. Both firms, however, would do better by not advertising because they would then earn (as in the case of charging a high price) the higher profit of 3.

Similarly, if we now changed, the heading of "low price" or "advertise' to "cheat" and the heading of "high price" or "don't advertise" to "don't cheat" in the columns and rows of the payoff matrix of Table 3, we could use the same payoffs in Table 5 to analyze the incentive for cartel members to cheat in the presence of the prisoners' dilemma. In this case, each firm adopts its dominant strategy of cheating and (as in the case of charging the low price or advertising) earns

a profit of 2. But by not cheating, each member of the cartel would earn the higher profit of 3. The cartel members therefore are facing the prisoners' dilemma.

TABLE 3
APPLICATION: PRICE COMPETITION
Dominant Strategy: Low Price

		Firm B	
		Low Price	High Price
Firm A	Low Price	(2, 2)	(5, 1)
	High Price	(1, 5)	(3, 3)

TABLE 4
APPLICATION: NONPRICE COMPETITION
Dominant Strategy: Advertise

		Firm B	
		Advertise	Don't Advertise
Firm A	Advertise	(2, 2)	(5, 1)
	Don't Advertise	(1, 5)	(3, 3)

TABLE 5
APPLICATION: CARTEL CHEATING
Dominant Strategy: Cheat

		Firm B	
		Cheat	Don't Cheat
Firm A	Cheat	(2, 2)	(5, 1)
	Don't Cheat	(1, 5)	(3, 3)

NASH EQUILIBRIUM: DOMINANT STRATEGY

To see how players choose strategies to maximize their payoffs, we begin with the simplest type of game in an industry (duopoly) .composed of two firms, firm A and firm B, and a choice of two strategies for each-advertise or don't advertise. Firm A, expects to earn higher profits if it advertises than if it doesn't. But the actual level of profits of firm A depends also on whether firm B advertises or not. Thus, each strategy by firm A (i.e., advertise or don't advertise) can be associated with each of firm B's strategies (also to advertise or not to advertise).

The four possible outcomes for this simple game are illustrated by the payoff matrix in Table 6. The first number of each of the four cells refers to the payoff (profit) for firm A, while the second is the payoff (profit) for firm B. From Table 6, we see that if both firms advertise, firm A will earn a profit of 4, and firm B will earn a profit of 3 (the top left cell of the payoff matrix). The bottom left cell of the payoff matrix, on the other hand, shows that if firm A doesn't advertise and firm B does, firm A will have a profit of 2, and firm B will have a profit of 5. The other payoffs in the second column of the table can be similarly interpreted.

If firm B does advertise (i.e., moving down the left column of Table 6), we see that firm A will earn a profit of 4 if it also advertises and 2 if it doesn't. Thus, firm A should advertise if firm B advertises. If firm B doesn't advertise (i.e., moving down the right column in Table 6), firm A would earn a profit of 5 if it advertises and 3 if it doesn't. Thus, firm A should advertise whether firm B advertises or not. Firm A's profits would always be greater if it advertises than if it doesn't

regardless of what firm B does. We can then say that advertising is the dominant strategy for firm A. The dominant strategy is the optimal choice for a player no matter what the opponent does.

The same is true for firm B. whatever firm A does (i.e., whether firm A advertises or not), it would always pay for firm B to advertise. We can see this by moving across each row of Table 6. Specifically, if firm A advertises, firm B's profit would be 3 if it advertises and 1 if it does not. Similarly, if firm A does not advertise, firm B's profit would be 5 if it advertises and 2 if it doesn't. Thus, the dominant strategy for firm B is also to advertise. In this case, both firm A and firm B have the dominant strategy of advertising, and this will, therefore, be the final equilibrium. Both firm A and firm B will advertise regardless of what the other firm does and will earn a profit of 4 and 3,

NASH EQUILIBRIUM

Not all games have a dominant strategy for each player. In fact, it is more likely in the real world that one or both players do not have a dominant strategy. An example of this is shown in the payoff matrix in Table 10. This is the same as the, payoff matrix in Table 6, except that the first number in the bottom tight cell was changed from 3 to 6. Now firm B has a dominant strategy, but firm A does not. The dominant strategy for firm B is to advertise whether firm A advertises or not, exactly as above, because the payoffs for firm B are the same as in Table 6. Firm A, however, has no dominant strategy now.

What is the optimal strategy for Firm A if Firm B chooses to advertise?
If Firm A chooses to advertise, the payoff is 4. Otherwise, the payoff is 2. The optimal strategy is to advertise.

TABLE 6

		Firm B	
		Advertise	Don't Advertise
Firm A	Advertise	(4, 3)	(5, 1)
	Don't Advertise	(2, 5)	(3, 2)

What is the optimal strategy for Firm A if Firm B chooses not to advertise?
If Firm A chooses to advertise, the payoff is 5. Otherwise, the payoff is 3. Again, the optimal strategy is to advertise.

TABLE 7

		Firm B	
		Advertise	Don't Advertise
Firm A	Advertise	(4, 3)	(5, 1)
	Don't Advertise	(2, 5)	(3, 2)

Regardless of what Firm B decides to do, the optimal strategy for Firm A is to advertise. The dominant strategy for Firm A is to advertise.

TABLE 8

		Firm B	
		Advertise	Don't Advertise
Firm A	Advertise	(4, 3)	(5, 1)
	Don't Advertise	(2, 5)	(3, 2)

What is the optimal strategy for Firm B if Firm A chooses to advertise?
If Firm B chooses to advertise, the payoff is 3. Otherwise, the payoff is 1. The optimal strategy is to advertise.

TABLE 9

		Firm B	
		Advertise	Don't Advertise
Firm A	Advertise	(4, 3)	(5, 1)
	Don't Advertise	(2, 5)	(3, 2)

What is the optimal strategy for Firm B if Firm A chooses not to advertise?
If Firm B chooses to advertise, the payoff is 5. Otherwise, the payoff is 2. Again, the optimal strategy is to advertise.

TABLE 10

		Firm B	
		Advertise	Don't Advertise
Firm A	Advertise	(4, 3)	(5, 1)
	Don't Advertise	(2, 5)	(6, 2)

Regardless of what Firm A decides to do, the optimal strategy for Firm B is to advertise. The dominant strategy for Firm B is to advertise.

TABLE 11

		Firm B	
		Advertise	Don't Advertise
Firm A	Advertise	(4, 3)	(5, 1)
	Don't Advertise	(2, 5)	(3, 2)

The dominant strategy for Firm A is to advertise and the dominant strategy for Firm B is to advertise. The Nash equilibrium is for both firms to advertise.

TABLE 12

		Firm B	
		Advertise	Don't Advertise
Firm A	Advertise	(4, 3)	(5, 1)
	Don't Advertise	(2, 5)	(3, 2)

What is the optimal strategy for Firm A if Firm B chooses to advertise?
If Firm A chooses to advertise, the payoff is 4. Otherwise, the payoff is 2. The optimal strategy is to advertise.

TABLE 13

		Firm B	
		Advertise	Don't Advertise
Firm A	Advertise	(4, 3)	(5, 1)
	Don't Advertise	(2, 5)	(6, 2)

What is the optimal strategy for Firm A if Firm B chooses not to advertise?
If Firm A chooses to advertise, the payoff is 5. Otherwise, the payoff is 6. In this case, the optimal strategy is not to advertise.

TABLE 14

		Firm B	
		Advertise	Don't Advertise
Firm A	Advertise	(4, 3)	(5, 1)
	Don't Advertise	(2, 5)	(6, 2)

The optimal strategy for Firm A depends on which strategy is chosen by Firms B. Firm A does not have a dominant strategy.

TABLE 15

		Firm B	
		Advertise	Don't Advertise
Firm A	Advertise	(4, 3)	(5, 1)
	Don't Advertise	(2, 5)	(6, 2)

What is the optimal strategy for Firm B if Firm A chooses to advertise?
If Firm B chooses to advertise, the payoff is 3. Otherwise, the payoff is 1. The optimal strategy is to advertise.

TABLE 16

		Firm B	
		Advertise	Don't Advertise
Firm A	Advertise	(4, 3)	(5, 1)
	Don't Advertise	(2, 5)	(6, 2)

What is the optimal strategy for Firm B if Firm A chooses not to advertise?
If Firm B chooses to advertise, the payoff is 5. Otherwise, the payoff is 2. Again, the optimal strategy is to advertise.

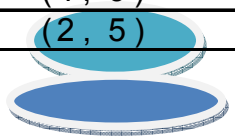
TABLE 17

		Firm B	
		Advertise	Don't Advertise
Firm A	Advertise	(4, 3)	(5, 1)
	Don't Advertise	(2, 5)	(6, 2)

Regardless of what Firm A decides to do, the optimal strategy for Firm B is to advertise. The dominant strategy for Firm B is to advertise.

TABLE 18

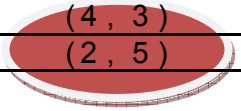
		Firm B	
		Advertise	Don't Advertise
Firm A	Advertise	(4, 3)	(5, 1)
	Don't Advertise	(2, 5)	(6, 2)



The dominant strategy for Firm B is to advertise. If Firm B chooses to advertise, then the optimal strategy for Firm A is to advertise. The Nash equilibrium is for both firms to advertise. The Nash equilibrium is a situation where each player chooses his or her optimal strategy, given the strategy chosen by the other player.

TABLE 19

		Firm B	
		Advertise	Don't Advertise
Firm A	Advertise	(4, 3)	(5, 1)
	Don't Advertise	(2, 5)	(6, 2)



Lesson 32

OLIGOPOLY: GAME THEORETIC APPROACH (CONTINUED)

GAMES IN ECONOMICS

Repeated Game: game is played repeatedly over a period of time. **Finitely repeated games** include games that take place over fixed or uncertain period of time.
Infinitely repeated games occur over and over again without boundary or limit. However, if number of periods is fixed, players will have incentive to 'cheat' in the last period due to lack of threat of retaliation, which will then allow them to cheat in all periods
Simultaneous games are games in which players make their strategy choices at the same time
Sequential games are games in which players make their decisions sequentially. In sequential games, the first mover may have an advantage.

In a **simultaneous-move game**, each decision maker makes choices without specific knowledge of competitor counter moves. In a **sequential-move game**, decision makers make their move after observing competitor moves. If two firms set prices without knowledge of each other's decisions, it is a simultaneous-move game. If one firm sets its price only after observing its rival's price, the firm is said to be involved in a sequential-move game. In a one shot game, the underlying interaction between competitors occurs only once; in a repeat game, there is an ongoing interaction between competitors.

A game theory strategy is a decision rule that describes the action taken by a decision maker at any point in time. A simple introduction to game theory strategy is provided by perhaps the most famous of all simultaneous-move one-shot games: The so-called **Prisoner's Dilemma**. The **Prisoner's Dilemma** game fascinates game theorists for various reasons. Oligopolistic firms often face a problem called the prisoners dilemma. This refers to a situation in which each firm adopts its dominant strategy but each could do better (i.e., earn larger profits) by cooperating. Theory shows that when games are repeated, the chances for cooperation or collusion increase.

Table 1
Infinitely Repeated Games
Infinitely repeated games occur over and over again without boundary or limit.

	Pepsi-Cola		
	Pricing Strategy	**Discount P**	**Regular P**
Coca-Cola	**Discount P**	**($4b,$2b)**	**($8b,$1b)**
	Regular P	**($2b,$5b)**	**($6b,$4b)**

NASH EQUILIBRIUM: COOPERATIVE VS NON COOPERATIVE GAMES

The "equilibrium" would be where each player makes a decision which represents the best outcome in response to what other players' decisions are. Nash equilibrium is a point where no

player can improve his position by selecting any other available strategy while other are playing their best options and not changing their strategies.

In Table 1, each firm's secure strategy is to offer a discount price regardless of the other firm's actions. The outcome is that both firms offer discount prices and earn relatively modest profits. This outcome is also called a Nash equilibrium because, given the strategy of its competitor, neither firm can improve its own payoff by independently changing its own strategy. In the case of Coca-Cola, given that Pepsi-Cola has chosen a discount pricing strategy, it too would decide to offer discount prices. When Pepsi-Cola offers discount prices, Coca-Cola can earn profits of $4b rather than $2b per week by also offering a discount. Similarly, when Coca-Cola offers discount prices, Pepsi-Cola can earn maximum profits of $2b per week, versus $1b per week, by also offering a discount.

Clearly, profits are less than if they colluded and both charged regular prices. As seen in Table 1, Coca-Cola would earn $6b per week and Pepsi-Cola would earn $4b per week if both charged regular prices. This is a business application of the Prisoner's Dilemma because the dual discount pricing Nash equilibrium is inferior from the firms' viewpoint to a collusive outcome where both competitors agree to charge regular prices. Of course, if firms collude and agree to charge high prices, consumers are made worse off. This is why price collusion among competitors is illegal in many countries.

A secure strategy, sometimes called the maximin strategy, guarantees the best possible outcome given the worst possible scenario. In the case of Prisoner's Dilemma, the worst possible scenario for each suspect is that the other chooses to confess. Each suspect can avoid the worst possible outcome of receiving a harsh 5 years in prison sentence only by choosing to confess. For each suspect, the secure strategy is to confess, thereby becoming a prisoner, because neither could solve the riddle posed by the Prisoner's Dilemma.

NASH BARGAINING

Though the Prisoner's Dilemma is posed within the scope of a bargaining problem between two suspects, it has obvious practical applications in business. Competitors like Coca-Cola and Pepsi-Cola confront similar bargaining problems on a regular basis. Suppose each has to decide whether or not to offer a special discount to a large grocery store retailer.

Table 1 show that if neither offers discount pricing, a weekly profit of $6b will be earned by Coca-Cola and $4b per week will be earned by its smaller competitor, Pepsi- Cola. This is the best possible scenario for both. However, if Coca-Cola is the only one to offer a discount, it will earn $8b per week, while Pepsi-Cola profits fall to $1b per week. If Pepsi-Cola offers a discount and Coca-Cola continues to charge the regular price, Pepsi-Cola profits will total $5b per week while Coca-Cola weekly profits fall to $2b. The only secure means Coca-Cola has for avoiding the possibility of only $2b per week profit is to grant a discount price to the retailer, thereby assuring itself of a weekly profit of at least $4b, Similarly, the only means Pepsi-Cola has of avoiding the possibility of profits of $1b, per week is to also grant a discount price to the grocery retailer, thereby assuring itself of at least $2b, in weekly profits. For both Coca-Cola and Pepsi-Cola, the only secure strategy is to offer discount prices, thereby assuring consumers of bargain prices and themselves of modest profits of $4b and $2b per week, respectively.

REPEATED GAMES: INFINITE

The study of one-shot pricing and product quality games might lead one to conclude that even tacit collusion is impossible. This is not true because competitors often interact on a continuous basis. In such circumstances, firms are said to be involved in repeat games.

When a competitive game is repeated over and over, firms receive sequential payoffs that shape current and future strategies. For example, in Table 11.2, both Coca-Cola and Pepsi Cola might tacitly or secretly agree to charge regular prices so long as the other party continues to do so. If neither firm cheats on such a collusive agreement, discounts will never be offered, and maximum profits will be earned. Although there is an obvious risk involved with charging regular prices, there is also an obvious cost if either or both firms offer discount pricing. If each firm is convinced that the other will maintain regular prices, both will enjoy high profits. This resolve is increased if each firm is convinced that the other will quickly match any discount pricing strategy. In fact, it is rational for colluding firms to quickly and severely punish colluding competitors who "cheat" by lowering prices.

However, although it is important to recognize that the repeat nature of competitor interactions can sometimes harm consumers, it is equally important to recognize that repetitive interactions in the marketplace provide necessary incentives for firms to produce high-quality goods. In any one-shot game, it would pay firms with high-quality reputations to produce low-cost or inferior quality goods. Both Coca-Cola and Pepsi-Cola have well-deserved reputations for providing uniformly high quality soft drinks. They have both invested millions of dollars in product development and quality control to ensure that consumers can depend upon the taste, smell, and feel of Coca- Cola and Pepsi-Cola products. Moreover, because the value of millions of dollars spent on brand-name advertising would be lost if product quality were to deteriorate, that brand name advertising is itself a type of quality assurance provided to customers of Coca-Cola and Pepsi-Cola.

FINITELY REPEATED GAMES

Finitely repeated games have limited duration.

Trigger Strategy is a system of behavior that remains the same until another player takes some course of action that gives rise to a different response.

Table 2: Supplier Strategy Game

	Intel		
	Supply Strategy	Intel Outside	Intel Inside
Dell	Don't Buy	($1b,$1b)	($2b,$3b)
	Buy	(-$3b,$8b)	($3b,$5b)

FINITELY REPEATED GAMES

A Finitely Repeated Game is one that occurs only a limited number of times, or has limited duration of time. Suppose Dell and Intel has agreed to have customer-supplier relationship strategy, as shown in Table 2. At present, Dell computers are marked with the logo "Intel Inside", indicating that Intel supplies Dell with microprocessors. Assume that Dell agrees to use Intel microprocessors in its computers sp long as Intel agrees not to market its own Intel brand of computers. If Intel breaks this supply agreement, it is understood that Dell will punish Intel by thereafter stop doing business with Intel. This means that Intel cheating will trigger a "do not buy" response from Dell in every future period.

So long as the supply agreement is not violated by both parties, Dell will earn $3b per year and Intel will earn $5b per year. If Intel breaks the supply agreement, Intel would earn $8b for one period while its rival Dell would suffer a loss of $3b. However, this one-time benefit for Intel is out-weighed by the loss forever of the $5b per period benefit that would have been earned from maintaining the supply agreement with Dell. Therefore, on account of trigger strategies, it can be ensured that the cost of breaking agreements exceeds any resulting benefits, where both costs and benefits are measured in present value terms.

REPEATED GAMES AND TIT-FOR-TAT STRATEGY

We have seen how two firms facing the prisoners' dilemma can increase their profit by cooperating. Such cooperation however, is not likely to occur in the type of prisoners' dilemma games discussed until now, which are played only once (i.e., that involve a single move or action by each player). Cooperation is more likely to occur in repeated games, or games involving many consecutive moves by each player. These types of games are more realistic in the real world. For example, oligopolists do not decide on their pricing strategy only once but many times over many years.

In repeated games (i.e., in games involving many consecutive moves and countermoves by each player), the best strategy for each player is tit-for-tat. Tit-for-tat behavior can be summarized as follows: Do to your opponent what he has just done to you. That is, you begin by cooperating and continue to cooperate as long as your opponent cooperates. If he betrays you, the next time you betray him back. If he then, cooperates, the next time you also cooperate. This strategy is retaliatory enough to discourage non-cooperation but forgiving enough to allow a pattern of mutual cooperation to develop. In computer simulation as well as, in actual experiments, a tit-for-tat behavior was found to be consistently the, best strategy (i.e., the one that result d in the largest benefit) for each player over time.

THREAT, COMMITMENTS, AND CREDIBILITY

Oligopolistic firms often adopt strategies to gain a competitive advantage over their rivals even if it means temporarily reducing their own profits. For example, an oligopolist may threaten to lower its prices if its rivals lower theirs, even if this means reducing its own profits.

One way to make this threat credible is for firm A to develop a reputation for carrying out its threats even at the expense of profits. This may seem irrational. However, if firm A, actually carries out its threat several times, it would earn a reputation for making credible threats, and this is likely to induce firm B to also charge a high price, thus possibly leading to higher profits for firm A in the long run. In that case, firm A would earn a profit of 5 and firm B a profit of 3 (the bottom right cell) as opposed to a profit of 3 for firm A and 4 for firm B (the bottom left cell). Note that even if firm B earns a profit of 3 by charging the high price (as compared with a profit of 4 by charging the low price), this is still higher than the profit of 2 that it would earn if firm A carries out the threat of charging the low price if firm B does (the top left cell of the table). By showing a commitment to carry out its threats, firm A makes its threats credible and increases its profits over time.

ENTRY DETERRENCE

One important strategy that an oligopolist can use to deter market entry is to threaten to lower its price and thereby impose a loss on the potential entrant. Such a threat, however, works only if it is credible. Entry deterrence can be examined with the payoff matrices of Tables 3 and 4.

The payoff matrix of Table 3 shows that firm A's threat to lower its price is not credible and does not discourage firm B from entering the market. The reason is that firm A earns a profit of 4 if it charges the low price and a profit of 7 if it charges the high price. Unless firm A makes a

credible commitment to fight entry even at the expense of profits, it would not deter firm B from entering the market. Firm A could make a credible threat by expanding its capacity before it is needed. The new payoff matrix might then look like the one indicated in Table 4.

The payoff matrix of Table 4 is the same as in, Table3, except that firm A's profits are now lower when it charges a high price because idle or excess capacity increases firm A's costs without increasing its sales. On the other hand, in the payoff matrix of Table 4, we assume that charging a low price would allow firm A to increase sales and utilize its newly built capacity so that costs and revenues increase, leaving firm A's profits the same as in Table 3. Building excess capacity in anticipation of future needs now becomes a credible threat because with excess capacity firm A will charge a low price and earns a profit of 4 instead of a profit of 3 if it charged the high price. By now charging a low price, however firm B would incur a loss of 2 if it entered the market, and so firm B would stay out of the market. Entry deterrence is now credible and effective.

ENTRY DETERRENCE
TABLE 3

		Firm B	
		Enter	Do Not Enter
Firm A	Low Price	(4, -2)	(6, 0)
	High Price	(7, 2)	(10, 0)

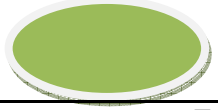
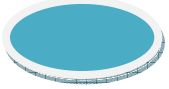


TABLE 4

		Firm B	
		Enter	Do Not Enter
Firm A	Low Price	(4, -2)	(6, 0)
	High Price	(3, 2)	(8, 0)



Lesson 33

PRICING PRACTICES

Price Discrimination: Meaning and Conditions

Price discrimination refers to the charging of different prices for different quantities of a product, at different times, to different customer groups or in different markets, when these price differences are not justified by cost differences. The incentive for this is that the firm can increase its total revenue and profits for a given level of sales and total costs by practicing price discrimination.

Relevant examples of price discrimination are (1) power (i.e., electrical and gas) companies charging lower prices to residential than to commercial users; (2) medical and legal professions charging lower fees to low-income than to high-income people; (3) companies charging lower prices abroad than at home for a variety of products and services, ranging from books and medicines to movies; (4) entertainment companies charging lower prices for afternoon than for evening performances of movies, theaters, and sports events; (5) service industries charging lower prices for children and the elderly for public transportation, and airline tickets; (6) hotels charging lower rates for seminars, workshops and conventions.

It should be remembered, however, that price differences based on cost differences in supplying a product or service in different quantities, at different times, to different customer groups, or in different markets, are not forms of price discrimination. To be price discrimination, the price differences must not be based on cost differences. Also, price discrimination does not have a negative connotation in economics that is, in economics, price discrimination is neutral and benefits some and harms others, and as such, it is not easy to determine whether, on balance, it is beneficial or harmful for society as a whole.

Three conditions must be met for a firm to be able to practice price discrimination.

1. Firm must be an imperfect competitor (a price maker)
2. Price elasticity must differ for units of the product sold at different prices. Price elasticity of demand must differ in submarkets.
3. Firm must be able to segment the market and prevent resale of units across market segments

DEGREES OF PRICE DISCRIMINATION

There are three types of price discrimination: first, second, and third degree. By practicing any type of price discrimination, the firm can increase its total revenue and profits by capturing all or part of the consumer's surplus.

1. First degree price discrimination creates different prices for each customer (maximum profits).
2. Second degree price discrimination gives quantity discounts.
3. Third degree price discrimination assigns different prices by customer age, gender, income, location etc.

The underlying motive for price discrimination can be understood using the concept of **consumers' surplus**. Consumers' surplus is the value of purchased goods and services above and beyond the amount paid to sellers. Consumers' surplus arises because individual consumers place different values on goods and services. Customers that place a relatively high

value on a product will pay high prices; customers that place a relatively low value on a product are only willing to pay low prices.

FIRST-DEGREE PRICE DISCRIMINATION

- Each unit is sold at the highest possible price
- Firm extracts all of the consumers' surplus
- Firm maximizes total revenue and profit from any quantity sold

Assuming a customer who would pay Rs 1000 for one pair of a favorite brand of shoes might be willing to pay Rs750 for a second pair. A monopolist knowing this could offer to sell that customer one pair at Rs 1000 but a second pair for Rs750. Or the monopolist might make an all or nothing offer to this customer: two pairs for Rs1750 or no sale. Either way the monopolist extracts almost all the consumer surplus from the customer. This is known as first degree price discrimination.

Unlike the simple monopolist who charges every customer the same price, the discriminating monopolist will find it profitable to supply any customer who is willing to pay at least marginal cost. This means that output will be at the efficient level! With price discrimination in the first degree there is no efficiency loss from under production, however the monopolist extracts almost all the consumer surplus from each customer.

SECOND-DEGREE PRICE DISCRIMINATION

- Charging a uniform price per unit for a specific quantity, a lower price per unit for an additional quantity, and so on
- Firm extracts part, but not all, of the consumers' surplus

Figure 1

In the absence of price discrimination, a firm that charges $2 and sells 40 units will have total revenue equal to $80.

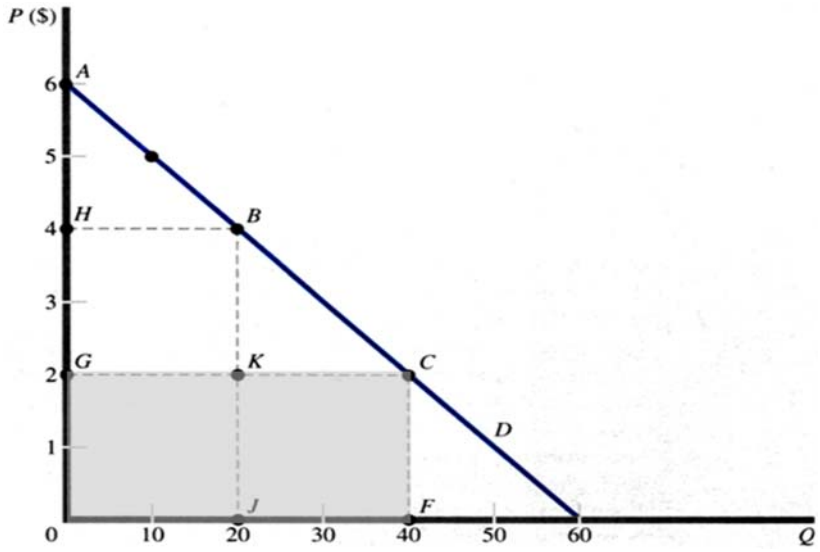


Figure 2

In the absence of price discrimination, a firm that charges $2 and sells 40 units will have total revenue equal to $80.

Consumers will have consumers' surplus equal to $80.

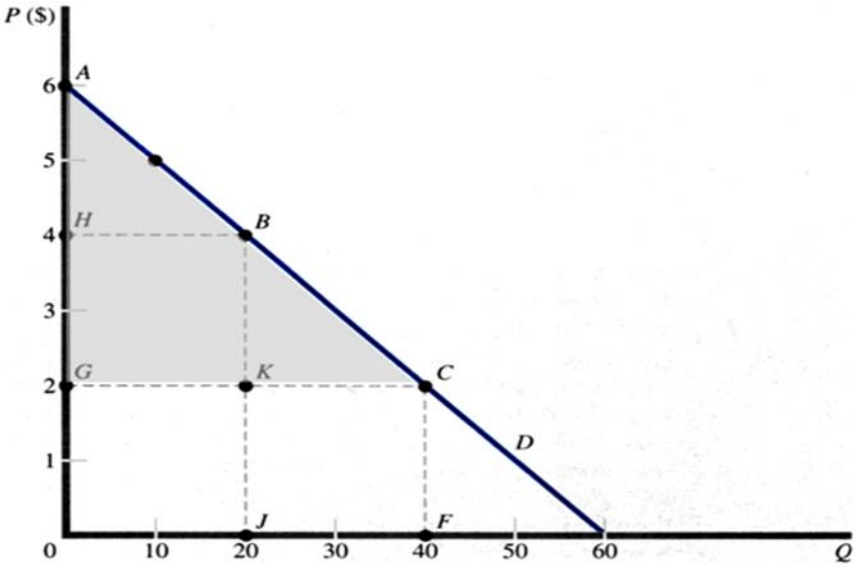
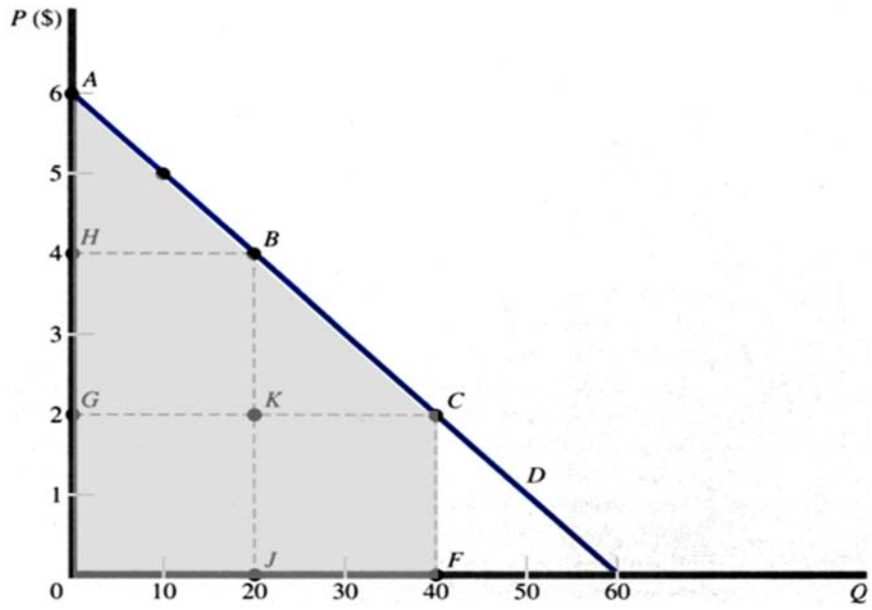


Figure 3

If a firm that practices first-degree price discrimination charges $2 and sells 40 units, then total revenue will be equal to $160 and consumers' surplus will be zero.



ELECTRICITY CONSUMER BILL: LESCO

Second-Degree Price Discrimination

Total units consumed = 770

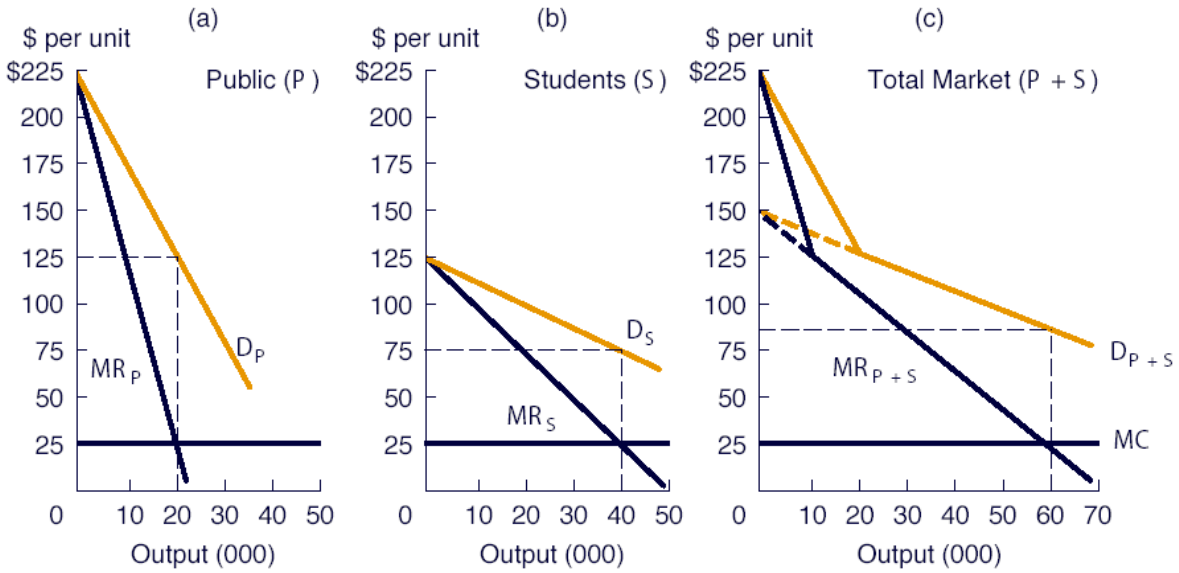
- **100 Units 4.20/unit**
- **200 6.34 /unit**
- **400 10.24/unit**
- **above 700 Units 12.77/unit**

THIRD-DEGREE PRICE DISCRIMINATION

- Charging different prices for the same product sold in different markets
- Firm maximizes profits by selling a quantity on each market such that the marginal revenue on each market is equal to the marginal cost of production

MR1 = MR2 = MC

Figure 4



- Price/Output Determination
  - To maximize profits, set MR = MC in each market.
- One-price Alternative
  - Without price discrimination, MR = MC for all customers as a group.
  - With price discrimination, MR = MC for each customer or customer group.

Profitable price discrimination benefits sellers at the expense of some customers

GRAPHIC ILLUSTRATION

The XYZ University pricing problem and the concept of price discrimination can be illustrated graphically. Figure 4 shows demand curves for the general public in part (a) and for students in part (b). The aggregate demand curve in part (c) represents the horizontal sum of the quantities demanded at each price in the public and student markets. The associated marginal revenue curve, *MRP+S*, has a similar interpretation. For example, marginal revenue equals $25 at an attendance level of 20,000 in the public market and $25 at an attendance level of 40,000 in the student market. Accordingly, one point on the total marginal revenue curve represents output of 60,000 units and marginal revenue of $25. From a cost standpoint, it does not matter whether tickets are sold to the public or to students. The single marginal cost curve *MC* = $25 applies to each market. Graphically solving this pricing problem is a two-part process.

The profit-maximizing total output level must first be determined, and then this output must be allocated between submarkets. Profit maximization occurs at the aggregate output level at which marginal revenue and marginal costs are equal. Figure 4(c) shows a profit-maximizing output of 60,000 tickets, where marginal cost and marginal revenue both equal $25. Proper allocation of total output between the two submarkets is determined graphically by drawing a horizontal line to indicate that $25 is the marginal cost in each market at the indicated aggregate output level. The intersection of this horizontal line with the marginal revenue curve in each submarket indicates the optimal distribution of sales and pricing structure. In this example, profits are maximized at an attendance (output) level of 60,000, selling 20,000 tickets to the public at a price of $125 and 40,000 tickets to students at a price of $75.

PRICE DISCRIMINATION EXAMPLE

Suppose that an XYZ University wants to reduce the athletic department's operating deficit and increase student attendance at home football games. To achieve these objectives, a new two-tier pricing structure for season football tickets is being considered. A market survey conducted by the school suggests the following market demand and marginal revenue relations:

Public Demand

Pp = 225 – 0.005Q

TRp =(225 – 0.005Q)Q

MRp = 225 – 0.01Qp

TC = 1,500,000 + 25Q

MC = 25

MRp = MC

225 – $0.01Qp = $25

0.01Qp = 200

Qp = 20,000

And Pp = 225 – 0.005(20,000)

= $125

Student Demand

Ps = 125 – 0.00125Q

TRs = (125 - 0.00125Q)Q

MRs = 125 – 0.0025Qs

TC = 1,500,000 + 25Q

MC = 25

From these market demand and marginal revenue curves, it is obvious that the general public is willing to pay higher prices than are students. The general public is willing to purchase tickets up to a market price of $225, above which point market demand equals zero. Students are willing to enter the market only at ticket prices below $125.

MRs = MC

125 – 0.0025Qs = $25

0.0025Qs = 100

Qs = 40,000

And Ps = $125 – $0.00125(40,000)

= $75

The ABC Arts Council total operating surplus (profit) is

Operating Surplus (Profit) = TRp + TRs – TC

= $125(20,000) + $75(40,000)– 1,500,000 – $25(60,000)

= $2.5 million

To summarize, the optimal price/output combination with price discrimination is 20,000 in unit sales to the general public at a price of $125 and 40,000 in unit sales to students at a price of $75. This two-tier pricing practice results in an optimal operating surplus (profit) of $2.5 million.

WITHOUT DISCRIMINATION EXAMPLE

Qp = 45,000 – 200Pp and Qs = 100,000 – 800Ps

Under the assumption *Pp = Ps, total demand (Qt) equals*

Qt = Qp + Qs

= 145,000 – 1,000*P*

and

P = $145 – $0.001Q

which implies that

*MR = ∂TR/∂*Q *= 145 – 0.002Q*

and

QP = 45,000 – 200(85) *QS = 100,000 – 800(85)*

= 28,000 = 32,000

Operating surplus (profit) = *TR – TC*

= 85(60,000) – 1,500,000 – 25(60,000)

= $2.1 million

We might notice that the total number of tickets sold equals 60,000 under both the two-tier and the single-price policies. This results because the marginal cost of a ticket is the same under each scenario. Ticket-pricing policies featuring student discounts increase student attendance from 32,000 to 40,000 and maximize the football program's operating surplus at $2.5 million (rather than $2.1 million). It is the preferred pricing policy when viewed from XYZ University perspective. However, such price discrimination creates both "winners" and "losers." Winners, in case of price discrimination are the students and XYZ. Losers include members of the general public, who pay higher football ticket prices or find themselves priced out of the market.

Lesson 34

PRICING PRACTICES (CONTINUED 1)

PRICING OF MULTIPLE PRODUCTS

Most modem firms produce a variety of products rather than a single product. This means that we have to consider demand and product interdependencies. We would examine the firm's pricing of multiple products with interdependent demands, plant capacity utilization and optimal product pricing, and the optimal pricing of joint products produced in fixed or in variable proportions.

DEMAND INTERRELATIONS

The products sold by a firm may be interrelated as substitutes or complements. Demand interrelationships influence the- pricing decisions of a multiple product firm through their effect on marginal revenue. For a two product (A and B) firm, the marginal revenue functions of the firm are

$$MR_A = \frac{\Delta TR_A}{\Delta Q_A} + \frac{\Delta TR_B}{\Delta Q_A}$$

$$MR_B = \frac{\Delta TR_B}{\Delta QB} + \frac{\Delta TR_A}{\Delta Q_B}$$

From the two equations above, we see that the marginal revenue for each product has two components, one associated with the change in the total revenue from the sale of the product itself, and the other associated with the change in the total revenue from the other product. The second term on the right hand side of each equation, therefore, reflects the demand interrelationships. For example, the term $(\Delta TR_B) / (\Delta Q_A)$ in Equation measures the effect on the firm's revenues from product B resulting from the sale of an additional unit of product A by the firm. Similarly, $(\Delta TR_A) / (\Delta Q_B)$ in Equation measures the effect on the firm's total revenue from product A resulting from the sale of an additional unit of product B by the firm.

Cross-marginal revenue terms that reflect demand interrelations can be positive or negative. For complementary products, the net effect is positive because increased sales of one product lead to increased revenues from another. For substitute products, increased sales of one product reduce demand for another; and the cross-marginal revenue term is negative. Accurate price determination in the case of multiple products requires a complete analysis of pricing decision effects. If the second term on the right hand side of each equation is positive, indicating that increased sales of one product stimulates sales of the other the two products are complementary. If, on the other hand, the second term in each equation is negative, indicating that increased sales of one product leads to reduced sales of the other, the two products are substitutes.

Optimal pricing and output decisions on the part of the firm, therefore, require that the total effect (i.e., the direct as well as the cross-marginal effects) of the change in the price of a product on the firm be taken into consideration. If the firm fails to do so, it will lead to suboptimal pricing and output decisions.

PRODUCTION INTERRELATIONS

Many products are related to one another through demand relationships, while others are related in terms of the production process. A by-product is any output that is usually produced as a direct result of an increase in the production of some other output. Multiple products are produced in variable proportions for a wide range of goods and services. In the refining process for crude oil, gasoline, diesel fuel, heating oil, and other products are produced in variable proportions. The cost and availability of any single by-product depends on the demand for others. By-products are also sometimes the unintended or unavoidable results of producing certain goods. For example pollution can be thought of as the necessary by-product of many production processes. Many agricultural products are jointly produced in a fixed ratio. Wheat and straw, beef and hides, milk and butter are all produced in relatively fixed proportions. In mining, gold and copper, silver and lead, and other precious metals and minerals are often produced jointly in fixed proportions.

JOINT PRODUCTS

Joint Products in Variable Proportions

- If products are produced in variable proportions, treat them as distinct products.
- For joint products produced in variable proportions, set $MR_A = MC_A$ and $MR_B = MC_B$.
- Allocation of common costs is wrong and illogical.

Joint Products in Fixed Proportions

- Some products are produced in a fixed ratio.
- If $Q = Q_A = Q_B$, set $MR_Q = MR_A + MR_B = MC_Q$.

Joint or Common Costs

Costs that are shared in the manufacturing and marketing of two or more products in a product line.

JOINT PRODUCT PRICING RULES

Joint Products without Excess By-product

- Profit-maximization requires setting $MR_Q = MR_A + MR_B = MC_Q$.
- Marginal revenue from each byproduct makes a contribution toward covering MC_Q.

Joint Production with Excess By-product

- Profit-maximization requires setting $MR_Q = MR_A + MR_B = MC_Q$.
- Primary product marginal revenue covers MC_Q.
- Byproduct MR = MC = 0.

Plant Capacity Utilization

A multi-product firm using a single plant should produce quantities where the marginal revenue (MR_i) from each of its k products is equal to the marginal cost (MC) of production.

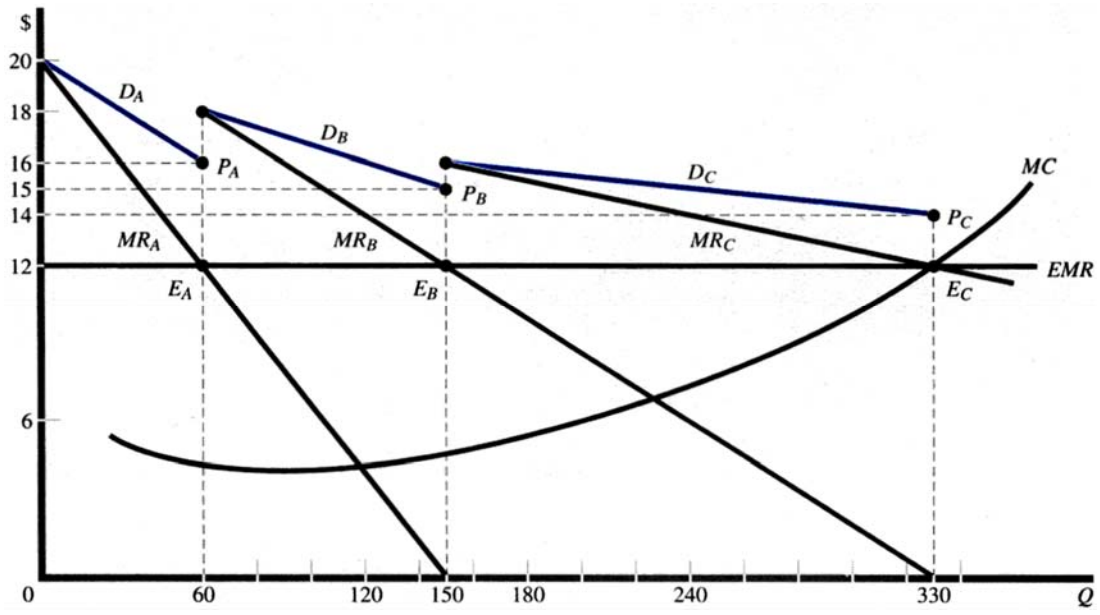
$$MR_1 = MR_2 = \cdots = MR_k = MC$$

One important reason that firms produce more than one product's to make fuller use of their plant and production capacities. A firm that would have idle capacity after producing the best level of output of a single product can search for other products to produce so as to make fuller use of its plant and production capacity. As long as the marginal revenue from these products exceeds their marginal cost, the profits of the firm will increase. Thus, instead of producing a single product at the point where MR = MC and be left with a great deal of idle capacity, the firm

will introduce new products or different varieties of existing products, in the order of their profitability, until the marginal revenue of the least profitable product produced equals its marginal cost to the firm. The quantity produced of the more profitable products is then determined by the point at which their marginal revenue equals the marginal revenue and marginal cost of the last unit of the least profitable product produced by the firm. The price of each product is then determined on its respective demand curve.

Figure1 shows the situation of a firm selling three products (A, B, and C) with respective demand curves D_A, D_B, and D_C, and corresponding marginal revenue, curves MR_A, MR_B, and MR_C. The firm maximizes profits when it produces the quantity of each product at which, $MR_A = MR_B = MR_C = MC$. This is shown by points E_A, E_B, and E_C, at which the equal marginal revenue (EMR) line from the level, at which $MR_C = MC$ crosses the MR_A, MR_B, and MR_C curves. In order to maximize profits, the firm should produce 60 units of product A and sell them at the price of P_A = \$16 on the D_A curve; 90 units of product B (the horizontal distance between points E_B and E_A, or 150 - 60) and sell them at P_B = \$15 on the D_B curve; and 180 units of product C (from 330 - 150) and sell them at P_C = \$14 on the D_C curve. Note that each successive demand curve is more elastic and that the price of each successive product introduced is lower, while its marginal cost is higher so that per unit profits decline.

Figure 1
Plant Capacity Utilization



JOINT PRODUCTS PRODUCED IN VARIABLE PROPORTIONS

Firms can often vary the proportions in which joint products are created. Even the classic example of fixed proportions in the joint production of beef and hides holds only over short periods. The marginal cost of either joint product produced in variable proportions equals the increase in total costs associated with a one-unit increase in that product, holding constant the quantity of the other joint product produced. Optimal price/output determination for joint products in this case requires a simultaneous solution of marginal cost and marginal revenue relations. The firm maximizes profit by operating at the output level where the marginal cost of producing each joint product just equals the marginal revenue it generates. The profit-maximizing combination of joint products *A* and *B*, for example, occurs at the output level where

MRA* = *MCB* and *MRB* = *MCB

Although it is possible to determine the separate marginal costs of goods produced in variable proportions, it is impossible to determine their individual average costs. This is because **common costs** are expenses necessary for manufacture of a joint product. Common costs of production—raw material and equipment costs, management expenses, and other overhead—cannot be allocated to each individual by-product on any economically sound basis. Only costs that can be separately identified with a specific by-product can be allocated. For example, tanning costs for hides and refrigeration costs for beef are separate identifiable costs of each by-product. Feed costs are common and cannot be allocated between hide and beef production. Any allocation of common costs is wrong and illogical.

JOINT PRODUCTS PRODUCED IN FIXED PROPORTIONS

Products that must be produced in fixed proportions should be considered as a package or bundle of output. When by-products are jointly produced in fixed proportions, all costs are common, and there is no economically sound method of cost allocation. Optimal price/output determination for output produced in fixed proportions requires analysis of the relation between marginal revenue and marginal cost for the combined output package. As long as the sum of marginal revenues obtained from all by-products is greater than the marginal cost of production, the firm gains by expanding output.

Figure 2
Joint Products in Fixed Proportions

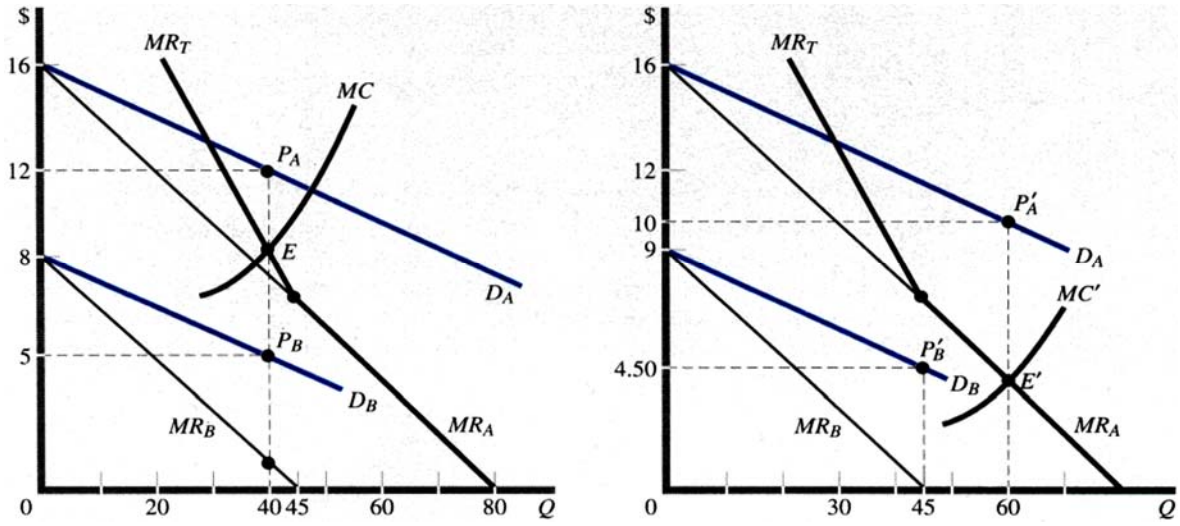


Figure 2 illustrates the pricing problem for two products produced in fixed proportions. Demand and marginal revenue curves for each by-product and the single marginal cost curve for production of the combined output package are shown. Vertical summation of the two marginal revenue curves indicates the total marginal revenue generated by both by-products. Marginal revenue curves are summed vertically because each unit of output provides revenues from the sale of both by-products. The intersection of the total marginal revenue curve MR_T with the marginal cost curve identifies the profit-maximizing output level. The optimal price for each by-product is determined by the intersection of a vertical line at the profit-maximizing output level with each by-product's demand curve.

An example of joint production in fixed proportions is cattle rising, which provides both beef and hides in the ratio of one-to-one. An example of joint production with variable proportions is provided by petroleum refining, which results in gasoline, fuel oils, and other products in proportions which, within a range, can be varied by the firm. Such production interdependence must be considered by the firm in order to reach optimal output and pricing decisions.

When products are jointly produced in fixed proportions, they should be thought of as a single production package. There is then no rational way of allocating the cost of producing the package to the individual products in the package. For example, the cost of raising cattle cannot be allocated in any rational way to beef and hides, since they are jointly produced. On the other hand, the jointly produced products may have independent demands and marginal revenues. For example, the demand and marginal revenue for beef are separate and independent of the demand and marginal revenue for hides. The best level of output of the joint product is then determined at the point where the vertical summation of the marginal revenues of the various jointly produced products equals the single marginal cost of producing the entire product package. This is shown in Figure 2.

In the left panel of Figure 2, D_A and D_B refer, respectively, to the demand curves of products A and B, which are jointly produced in the proportion of one-to-one. We could think of product A as beef and product B as hides which result in the ratio of one-to-one from the slaughter of each cow. Thus, the horizontal axis of the figure measures at the same time the quantity (Q) of cattle, beef, and hides. Despite the fact that beef and hides are jointly produced, their, demand curves are independent because they are unrelated in consumption. The corresponding marginal revenue curves are MR_A and MR_B in the figure. The total marginal revenue (MR_T) curve is obtained by summing vertically the MR_A and MR_B curves because the firm receives marginal revenues from the sale of both products. Note that starting at the output level of Q = 45 units, at which $MR_B = 0$, the MR_T curve coincides with the MR_A curve. The best level of output of both beef and hides is 40 units and is given by point E, at which the MC curve for cattle (both beef and hides together) crosses the MR_T curve of the firm. At Q = 40, P_A = \$12 on the D_A curve and P_B = \$5 on the D_B curve.

In the left panel of Figure, both MR_A and MR_B are positive at the best level of output of Q = 40. In contrast, in the right panel of Figure MR_B is negative at the best level of output of Q = 60 given by point E, at which the lower MC' curve crosses, the same MR_T curve. This means that selling more than 45 units of product B (hides) reduces the firm's total revenue and profits. In such a case the firm -produces 60 units of the joint product (cattle), sells 60 units of product A (beet) at P'_A = \$10 but sells only 45 units of product B (hides) at P'_B = \$4.50 (at which TR_B is maximum and $MR_B = 0$). That is, the firm withholds from the market and dispose & of the extra 15 units of product B jointly produced with the 60 units of product A in order not to sell, them at a negative marginal revenue. An example of this was provided by the destruction of excess pineapple juice that jointly resulted from the production of sliced pineapples for canning. Until use was found for it, the excess pineapple juice was simply destroyed or otherwise held off the market, in order not to depress its price below the point at which its marginal revenue became negative. However, seeing a profit-making opportunity, some producers advertised heavily to shift the demand curve for pineapple juice outward. New products were also created, such as pineapple-grapefruit juice, to increase demand for the waste by-product i.e. pineapple juice.

JOINT PRODUCT PRICING EXAMPLE

Joint Products without Excess By-Product

A Paper Company produces newsprint and packaging materials in a fixed 1:1 ratio, or 1 ton of packaging materials per 1 ton of newsprint.

TC = $2,000,000 + $50Q + $0.01Q2
MC = ∂TC/∂Q = $50 + $0.02Q

Newsprint

Pa = $400 – $0.01Qa
MRa = ∂TR/∂Qa
= $400 – $0.02Qa

Packaging Materials

Pb = $350 – $0.015Qb
MRb = ∂TRb/∂Qb
= $350 – $0.03Qb
TR = TRa + TRb = PaQa + PbQb

Substituting for *Pa and Pb results in the total revenue function*

TR = ($400 – $0.01Qa)Qa + ($350 – $0.015Qb)Qb
= $400*Qa – $0.01Qa²* + $350*Qb – $0.015Qb²*

Because one unit of product *A and one unit of product B are contained in each unit of Q, Qa = Qb = Q. This allows substitution of Q for Qa and Qb to develop a total revenue function in* terms of *Q, the unit of production:*

TR = $400Q – $0.01Q² + $350Q – $0.015Q²
= $750*Q – $0.025Q²*

The profit-maximizing output level is found by setting *MR = MC and solving for Q:*

MR = MC
$750 – $0.05*Q = $50 + $0.02Q*
0.07*Q = 700*
Q = 10,000 units

At the activity level *Q = 10,000 units, marginal revenues for each product are positive:*

MRa = $400 – $0.02Qa *MRb = $350 – $0.03Qb*
= $400 – $0.02(10,000) = $350 – $0.03(10,000)
= $200 (at 10,000 Units) = $50 (at 10,000 Units)

Each product makes a positive contribution toward covering the marginal cost of production,

Where
MC = $50 + $0.02Q
= $50 + $0.02(10,000)
= $250

Pa = $400 – $0.01Qa *Pb = $350 – $0.015Qb*
= $400 – $0.01(10,000) = $350 – $0.015(10,000)
= $300 = $200
Pa = $400 – $0.01Qa *Pb = $350 – $0.015Qb*
= $400 – $0.01(10,000) = $350 – $0.015(10,000)
= $300 = $200

π = PaQa + PbQb – TC
= $300(10,000) + $200(10,000) – $2,000,000
– $50(10,000) – $0.01(10,0002)

= $1,500,000 = $1.5 million

The Paper Company should produce 10,000 units of output and sell the resulting 10,000 units of product A at a price of $300 per ton and 10,000 units of product B at a price of $200 per ton.

Joint Production with Excess By-Product

Suppose that an economic recession causes the demand for product *B (packaging materials)* to fall dramatically, while the demand for product *A (newsprint) and marginal cost conditions* hold steady. Assume new demand and marginal revenue relations for product *B are:*

Pb = $290 – $0.02Qb

MRb = ∂TRb/∂Qb

= $290 – $0.04Qb

MR = MRa + MRb

= $400 – $0.02*Qa + $290 – $0.04Qb*

= $690 – $0.06*Q*

If all production is sold, the profit-maximizing level for output is found by setting *MR = MC* and solving for *Q:*

MR = MC

$690 – $0.06*Q = $50 + $0.02Q*

0.08*Q = 640*

Q = 8,000

At *Q = 8,000, the sum of marginal revenues derived from both by-products and the marginal* cost of producing the combined output package each equal $210, because

MR = $690 – $0.06Q *MC = $50 + $0.02Q*

= $690 – $0.06(8,000) = $50 + $0.02(8,000)

= $210 = $210

However, the marginal revenue of product *B is no longer positive:*

MRa = $400 – $0.02Qa *MRb = $290 – $0.04Qb*

= $400 – $0.02(8,000) = $290 – $0.04(8,000)

= $240 = –$30

Even though *MR = MC = $210, the marginal revenue of product B is negative at the Q = 8,000* activity level.

MRa = MC

$400 – $0.02*Q = $50 + $0.02Q*

$0.04*Q = $350*

Q = 8,750 units

MRb = MCb

$290 – $0.04*Qb = $0*

$0.04*Qb = $290*

Qb = 7,250

Optimal prices and the maximum total profit for the Firm are as follows:

Pa = $400 – $0.01Qa *Pb = $290 – $0.02Qb*

= $400 – $0.01(8,750) = $290 – $0.02(7,250)

= $312.50 = $145

π = *PaQa + PbQb – TC*

= $312.50(8,750) + $145(7,250) – $2,000,000

– $50(8,750) – $0.01(8, 7502)

= $582,500

Lesson 35

PRICING PRACTICES (CONTINUED 2)

Transfer Pricing: Meaning and Nature of Transfer Pricing

Transfer Pricing is the pricing strategy in which a firm optimally sets the internal price at which an upstream division sells inputs to a downstream division. The transfer pricing problem results from the difficulty of establishing profitable relationships among divisions of a single company when each separate business unit stands in **vertical relation** to the other. A vertical relation is one where the output of one division or company is the input to another. **Vertical integration** occurs when a single company controls various links in the production chain from basic inputs to final output. Media powerhouse AOL-Time Warner, Inc., is vertically integrated because it owns AOL, an Internet service provider (ISP) and cable TV systems, plus a number of programming properties in filmed entertainment (e.g., Warner Bros.) and television production (e.g., HBO, CNN). Similarly we can give the example of Textiles producing.

To maximize profits for the vertically integrated firm, it is essential that a profit margin or markup only be charged at the final stage of production. All intermediate products transferred internally must be transferred at marginal cost. Decentralization and the establishment of semiautonomous profit centers also gave rise to the need for transfer pricing, or the need to determine the price of intermediate products sold by one semiautonomous division of a large-scale enterprise and purchased by another semiautonomous division of the same enterprise. For example, if a steel company owned its own coal mine, the questions would arise as to how much coal the coal mine should sell to the parent steel company and how much to outsiders, and at what prices. Similarly, the parent steel company must determine how much coal to purchase from its own coal mine and how much from outsiders, and at what prices.

To simplify our discussion, we assume throughout that the firm has two divisions, a production division (indicated by the subscript p) and a marketing division (indicated by the subscript m). The production division sells the intermediate product to the marketing division, as well as to outsiders, if an outside market for the intermediate product exists. The marketing division purchases the intermediate product from the production division, completes the production process, and markets the final product for the firm. Also, to simplify the presentation, we will assume throughout that 1 unit of the transfer or intermediate product is required to produce each unit of the final product sold by the marketing division.

- The transfer price creates revenues for the selling sub-unit and purchase costs for the buying sub-unit, affecting each sub-unit's operating income
- Intermediate Product – the product or service transferred between sub-units of an organization

EFFECTS OF TRANSFER PRICES

- Reallocate total company profits among business segments
- Influence decision making by purchasing, production, marketing, and investment managers

Transfer Pricing Problem

- Pricing transfer of products among divisions of a single firm can become complicated.

Products without External Markets

- Marginal cost is the appropriate transfer price.

Products with Competitive External Markets

- Market price is the optimal transfer price.

Products with Imperfectly Competitive External Markets

- Optimal transfer price is the marginal revenue derived from combined internal and external markets.

TRANSFER PRICING FOR PRODUCTS WITHOUT EXTERNAL MARKETS

Supply is offered by various upstream suppliers to meet the demand of downstream users. Goods and services must be transferred and priced each step along the way from basic raw materials to finished products. An effective transfer pricing system leads to activity levels in each division that are consistent with profit maximization for the overall enterprise. This observation leads to the most basic rule for optimal transfer pricing: When transferred products cannot be sold in external markets, the marginal cost of the transferring division is the optimal transfer price.

When there is no external demand for the intermediate product, the production division can sell the intermediate product only internally to the marketing division of the firm, and the marketing division can purchase the intermediate product only from the production division of the firm. Since 1 unit of the intermediate product is used to produce each unit of the final product, the output of the intermediate product and of the final product are equal Figure 1 shows how the transfer price of the intermediate product is determined when there is no external market for the intermediate product.

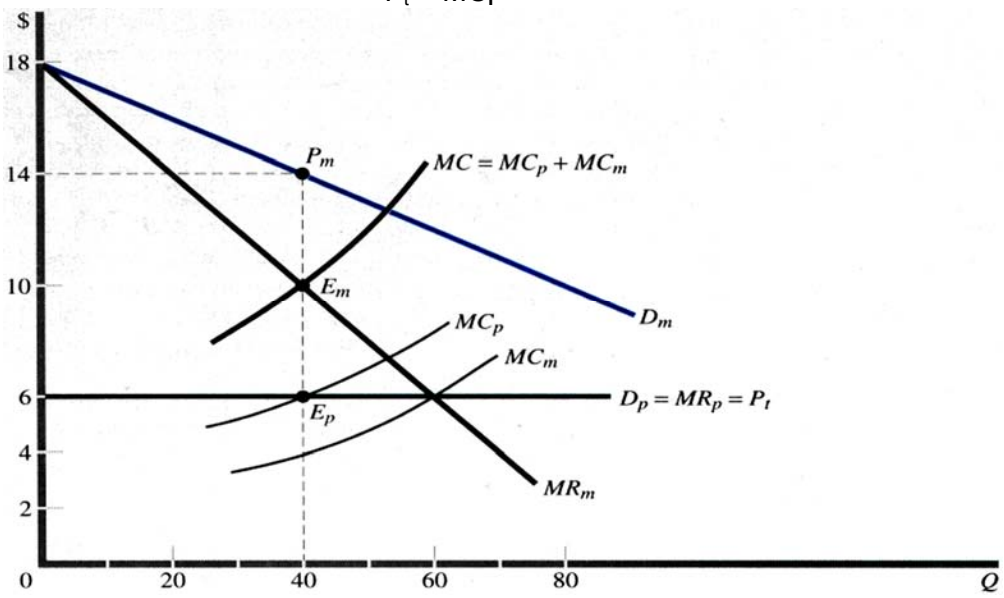
In Figure 1 MC_P and MC_m are the marginal cost curves of the production and marketing divisions of the firm, respectively, while MC is the vertical summation of MC_P and MC_m, and it represents the total marginal cost curve for the firm as a whole. The figure also shows the external demand curve for the final product sold by the marketing division, D_m, and its corresponding marginal revenue curve, MR_m. The firm's profit maximizing level of output for the final product is 40 units and is given by point E_m, at which $MR_m = MC$. Therefore, $P_m = \$14$.

Figure 1

Transfer Price = P

MC of Intermediate Good = MC_P

$P_t = MC_P$



Since 40 units of the intermediate product are required (i.e., are demanded by the marketing division of the firm in order to produce the best level of 40 units of the final product), the transfer price for the intermediate product, P_t is set equal to the marginal cost of the intermediate product (MC_P) at $Q_P = 40$. Thus, $P_t = \$6$ and is given by point E_P at which $Q_P = 40$. The demand and marginal revenue curves faced by the production division of the firm are then equal to the transfer price (that is, $D_P = MR_P = P_t$). Note that $Q_P = 40$ is the best level of output of the intermediate product by the production division of the firm because at $Q_P = 40$, $D_P = MR_P = P_t = MC_P = \$6$. Thus, we can conclude that the correct transfer price for an intermediate product for which there is no external market, is the marginal cost of production.

TRANSFER PRICING WITH PERFECTLY COMPETITIVE EXTERNAL MARKETS

The transfer pricing problem is only slightly more complicated when transferred inputs can be sold in external markets. When transferred inputs can be sold in a perfectly competitive external market, the external market price represents the firm's opportunity cost of employing such inputs internally. As such, it would never pay to use inputs internally unless their value to the firm is at least as great as their value to others in the external market. This observation leads to a second key rule for optimal transfer pricing: When transferred products can be sold in perfectly competitive external markets, the external market price is the optimal transfer price. If upstream suppliers wish to supply more than downstream users desire to employ at a perfectly competitive price, excess input can be sold in the external market. If downstream users wish to employ more than upstream suppliers seek to furnish at a perfectly competitive price, excess input demand can be met through purchases in the external market. In either event, an optimal amount of input is transferred internally.

When an external market for the intermediate product does exist, the output of the production division need not be equal to the output of the final product. The transfer price, however, depends on whether or not the external market for the intermediate product is perfectly competitive. The determination of the transfer price when the external market is perfectly

competitive is shown in Figure 2.

Figure 2

Transfer Price = P_t

MC of Intermediate Good = MC'_p

$P_t = MC'_p$

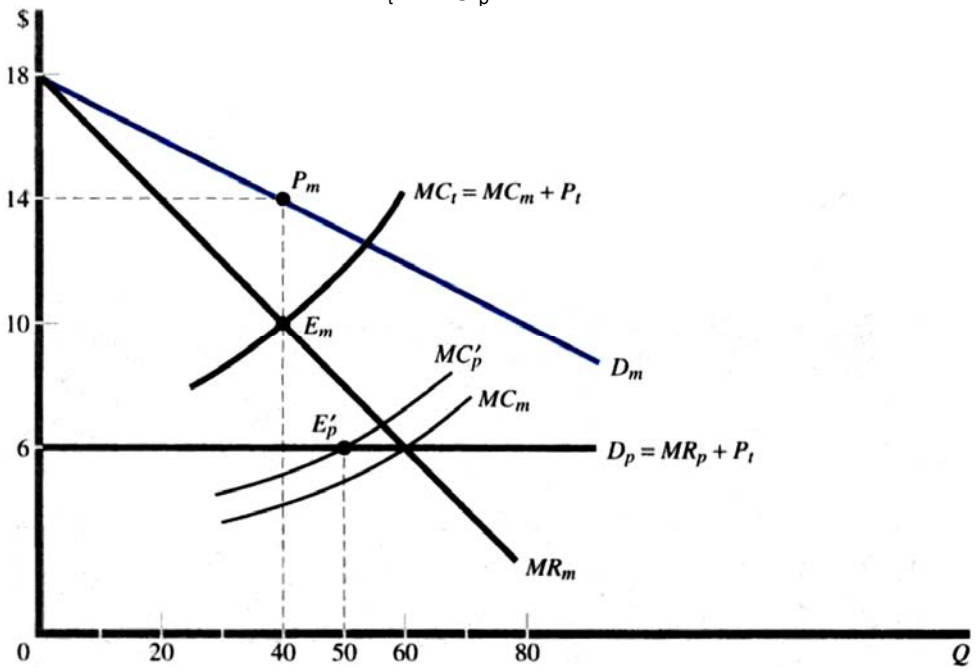


Figure 2 is identical to Figure 1, except that the marginal cost curve of the production division MC'_p is lower than in figure 1. The production division then produces more of the intermediate product than the marketing division demand's and sells the excess in the perfectly competitive external market for the intermediate product. With a perfectly competitive market for the intermediate product, the production division faces horizontal demand curve D_P for its output at the given market price P_t for the intermediate product. Since D_P is horizontal, $D_P = MR_P = P_t$. The profit-maximizing level of output of the intermediate product by the production division of the firm is 50 units and is given by point E'_P at which $D_P = MR_P = P_t = MC'_P = \$6$.

Since the marketing division can purchase the *intermediate* product either internally or externally at $P_t = \$6$, its total marginal cost curve is given by MC_t which is the, vertical sum of its own marginal cost (MC_m) and the price of the intermediate product (P_t). Thus, the best level of, output of the *final* product by the marketing division of the firm is 40 units (the same as when there was no external market for the intermediate product) and is given by point E_m at which $MR_m = MC_t$. At $Q_m = 40$, $P_m = \$14$.

Thus, the production division of the firm produces 50 units of the intermediate product and sells 40 units internally to the marketing division at $P_t = \$6$ and sells the remaining 10 units in the external market, also at $P_t = \$6$. The marketing division will not pay more than the external price of \$6 per unit for the intermediate product, while the production division will not sell the intermediate product internally to the marketing division for less than \$6 per unit. Thus, when a perfectly competitive external market for the intermediate product exists, the transfer price for intra-company sales of the intermediate product is given by the external competitive price for the intermediate product.

The analysis shown graphically in Figure 2 can also be shown algebraically, as follows. The demand and marginal revenue curves for the final product faced by the marketing division in Figure 2 can be represented algebraically as:

$Qm = 180 – 10Pm$ or $Pm = 18 – 0.1Qm$

And $MRm = 18 – 0.2Qm$

$MC'p = 1 + 0.1Qp$ and $MCm = 0.1Qm$

The perfectly competitive external price for the transfer product is Pt = $6, we can find the best level of output of the inter product for the production division by setting its MC = Pt i-e

$MC'p = 1 + 0.1Qp = \$6 = Pt$ so that $0.1Qp = 5$

And $Qp = 50$

The best level of output for the marketing division is determined by finding the total MC of the marketing division (MCt) and setting it equal to its MR, i-e,

$MCt = MCm + Pt = 0.1Qm + 6$

Then $MCt = 0.1Qm + 6 = 18 – 0.2Qm = MRm$

$0.3Qm = 12$ so that $Qm = 40$

And $Pm = 18 – 0.1(40) = \$14$

Thus, the production division sells 40 units of the intermediate product internally to the marketing division and the remaining 10 units on the external competitive market, all at P_t = $6. The marketing division uses the 40 units of the intermediate product purchased internally from the production division at P_t = $6 to produce 40 units of the final product to be sold on the external market at P_m = $14.

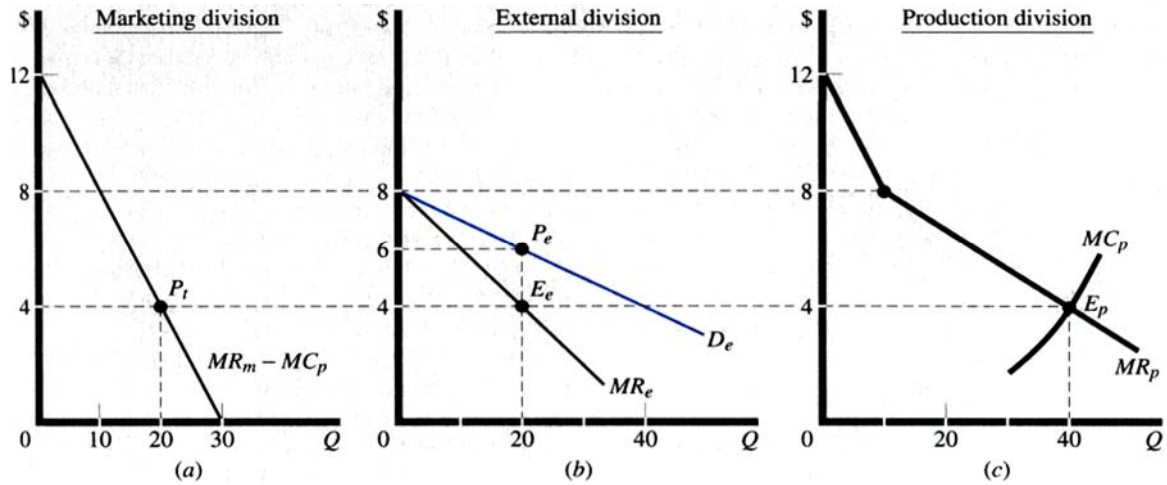
TRANSFER PRICING WITH IMPERFECTLY COMPETITIVE EXTERNAL MARKETS

The typical case of vertical integration involves firms with inputs that can be transferred internally or sold in external markets that are not perfectly competitive. Again, it never pays to use inputs internally unless their value to the firm is at least as great as their value to others in the external market. This observation leads to a third and final fundamental rule for optimal transfer pricing: When transferred products can be sold in imperfectly competitive external markets, the optimal transfer price equates the marginal cost of the transferring division to the marginal revenue derived from the combined internal and external markets.

When an imperfectly competitive external market for the intermediate product exists, the transfer price of the intermediate product for intra firm sales will differ from the price of the intermediate product on the imperfectly competitive external market. The determination of the internal and external prices of the intermediate product by the production division of the firm becomes one of third-degree price discrimination. This is shown in Figure 3.

Figure 3

Transfer Price = P_t = $4 **External Market Price = P_e = $6**



Panel a of Figure 3 shows the marginal revenue of the marketing division of the firm (that is, MR_m) after subtracting from it the transfer price of the intermediate product (P_t), which is equal to the marginal cost of the production division (MC_P). Thus, the MR_m - MC_P curve in the left panel shows the net marginal revenue of the marketing division. Panel b presents the negatively sloped demand curve for the intermediate product of the firm on the imperfectly competitive external market (D_e) and its corresponding marginal revenue curve (MR_e). In panel c, on one hand, the MR_P curve is the total revenue curve of the production division of the firm, which is equal to the horizontal summation of the net marginal revenue curves for internal sales to the marketing division of the firm and to the external market (that is, $MR_P = MR_m - MC_P + MR_e$.

The best level of output of the intermediate product by the production division of the firm is 40 units and is given by point E_P at which $MR_P = MC_P$ in panel c. The optimal distribution of the 40 units of the intermediate product produced by the production division of the firm is 20 units internally to the marketing division of the firm (given by point P_t in panel a) and 20 units to the external market (given by point E_e in panel b), so that $MR_m - MC_p = MR_e = MR_p = MC_p = \$4$. Thus, the production division of the firm operates as the monopolist seller of the intermediate product in the segmented internal and external markets for the intermediate product.

PRICING RULES-OF-THUMB

Competitive Markets

- Profit maximization always requires setting
- $M\pi = MR - MC = 0$, or MR=MC, to maximize profits.
- In competitive markets, P=MR, so profit maximization requires setting P=MR= MC.

Imperfectly Competitive Markets

- With imperfect competition, P > MR, so profit maximization requires setting MR=MC.

Optimal Price

$MR = P[1 + (1/\varepsilon_P)]$

$P[1 + (1/\varepsilon_P)] = MC$

Optimal $P^* = MC/[1 + (1/\varepsilon_P)]$

MARKUP PRICING AND PROFIT MAXIMIZATION

Although profit maximization requires that prices be set so that marginal revenues equal marginal cost, it is not necessary to calculate both to set optimal prices. Just using information on marginal costs and the point price elasticity of demand, the calculation of profit-maximizing prices is quick and easy. Many firms derive an optimal pricing policy using prices set to cover direct costs plus a percentage markup for profit contribution. Flexible markup pricing practices that reflect differences in marginal costs and demand elasticities is considered to be an efficient method for ensuring that *MR* = *MC* for each line of products sold.

Optimal Markup on Cost

- Markup pricing is an efficient means for achieving profit maximization.
- Markup on cost uses cost as a basis.
- Optimal markup on cost = $-1/(\varepsilon_P + 1)$.

Optimal Markup on Price

- Markup on price uses price as a basis.
- Optimal markup on price = $-1/\varepsilon_P$

Markup or Full-Cost Pricing

- Fully Allocated Average Cost (C)
  - Average variable cost at normal output
  - Allocated overhead
- Markup on Cost (m) = (P - C)/C
- Price = P = C (1 + m)

Markup and Demand Elasticity

There is an inverse relationship between markup and demand elasticity. For example, if Ep = -2 then
m = 100%
If Ep = -5 then m = 25%

- More elastic the demand, lower markup.
- Less elastic the demand, higher markup.

Lesson 36

ALTERNATIVE THEORIES OF THE FIRM

A CRITIQUE OF THE NEOCLASSICAL THEORY OF THE FIRM

Traditional theory of the firm assumes profit maximization as the only objective of the business firm. Although the conventional Profit maximization theory of the firm still holds its ground firmly, several alternative theories of the firm were purposed during the early 1960s by economists, particularly by Simon, Baumol, Marris, Williamson, Cyret and March.

THE BASIC ASSUMPTIONS OF THE NEOCLASSICAL THEORY

The basic assumptions of the neoclassical theory of the firm may be outlined as follows.

1. The traditional theory of the firm assumes a single owner-entrepreneur. There is no separation between ownership and management. The owner-entrepreneur takes all the decisions. All organizational problems are assumed resolved by payments to the factors employed by the firm. The entrepreneur is furthermore assumed to have unlimited information, unlimited time at his disposal, and unlimited ability to compare all the possible alternative actions and choose the one that maximizes his profit. This behavior is described by postulating that the entrepreneur acts with global rationality. There is no time, information or other constraints in achieving the single goal of profit maximization.

2. The firm has a single goal, that of profit maximization. The firm is a profit maximize. It is assumed to make as much profit as possible. This means that the model is an optimizing' model. The firm attempts to achieve the best possible performance, rather than simply seeking "feasible" performance which meets some set of minimum criteria. In most situations these are consistent with each other. Shareholder wealth is maximized by selecting the most profitable set of plant and equipment and then operating it in the most profitable way. BUT THERE MAY BE EXCEPTIONS - making maximum short term profit might cause entry of the new firms or government intervention.

3. This goal is attained by application of the marginalist principle. It assumes perfect certainty. Cost and demand conditions are perfectly known.

"MANAGERIAL" CRITICISMS OF THE PROFIT-MAXIMISING MODEL

Berle and Means (1932) were the first to comment:

- firms are owned by shareholders but controlled by managers
- owners' and managers' interests are different
- managers have discretion to use the firm's resources in their own interests

There is some empirical evidence that profits are higher in owner-controlled firms than in firms where management is divorced from ownership. This supports the view that managers have at least some discretion in pursuing goals other than profit maximization. However, the evidence is far from conclusive. In any case such empirical findings do not imply that managers have unlimited discretion, or that the goal of profit maximization is not valid.
The assumptions of profit-maximisation has been criticised in a number of ways; so we have two schools of thought:

1. The "Managerial School"
2. The "Behavioural School"

ALTERNATIVE THEORIES OF THE FIRM

1. **Sales Revenue Maximization Model –** William Baumol
2. **Theory of Managerial Utility –** Oliver Williamson Integrates the growth maximization model and the profit/sales maximization models and the maximization of the present value of future sales
3. **Maximizing Growth** – Robin Marris the managerial utility depends on the firm's rate of growth. Supply led growth vs. Demand led growth, or
4. **Behavioral theories** - Herbert Simon, Richard Cyert and James March Firms – multi-goal, multi-decision, organizational coalitions, Imperfect knowledge and bounded rationality, Managers cannot meet the aspiration levels of all stakeholders, Managers cannot maximize, instead they have to satisfice.

The first three Alternative Theories of the Firm are representing Managerial School of thought while the last one is representing Behavioural School of thought.

SALES MAXIMIZATION MODEL: W.J.BAUMOL

Baumol offers several justifications of sales maximization as a goal of the firm. The separation of ownership from management, characteristic of the modern firm, gives discretion to the managers to chase goals which maximize their own utility and deviate from profit maximization, which is the desirable goal of owners. Given this judgment, Baumol argues that sales maximization seems the most reasonable goal of managers. From his experience as a consultant to large firms Baumol found that managers are worried about maximization of the sales rather than profits. He gave several reasons for this attitude of the top management:

- Firstly, there is evidence that salaries and other (slack) earnings of top managers are correlated more closely with sales than with profits.
- Secondly, the banks and other financial institutions keep a close eye on the sales of firms and are more willing to finance firms with large and growing sales.
- Thirdly, personnel problems are handled more satisfactorily when sales are growing. The employees at all levels can be given higher earnings and better terms of work in general. Declining sales, on the other hand, will make necessary the reduction of salaries and other payments.

Baumol argues that the top managers become to a certain extent risk-avoiders, and this attitude may act as a curb on economic growth. However, the desire for steady performance has a stabilizing effect on economic activity. In general, large firms have research units which develop new ideas of products or techniques of production. The application of these projects is spread over time so as to avoid wide swings in the economic performance of the firm.

Although Baumol recognizes the interdependence of firms the main feature of oligopolistic markets he argues that in day-to-day decision making management often, acts explicitly or implicitly on the basis that its decisions will produce no changes in the behavior of those with whom they are competing……It is only when the firm makes more radical decisions, such as the launching of a major advertising campaign or the introduction of a radically new line of product, only then the management considers the competitive response. But often even in fairly crucial decisions, and almost always in routine policy-making, only the most rapid attention is paid to competitive reactions.

W. J. Baumol suggested sales revenue maximization as an alternative goal to profit maximization. He presented two basic models the first is a static single period model the second is a multi-period dynamic model of growth of sales revenue maximization. Each model has two

versions, one without and one with advertising activities. We will discuss here only the static single period model.

BAUMOL'S STATIC MODELS

The basic assumptions of the static models:

1. The time-horizon of a firm is a single period.

2. During this period the firm attempts to maximize its total sales revenue subject to a profit constraint. Managers seek to maximize sales, after ensuring that an adequate rate of return has been earned, rather than to maximize profits

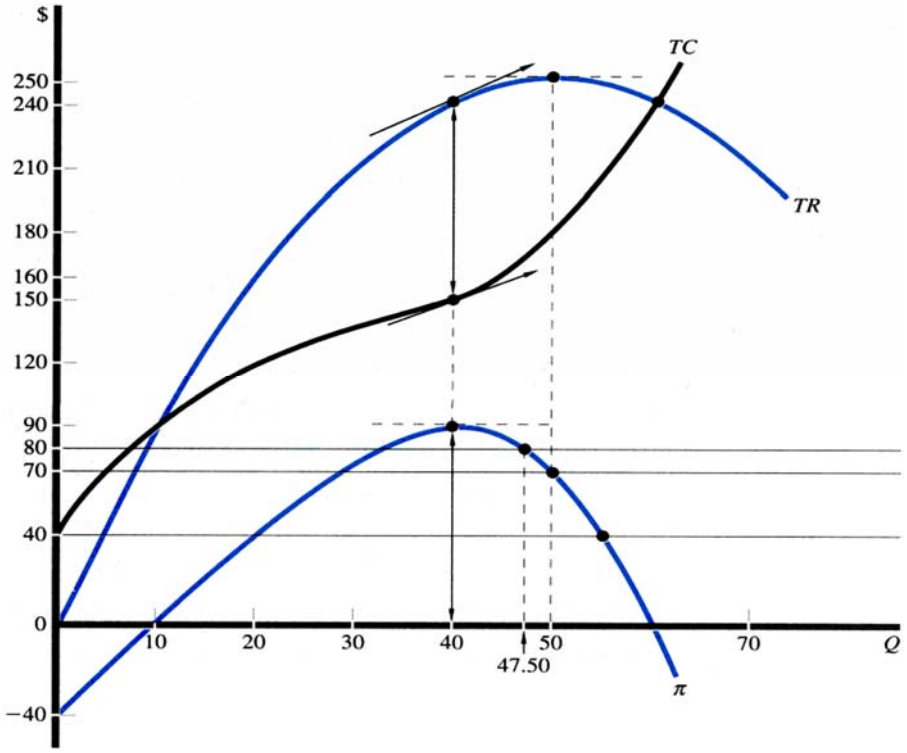
3. The minimum profit constraint is exogenously determined by the demands and expectations of the shareholders, the banks and other financial institutions. The firm must realize a minimum level of profits to keep shareholders happy and avoid a fall of the prices of shares on the stock exchange. If profits are below this exogenously determined minimum acceptable level the managers run the risk of being dismissed, since shareholders may sell their shares and take-over raiders may be attracted by a fall of the prices of shares.

4. Baumol assumes that cost curves are U-shaped and the demand curve of the firm is downward-sloping.

A STATIC MODEL, WITHOUT ADVERTISING

The total-cost and total-revenue curves under Baumol's assumptions are shown in figure 1. Total sales revenue is at its maximum level at the highest point of the TR curve, where the price elasticity of demand is unity and the slope of this TR curve (the marginal revenue) is equal to zero. Sales (or total revenue, TR) will be at a maximum when the firm produces a quantity that sets marginal revenue equal to zero (MR = 0).

Figure 1



In figure 1 TR refers to the total revenue, TC to the total cost, and Π to the total profits of the firm. Π = TR – TC and is maximized at \$90 at Q = 40 units where the positive difference between TR and TC is greatest (i.e. where the TR and TC curves are parallel i.e. where the MR = MC since the tangents are the slopes of the TR and TC curves respectively). On the other hand, TR is maximum at \$250 where Q = 50, at which the slope of the TR curve or MR is zero and Π = \$70. If the firm had to earn a profit of at least \$70 to satisfy the minimum profit constraint, the firm would produce 50 units of output and maximize TR at \$250 and with Π = \$70. The same would be true as long as the minimum profit requirement of the firm was equal to or smaller than \$70. That means, when $\Pi \leq$ \$70, the profit constraint is non-binding. With a minimum profit requirement between \$70 and \$90, however, the profit constraint would be binding or operative. For example to earn a profit of \$80, the firm would have to produce 47.5 units. Finally, if the minimum profit requirement were higher than \$90, all that the firm can do is to produce Q = 40 units and maximize Π at \$90 with TR = \$240.

In short two types of equilibria appear to be possible: one in which the profit constraint provides no effective barrier to sales maximization and one in which it does. In this second type of equilibrium, the firm will produce an output which yields a satisfactory target Π.
Assuming that the profit constraint is operative, the following predictions of Baumol's single-period model (without advertising) emerge:

- The sales maximizer will produce a higher level of output as compared to a profit maxi mizer.
- The sales maximizer sells at a price lower than the profit maximizer. The price at any level of output is the slope of the line through the origin of the relevant point of the total-revenue curve.
- The sales maximize will earn lower profits than the profit maximizer.
- The sales maximize will never choose a level of output at which price elasticity (e) is less than unity, because from the expression

$$\mathbf{MR = P [1 – 1/e]}$$

At the point of maximum the slope of the total revenue curve is

$$\mathbf{\frac{\delta R}{\delta X} = MR = 0}$$

Therefore

$$\mathbf{0 = P [1 – 1/e]}$$

Given P > 0, we have

$$\mathbf{1 – 1/e = 0}$$

Or

$$\mathbf{e = 1}$$

we can see, that if the absolute value of e < 1, the MR < 0, denoting that TR is declining. The maximum sales level will be where the absolute value of e = 1 (and MR = 0) and will be earned only if the profit constraint is not binding (or not operative).

BAUMOL SALES MAXIMIZATION MODEL: EXAMPLE

Assume that a Demand function faced by an oligopolist is given as: $Q = 100 – 10P$
or $P = 10 – 0.1Q$
$TR = PQ = (10 – 0.1Q)Q = 10Q – 0.1Q^2$

If the total cost function of oligopolist is given as:
TC = 70 + 2Q
Then π = TR – TC
= 10Q – 0.1Q² - 70 – 2Q = - 70 + 8Q - 0.1Q²
FOC: d π / Dq = 10 - 0.2Q = 0
So that Q = 40 and P = 10 – 0.1(40) = $6
TR = = 10(40) – 0.1(40)² = $240 and
π = - 70 + 8(40) - 0.1(40)² = $90

On the other hand, The firm maximizes sales or total revenue where
d (TR) / dQ = 10 – 0.2 Q = 0 and solving for Q:
Q = 50 and P = 10 – 0.1(50) = $5.
TR = = 10(50) – 0.1(50)² = $250 and
π = - 70 + 8(50) - 0.1(50)² = $80

To find Q, P, and TR if the minimum profit constraint of the firm is π = $85, we proceed as:
π = - 70 + 8Q - 0.1Q² = $85
0.1Q² - 8Q + 155 = 0

Using the quadratic formula to find the 2 roots (we get: 32.93 and 47.07) and selecting the largest output, we have
P = 10 – 0.1(47.07) = $5.29
and TR = 10(47.07) – 0.1(47.07)² $249.14
So that π = - 70 + 8(47.07) - 0.1(47.07)²
= $85 (the minimum π required)

Lesson 37

ALTERNATIVE THEORIES OF THE FIRM (CONTINUED 1)

SALES MAXIMIZATION MODEL: W.J.BAUMOL

BAUMOL'S STATIC MODEL with Advertising

The assumptions of the model: As in the previous model the goal of the firm is sales revenue maximization subject to a minimum profit constraint which is exogenously determined. The new element in this model is the introduction of advertising as a major instrument (policy variable) of the firm. Baumol argues that in the real world non-price competition is the typical form of competition in oligopolistic markets.

$$TR = f(Q, A)$$

So the TR is now, not a function of Q alone, it's a function of A (advertising expenditure) too.

ADDED ASSUMPTION OF THE ADVERTISING MODEL

The central assumption of the advertising model is that sales revenue increases with advertising expenditure (that is, ∂TR/∂A > 0, where a = advertising expenditure). This implies that advertising will always shift the demand curve of the firm to the right and the firm will sell a larger quantity and earn larger revenue. The price is assumed to remain constant. This, however, is a simplifying assumption which may be relaxed in a more general analysis. Another simplifying assumption is that production costs are independent of advertising. Baumol recognizes that this is an unrealistic assumption, since with advertising the physical volume of output increases and the firm might move to a cost structure where production cost is different (increasing or decreasing). But he claims that this assumption is simplifying and can be relaxed without significantly changing the analysis.

Baumol argues that a firm in an oligopolistic market will prefer to increase its sales by advertising rather than by a cut in price. While an increase in physical volume induced by a price cut may or may not increase the sales revenue, depending on whether demand is elastic or inelastic, an increase in volume brought about by an increase in advertising will always increase sales revenue, since by assumption the marginal revenue of advertising is positive (∂TR/∂a > 0).

With advertising introduced into the model, it is no longer possible to have equilibrium where the profit constraint is not operative. While with price competition alone it is possible to reach an equilibrium (that is, maximize sales) where Π is not operative, with non-price competition such an unconstrained equilibrium is impossible. Unlike a price reduction, increased advertising always increases sales revenue. Consequently it will always pay the sales maximizer to increase his advertising expenditure until he is stopped by the profit constraint. Consequently the minimum profit constraint is always operative when advertising is introduced in the model.

The sales maximizer will normally have higher advertising expenditures than a profit maximizer. In any case advertising cannot be less in a sales-maximizing model.

Baumol's single-product model with advertising is shown in figure 1. Advertising expenditure is measured on the horizontal axis and the advertising function is shown as a 45° line. Costs, total revenue and profits are measured on the vertical axis. Production costs are shown as being independent of the level of advertising (curve CC').

Figure 1

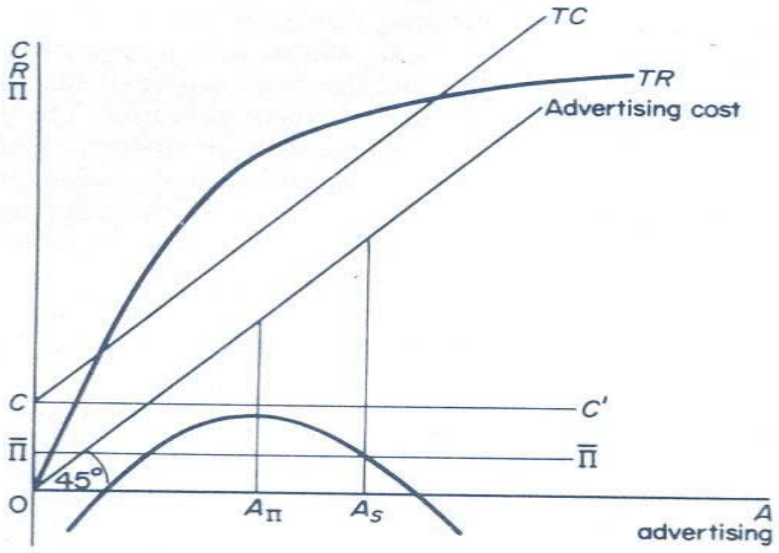


Figure 15.5

If these costs are added to the advertising cost line, we obtain the total-cost curve (TC) as a function of advertising outlay. Subtracting the total cost from the total revenue at each level of output we obtain the total-profit curveΠ. The interrelationship between output and advertising and in particular the (assumed) positive marginal revenue of advertising permits us to see clearly that an unconstrained sales maximization is now not possible. If price is such as to enable the firm to sell an output yielding profits above their minimum acceptable level, it will pay the firm to increase advertising and reach a higher level of sales revenue. The advertising expenditure of the sales maximizer ($0A_s$) is higher than that of the profit maximizer ($0A_\Pi$), and the profit constraint (Π) is operative at equilibrium.

R = f (Q, A) = TR function
C = f (Q) = total production function
Π' = minimum acceptable profit
A(a) = total cost of advertising function
Maximize: R = f (Q, A)
Subject to the Min Profit constraint
$\Pi = R - C - A \geq \Pi'$
$\partial R/\partial Q > 0, \partial C/\partial Q > 0, Q > 0$ Baumol's assumption

$\partial R/\partial Q > 0$ ensures that constraint is operative

CRITICISM OF BAUMOL'S MODEL

- The sales-maximization hypothesis cannot be tested against competing behavioral hypotheses unless the demand and cost functions of individual firms are measured. However, such data are not disclosed by firms to researchers, and are commonly unknown to the firms.

- It has been argued that in the long run the sales-maximization and the profit-maximization hypotheses yield identical solutions; because profits attain their normal level in the long-run and the minimum profit constraint will coincide with the maximum

attainable ('normal') level of profit. This argument cannot be accepted without any empirical evidence to support it.

- The sales-maximization theory does not show how equilibrium in an industry, in which all firms are sales maximizers, will be attained. The relationship between the firm and the industry is not established by Baumol.

- Baumol's hypothesis is based on the implicit assumption that the firm has market power, that is, it can have control on its price and expansion policies. The firm can take decisions without being affected by competitors' reactionss

- The assumption that the MR of advertising is positive (δTR/δA > 0) is not justified by Baumol. And casual observation shows that this may not be so.

MARRIS'S MODEL OF MAXIMIZATION OF GROWTH RATE (1964)

The goal of the firm in Marris's model is the maximization of the balanced rate of growth of the firm, that is, the maximization of the rate of growth of demand for the products of the firm, and of the growth of its capital supply:

$$\text{Maximize } g = g_D = g_C$$

Where g = balanced growth rate
g_D = growth of demand for the products of the firm
g_C = growth of the supply of capital

In pursuing this maximum balanced growth rate the firm has two constraints. Firstly, a constraint set by the available managerial team and its skills. Secondly, a financial constraint set by the desire of managers to achieve maximum job security. The rationalization of this goal is that by jointly maximizing the rate of growth of demand and capital the managers achieve maximization of their own utility as well as of the utility of the owners / shareholders.
It is usually argued by managerial theorists that the division of ownership and management allows the managers to set goals which do not necessarily coincide with those of owners. The utility function of managers includes variables such as salaries, status, power and job security, while the utility function of owners includes variables such as profits, size of output, size of capital, share of the market and public image. Thus managers want to maximize their own utility:

$$U_M = f \text{ (salaries, power, status, job security)}$$

While the owners seek the maximization of their utility

$$U_0 = f^*\text{(profits, capital, output, market share, public esteem)}$$

Marris argues that the difference between the goals of managers and the goals of the owners is not so wide as other managerial theories claim, because most of the variables appearing in both functions are strongly correlated with a single variable: the size of the firm. There are various measures of size: capital, output, revenue and market share.

From Marris's discussion it follows that the utility function of owners can be written as follows:

$$U_{\text{owners}} = f^*(g_C)$$

Where g_C = rate of growth of capital

It is not clear why owners should prefer growth to profits, unless g_C and profits are positively related. At the end of this article Marris argues in fact that g_C and Π are not always positively related. Under certain circumstances g_C and Π become competing goals. Furthermore from

Marris's discussion of the variables of the managerial utility function it seems that he implicitly assumes that, salaries, status and power of managers are strongly correlated with the growth of demand for the products of the firm: managers will enjoy higher salaries and will have more prestige the faster the rate of growth of demand. Therefore the managerial utility function may be written as follows:

$$U_M = f(g_D, s)$$

Where g_D = rate of growth of demand for the products of the firm

s = a measure of job security.

Marris treats s' as an exogenously determined constraint by assuming that there is a saturation level for job security.

$$U_M = f(g_D)\ s'$$

Where s' is the security constraint.

Thus in the initial model there are two constraints - the managerial team constraint and the job security constraint - reflected in a financial constraint.

CONSTRAINTS

The managerial constraint

The managerial constraint and the R & D capacity of the firm set limits both to the rate of growth of demand (g_D) and the rate of growth of capital supply (g_c).

The job security constraint

The risk of dismissal is largely avoided by: (a) Non-involvement with risky investments. The managers choose projects which guarantee a steady performance, rather than risky ventures which may be highly profitable, if successful, but will endanger the managers' position if they fail. Thus the managers become risk-avoiders. (b) Choosing a prudent financial policy. The latter consists of determining optimal levels for three crucial financial ratios the leverage (or debt ratio) the liquidity ratio, and the retention ratio.

The three financial ratios

The three financial ratios are combined (subjectively by the managers) into a single parameter a' which is called the 'financial security constraint. This is exogenously determined, by the risk attitude of the top management. Marris does not explain the process by which a is determined It is stated that it is not a simple average of the three ratios, but rather a weighted average, the weights depending on the subjective decisions of managers. Two points should be stressed regarding the overall financial constraint a'.

Firstly, Let,

$$a_1 = \text{liquidity ratio} = \frac{L}{A}$$

$$a_2 = \text{Leverage ratio} = \frac{D}{A}$$

$$a_3 = \text{retention ratio} = \frac{\Pi_R}{\Pi}$$

Marris postulates that overall a' is negatively related to a_1 and positively to a_2 and a_3. Secondly, Marris implicitly assumes that there is a negative relation between job security (s) and the financial constraint a'. The financial security constraint sets a limit to the rate of growth of the capital supply, g_c in Marris model.

THE MODEL: EQUILIBRIUM OF THE FIRM

The managers aim at the maximization of their own utility, which a function of the growth of demand for the products of the firm (given the security constraint)

$$U_{managers} = f(g_D)$$

The owners-shareholders aim at the maximization of their own utility which Marris assumes to be a function of the rate of growth of the capital supply (and not of profits, as the traditional theory postulated)

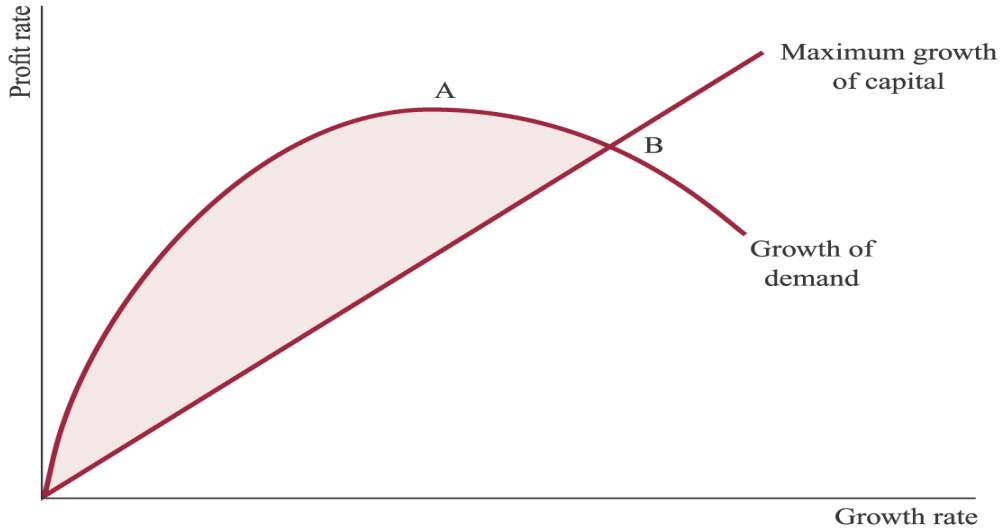
$$U_{owners} = f^*(g_C)$$

The firm is in equilibrium when the maximum balanced growth raft, attained that is, the condition for equilibrium is

$$g_D = g_C = g^* \text{ maximum}$$

The first stage in the solution of the model is to derive the 'demand' and 'supply' functions, that is, to determine the factors that determine g_D and g_C Marris establishes that the factors that determine g_D and g_c can be expressed in terms f two variables, the diversification rate, d, and the average profit margin, m. The firm will first determine (subjectively) its financial policy, that is, the value of the financial constraint a, and subsequently it will choose the rate of diversification do and the profit margin m, which maximize the balanced-growth rate g*.

Figure 1



- As the rate of growth of demand is increased, profitability is increased as well until a certain point. Then managerial constraints on growth tend to take place. Demand-growth curve clearly shows the fact that after a certain point, A, in the graph, (peak of the demand-growth curve) a higher growth rate can only be acquired at the cost of a reduction in the rate of profit.
- The optimal position for the managers is point B, the growth (utility) maximization point. Both the supply-growth and demand-growth constraints are satisfied at this point as the two curves intersect.
- The direct relationship between profitability and the growth in the case of supply growth is clearly shown in the figure. The maximum growth of capital function shows the relationship between the firm's rate of profit and the maximum rate at which the firm is able to increase its capital
- This model suggests several testable hypothesis one of which is: "owner controlled firms achieve lower growth and higher profits".

THE FOLLOWING ARE POLICY VARIABLES IN THE MARRIS MODEL:

Firstly, ā implies freedom of choice of the financial policy of the firm of the financial policy of the firm. The firm can change its g* by changing 3 security ratios: a_1 , a_2 and a_3.

Secondly, the firm can choose its diversification rate. d, either by a change in the style of its

existing range of products or by expanding the range of its products.
Thirdly, in Marris's model price is price is given by the oligopolistic structure of the industry. Price is not actually a policy variable of the firm. The determination of the price, in the market is very briefly mentioned in Marris"s article.

THE RATE OF GROWTH OF DEMAND: g_D

It is assumed that the firm grows by diversification. Growth by merger or take over is excluded from this model.
The rata of growth of demand for the products of the firm depends on the diversification rate, d, and the percentage of successful new products, k, that is,

$$g_D = f_1(d, k)$$

Where d the diversification rate defined as the number of new products introduced per time period, and k = the proportion of successful new products.

THE RATE OF GROWTH OF CAPITAL SUPPLY:g_C

It is assumed that the shareholder owners aim at the maximization of the rate of growth of the corporate capital, which is taken as a measure of the size of the firm. Corporate capital is defined as the sum of fixed assets, inventories, short-term assets and cash reserves. It is not stated why the shareholders prefer growth to profits in periods, during which growth is not steady.
The rate of growth is financed from internal and external sources. The source of internal finance for growth is profits. External finance may be obtained by the issue of new bonds or from bank loans. The optimal relation between external and internal finance, is still strongly disputed in economic literature.
Under Marris's assumptions the rate of growth of capital supply is proportional to the level of profit.

$$g_C = \bar{a}(\Pi)$$

Where $\bar{a}$ = the financial security coefficient

Π = level of total profits

The security coefficient $\bar{a}$ is assumed constant and exogenously determined in this model.
Diversification may take the form that the firm introduces a completely new product, which has no close substitutes, which creates new demand and thus competes.

MARRIS MODEL CONTRIBUTIONS

- Marris main contribution is the incorporation of financial policies of the firm into the decision-making process of the firm. This is done by introducing the financial coefficient a' = a* in the model as an additional policy variable, however it is determined exogenously.
- Besides his theory provides reconciliation between the conflicting utility functions of the managers and owners.

CRITICISM OF MARRIS MODEL

- Marris assumes cost structure and price to be given. Therefore, he assumes that profit is given which is not true. In fact, price determination has been the major point of contention in the theory of firms whereas Marris has ignored this aspect completely. This is one of the serious drawback.
- Most industries are oligopolistic and hence businesses are interdependent. Marris ignores this interdependence among firms' decisions. This implies that product differentiation goes unnoticed.
- He fails to explain oligopolistic interdependence in non collusive firms. His theory, therefore, has limited applicability.

Lesson 38

ALTERNATIVE THEORIES OF THE FIRM (CONTINUED 2)

WİLLİAMSON'S THEORY OF MANAGERİAL UTİLİTY MAXİMİZATİON

The assumptions of profit-maximisation has been criticised in a number of ways; so as a result two schools of thought emerged, namely:

1. The "Managerial School"
2. The "Behavioural School"

Oliver Williamson model is one of the Managerial School model. Williamson's model of utility maximization of managerial utility function is a finale of the managerial utility models. He was awarded Nobel Prize in the Economic Science for the year 2009. Press release of the Nobel Prize describes, "....This year's (2009) Laureates have been instrumental in establishing economic governance as a field of research. **Elinor Ostrom** has provided evidence on the rules and enforcement mechanisms that govern the exploitation of common pools by associations of users. **Oliver Williamson** has proposed a theory to clarify why some transactions take place inside firms and not in markets. Both scholars have greatly enhanced our understanding of non-market institutions."

Managers have different motives, desires and aspirations which they want to maximize rather than maximizing profit. Their perks include big company cars, lavish offices, luxurious bungalows in posh areas, foreign trips etc.

WHAT DO MANAGERS WANT?

- UTILITY = happiness, satisfaction
- What gives them utility?
- Utility = f (Salary, power, status, professional excellence)

  Utility = f(S, M, $\mathbf{I_D}$)

ASSUMPTIONS OF UTILITY MAXIMIZATION MODEL

1. Managers can act independently
2. The firm's market is not highly competitive, that is, the firm can make large (supernormal) profits
3. Managerial utility (U) is obtained from a combination of additional expenditure on staffing (*S*), managers' salaries and fringe benefits (*M*), and discretionary investment (I_D).

Williamson's model suggests that managers' self-interest focuses on the achievement of goals in 3 particular areas, namely:

$$\mathbf{U = f_1(S, M, I_D)}$$

Where S = staff expenditure, including managerial salaries (administrative and selling expenditure)

M = managerial emoluments

I_D = discretionary investment

BASIC RELATIONS AND DEFINITIONS

The demand of the firm

It is assumed that the firm has a known downward-sloping demand curve, defined by the function

$\mathbf{X = f^*(P, S, \varepsilon)}$

Or

$\mathbf{P = f_2(X, S, \varepsilon)}$

Where

X = output

P = price
S = staff expenditure
ε = the condition of the environment (a demand-shift parameter reflecting autonomous changes in demand)
It is assumed that the demand is negatively related to price, but positively related to staff expenditure and to the shift factor ε. Thus

$$\frac{\delta P}{\delta X} < 0; \qquad \frac{\delta P}{\delta S} > 0; \qquad \frac{\delta P}{\delta \varepsilon} > 0$$

An increase in staff expenditure is assumed to cause a shift in the demand curve upwards and thus allow the charging of a higher price. The same holds for any other change in the environment (ε, for example an increase in income) which shifts upwards the demand curve of the firm.

THE PRODUCTION COST
The total production cost (C) is assumed to be an increasing function of output
$\mathbf{C = f_3 (X)}$
Where
$$\frac{\delta C}{\delta X} > 0$$

ACTUAL PROFIT Π
The actual profit is revenue from sales (R), less the production costs (C), and less the staff expenditure (S)
$\mathbf{\Pi = R - C - S}$
This is the profit reported to the tax authorities. It is the actual profit less the managerial emoluments (M) which are tax deductible.
$\mathbf{\Pi_R = \Pi - M = R - C - S - M}$

MINIMUM PROFIT Π_O
This is the amount of profits (after tax) which is required for an acceptable dividend policy by the shareholders. If shareholders do not receive some profit they will be inclined to sell their shares or to vote for a change in the top management. Both actions obviously reduce the job security of managers. Hence they will make sure to have a minimum profit Π_O which is sufficient to keep shareholders satisfied. For this the reported profits must be at least as high as the minimum profit requirement plus the tax that must be paid to the government.
$\mathbf{\Pi_R \geq \Pi_O - T}$
Where T = tax
The tax function is of the form
$\mathbf{T = T' + t - \Pi_R}$
Where t = marginal tax rate (or unit profit tax)
T' = a lump sum tax.

DISCRETIONARY INVESTMENT = I_D
Discretionary investment is the amount left from the reported profit, after subtracting the minimum profit (Π_O) and the tax (T)
$\mathbf{I_D = \Pi_R - \Pi_O - T}$

DISCRETIONARY PROFIT = Π_D

This is the amount of profit left after subtracting from the actual profit (Π) the minimum profit requirement (Π_O) and the tax (T)

$\Pi_D = \Pi - \Pi_O - T$

THE MODEL: A SIMPLIFIED MODEL OF MANAGERIAL DISCRETION

We will present the model in two stages to simplify the exposition. In the first stage we assume that there are no managerial emoluments (M = 0), so that the actual profit is the reported profit for tax purposes. The simplified model may be stated as follows:

Maximize $U = f(S, I_D)$

Subject to $\Pi \geq \Pi_O + T$

Since there are no emoluments, discretionary investment absorbs all the discretionary profit. Thus we may write the managerial utility function as:

$U = f [S, (\Pi - \Pi_O - T)]$

Since $I_D = \Pi_R - \Pi_O - T$

For simplicity we may assume that there is no lump-sum tax so that $T = t\Pi$. Thus the managerial utility function becomes

$U = f [S, (1 - t) \Pi - \Pi_O)]$

Where $(1 - t) \Pi - \Pi_O = \Pi_D$ **is the discretionary profit**

The graphical presentation of the equilibrium of the firm in Williamson's model requires the construction of the indifference curves map of managers and the curve showing the relationship between the two variables appearing in the utility function **S** and Π_D.

The indifference curves of managers are drawn on a graph on whose axes we measure staff expenditure (S) and discretionary profit (Π_D.) Each indifference curve shows combinations of **S** and Π_D which give the same satisfaction to the managers. It is assumed that the indifference curves of managers are of the usual shape: they are convex to the origin implying diminishing marginal rate of substitution of staff expenditure and discretionary profit. It is further assumed that the indifference curves do not intersect the axes. This assumption restricts the choice of managers to positive levels of both staff expenditures and discretionary profits, implying that the firm will choose values of Π_D and **S** that will yield positive utility with respect to each component of its utility function.

Figure 1

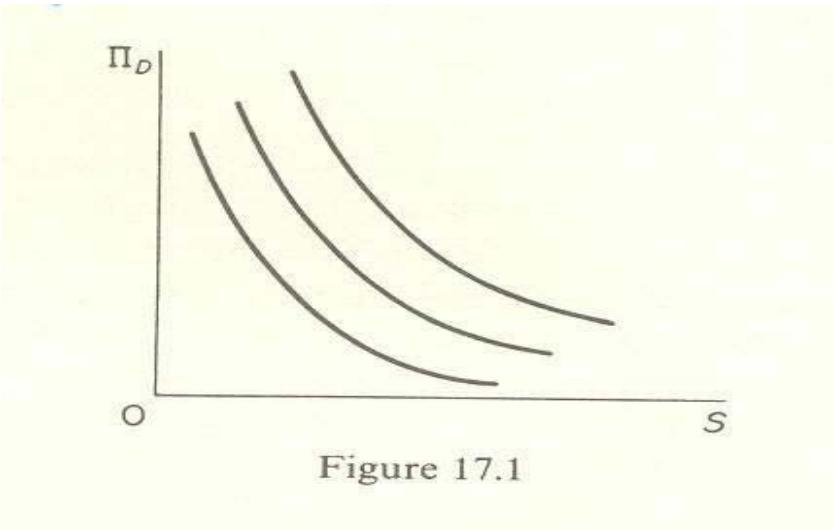


Figure 17.1

Figure 2

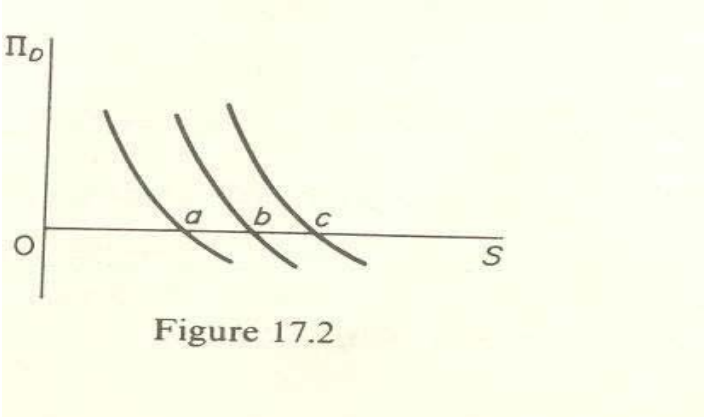


Figure 17.2

The relationship between **S**, staff expenditure, and Π_D, discretionary profit, is determined by the profit function:

$$\Pi = f(X) = f(P, S, \varepsilon)$$

Since **t** and Π_0 are exogenously given (by the tax laws and the demand for dividends of shareholders). Assuming that output is chosen optimally (according to the marginalistic rule MC = MR) and that the market environment is given (ε), the relationship between Π_D and **S** is shown in figure 3.

Figure 3

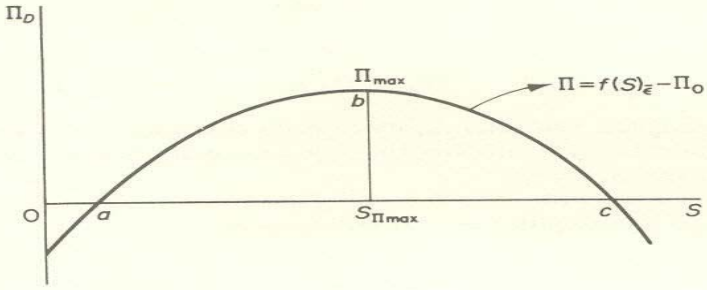
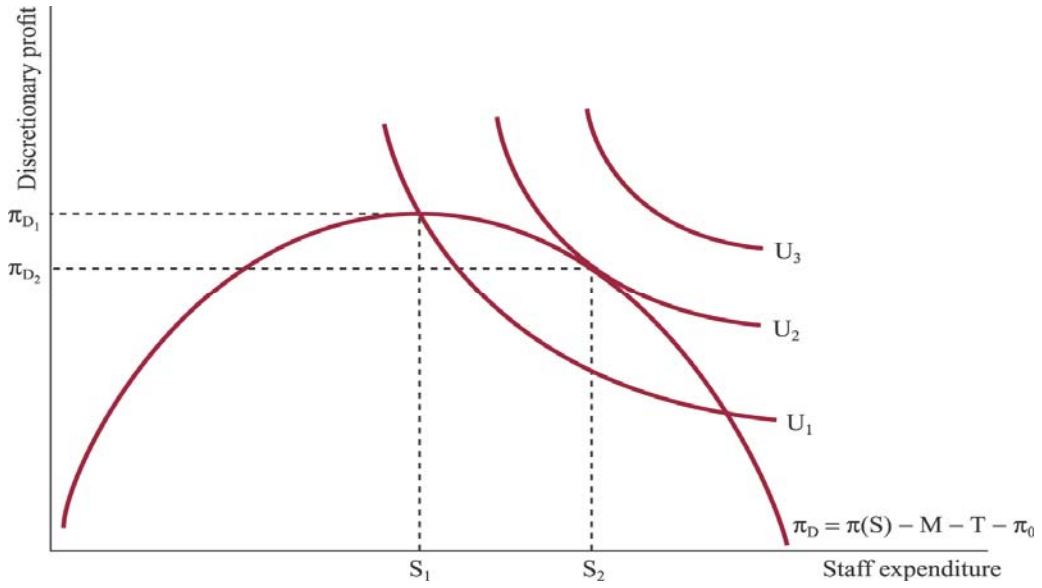


Figure 17.3

Figure 4



The overall reported profit made by the shareholders if they are to continue supporting the managers and it must also be sufficient to pay the firm's tax bill. Once these two prior commitments have been paid, the rest of the firm's reported profits are the at the managers discretion.

There is substitutability between **S** (staffing expenditure) and Π_D (discretionary profit). **S** and Π_D curve shows the trade-off between profit and staffing expenditure. This implies that managers can obtain a certain level of utility **U** from the various combinations of **S** and Π_D as shown by the managerial indifference map in Figure 1 and Figure 4. The indifference curve U_1 presents various combinations of **S** and Π_D that yield the same level of managerial satisfaction.

Initially as we move from left to right along the **S** and Π_D curve, both the level of profit and staffing expenditure increase. After point S_1 Π_{D1} , however, any further increase in expenditure increase is associated with a decrease in profit. This is because the firm is now spending more on staff than is compatible with maximizing profit. As a result, S_1 Π_{D1} is the profit maximizing point for the firm. If all the firm's reported profits were taken up by the shareholders' requirements and /or taxes then this would be the firm's optimal point. The model assumes, however, that this is not the case and that the managers will increase the amount of money spent on staff beyond this point, in order to generate increased utility for themselves. The question is, how will they maximize that utility?

U_1, U_2 and U_3 are managerial indifference curves. Along any of these curves, the managers gain equal amounts of utility. U_2 is preferred to U_1 as it yields a higher level of utility, and similarly U_3 is preferred to U_2. The managers' ability to obtain a particular level of utility is, however, constrained by what the firm can afford. In Figure 4, is out of reach of the managers. The managers will choose the point S_2 Π_{D2} on the graph where the **S-** Π_D curve is tangent to U_2 which is the highest level of utility obtainable given the relative positions of the curves. S_2 Π_{D2} is therefore the managerial utility maximization point on the graph. At this point, staffing expenditure is higher, and profit lower, than if the firm were pursuing a profit maximization strategy.

THE GENERAL MODEL OF MANAGERIAL DISCRETION

Formally the model may be stated as follows

Maximize $U = f(S, M, \Pi_R - \Pi_O - T)$

Subject to $\Pi_R \geq \Pi_O + T$

It is assumed that the marginal utility of each component of the utility function is diminishing but positive. This implies that the firm will always choose positive values for these components **(S, M, I_D).**

With the above assumption the constraint becomes redundant, and we may treat the problem as one of straightforward maximization.

Substituting

$\Pi_R = \Pi - M = R - C - S - M$

And

$T = T' + t(R - C - S - M)$

We obtain

$U = f [S, M, \{(1 - t)(R - C - S - M) - \Pi_O \}]$

We may also substitute M as Follows. Define σ as the ratio of retained to actual profit

$\sigma = \frac{\Pi_R}{\Pi}$

So that $\Pi_R = \Pi^*\sigma$

Substituting this expression in the definition of retained profit and rearranging we obtain

$\Pi_R = \Pi - M = \Pi^*\sigma$

Solving for **M** we find

$M = (1 - \sigma)\Pi = (1 - \sigma)(R - C - S)$

Where **$(1 - \sigma)$** is the proportion of profits absorbed by emoluments. Thus the managerial utility function becomes

$U = f [S, \{(1 - \sigma)(R - C - S)\}, \{\sigma(1 - t)(R - C - S) - \Pi_O \}]$

IMPLICATIONS OF THE MODEL

The implications of this model become clear if we compare it with the model of a profit maximization.

For the profit maximize:

$\Pi = R - C - S$ **and**

$\Pi_R = \Pi$

That is $\sigma = 1$. The profit maximizer will choose the values of X and S that maximize his profit

$\Pi = R - C - S$

From the first-order conditions we have

(a) $\partial\Pi / \partial X = \partial R / \partial X - \partial C / \partial X$ **or** $\partial R / \partial X = \partial C / \partial X$

(b) $\partial\Pi / \partial S = \partial R / \partial S$ **or** $\partial R / \partial S = 1$

Table 1

	Williamson	**Π Maximize**
Equil Condition	$\partial R / \partial X = \partial C / \partial X$	$\partial R / \partial X = \partial C / \partial X$
	$\partial R / \partial S < 1$	$\partial R / \partial S = 1$
	$\sigma < 1$	$\sigma = 1$

Differences of S, M, I_D at Equil	M > 0 S > 0	M = 0 S = 0
	$I_D > 0$	$I_D = 0$

In short, staffing expenditure, managerial slack and discretionary investment spending will be larger for a firm that maximizes utility than for a firm that maximizes profits.
Williamson conducted various case studies to infer that his model is better suited for the real world phenomena, such as:

1. Increase in S and M in booms, and drastic cut of these expenditures in recessions
2. Reaction of firms to taxation changes
3. Changes of the level of X, S and M in response to changes in the fixed cost of the firm
4. Drastic cuts in S by newly appointed top management, without affecting the productivity of the firm

CRITICISM OF WİLLİAMSON'S MANAGERİAL UTİLİTY MAXİMİZATİO MODEL

1. The available evidence is not enough for the verification of the theory
2. Williamson's model fails to deal with the core problem of oligopolistic interdependence of strong oligopolistic rivalry. Williamson's model hold is said to hold only where rivalry is not strong. When rivalry is strong a profit-maximizing model is more appropriate.

Lesson 39

ALTERNATIVE THEORIES OF THE FIRM (CONTINUED 3)

The assumptions of profit-maximisation has been criticised in a number of ways; so we have:
1. The “Managerial School”
2. The “Behavioural School”

BEHAVIORAL SCHOOL THEORIES were mainly presented by: - Herbert Simon, Richard Cyert and James March.
The main features of the Behavioural School are:

- Firms are multi-goal, multi-decision, multi-product organizational coalitions
- Imperfect knowledge and bounded rationality. Managers have imperfect knowledge.
- Managers cannot meet the aspiration levels of all stakeholders. Managers can never really know if they are maximizing profits, sales or growth (of the firm) or not
- Managers cannot maximize, instead they have to satisfice

CARNEGIE SCHOOL

The ‘Carnegie School’ is often identified with the pioneering work in Behavioral Economics done by Herbert Simon, James G. March and Richard Cyert in the 1950s and 1960s. The **Carnegie behavioralists** are known for their interest in understanding how individuals and organizations act and make decisions in the real world, and their challenges to the neoclassical theory of optimization and maximization in decision making organizations. Concepts such as bounded rationality and satisficing were developed to describe individuals and organizations acting in the face of ‘the uncertainties and ambiguities of life’.

CARNEGIE MELLON UNIVERSITY (ORIGINALLY CARNEGIE INSTITUTE OF TECHNOLOGY)

The background for the Carnegie School was the Ford Foundation’s mission to establish a broad and interdisciplinary behavioral social science in the late 1940s and early 1950s, at Carnegie Mellon University where Herbert Simon, James March and Richard Cyert were working. Their Students Williamsons, Feldman were also working with them.

FIRM’S GOALS: THE SATISFICING MODEL

- **Production Goal:** A goal that output must lie within a certain satisfactory range.
- **Inventory Goal:** Aims at maintaining a balanced inventory of both raw materials and finished goods. A balanced stock of both raw materials and finished goods ensures continuity of production and supply of goods to the customers and also keeps the input suppliers satisfied.
- **Sales Goal:** A goal that there must be a satisfactory level of sales however defined.
- **Market share Goal:** A goal indicating a satisfactory size of market share as a measure of comparative success as well as of the growth of the firm.
- **Profit Goal:** Still an important goal, but one amongst a number rather than necessarily the goal of overriding importance.

THE BEHAVIORAL MODEL OF CYERT AND MARCH

The behavioral theories of the firm started developing in the early 1950s. Some of the influential work may be traced in Simon's article 'A Behavioral Model of Rational Choice', published in the Quarterly Journal of Economics in 1955. The theory has subsequently been elaborated by Cyert and-March, with whose names it is connected. The writers founded their theory on four case studies and two 'laboratory'-experimental studies.

The assumptions underlying the behavioral theories about the complex nature the firm introduces an element of realism into the theory of the firm. The firm is not treated as a single-goal single-decision unit, as in the traditional theory, but as a multi-goal, multi-decision organizational coalition. The firm is considered as a coalition of different groups which are connected with its activity in various ways: managers, workers, shareholders, customers, suppliers, bankers, tax inspectors and so on. Each group has its own set of goals or demands. For example, workers want high wages good pension schemes, good conditions of work. The managers want high salaries power, prestige. The shareholders want high profits, growing capital and market size The customers want low prices and good quality and service. The suppliers want stable contracts for the materials they sell to the firm, and so on. The most important groups however, within the framework of the behavioral theories are those most directly and actively connected with the firm, namely the managers, the workers and the share holders.

GOALS OF THE FIRM: SATISFICING BEHAVIOUR

The goals of the firm are set ultimately by the top management. There are five main goals of the firm: (a) Production goal. (b) Inventory goal. (c) Sales goal. (d) Share-of-the market goal. (e) Profit goal.

The **production goal** originates from the production department. The main goal of the production manager is the smooth running of the production process. Production should be distributed evenly over time, irrespective of possible seasonal fluctuations of demand, so as to avoid excess capacity and lay-off of workers at some periods, and overworking the plant and resorting to rush recruitment of workers at other times, with the consequence of higher costs, due to excess capacity and dismissal payments or too frequent breakdowns of machinery and wastes of raw materials in period of 'rush' production.

The inventory goal originates mainly from the inventory department, if such a department exists, or from the sales and production departments. The sales department wants an adequate stock of output for the customers, while the production department requires adequate stocks of raw materials and other items necessary for a smooth flow of the output process.

The sales goal and possibly the share-of-the-market goal originate from the sales department. The same department will also normally set the 'sales strategy,' that is, decide on the advertising campaigns, the market research programs, and so on.

Market share Goal A goal indicating a satisfactory size of market share as a measure of comparative success as well as of the growth of the firm.

The profit goal is set by the top management so as to satisfy the demands of share holders and the expectations of bankers and other finance institutions; and also to create funds with which they can accomplish their own goals and projects, or satisfy the other goals of the firm.
The number of goals of the firm may be increased, but the decision-making process becomes increasingly complex. The efficiency of decision-making decreases as the number of goals increases. The law of diminishing returns holds for managerial work as for all other types of labor.

Cyert and March argue that satisficing behavior is rational given the limitations, internal and external, within which the operation of the firm is confined. Simon introduced the concept of 'bounded rationality' to justify the satisficing behavior of the large corporate firms. The goals, irrespective of where they originate, are finally decided by the top management and approved,

normally, by the board of directors. They take the form of aspiration levels, and, if attained, the performance of the firm is considered as 'satisfactory'. The goals-targets do not normally take the form of maximization of the relevant magnitudes. The firm is not a maximizing but rather a satisficing organization.

This behavior is characterized by Simon as a behavior of 'limited' or 'bounded' rationality, as opposed to 'global' rationality of the entrepreneur-firm of the traditional theory. Traditional theory conceived of the entrepreneur as a person with unlimited and costless information, unlimited computational ability and with unlimited time at his disposal. The behavioral theory recognizes explicitly the fact that in the modern real world the entrepreneurial work is executed by the group of top management. These are people with limited time at their disposal, have limited and imperfect information and limited computational ability. Hence it is impossible for them to examine all possible alternatives open to them and choose the one that maximizes profits. Instead they examine only a small number of alternatives and choose the 'best' given their limited time, information and computational abilities.

Traditional theory defined the rational firm as the firm that maximizes profit (short-run and long-run). The behaviorist school is the only theory that postulates a satisficing behavior of the firm, which is rational given the limited information and limited computational abilities of the managers.

MEANS FOR THE RESOLUTION OF THE CONFLICT

Money payments

Money payments are a major source of satisfying the demands of the various groups of the coalition-firm. In the traditional theory of the firm, money payments are the only means for achieving the goals of the firm: the entrepreneur all demands by money payments, that is, by paying to the factors their market prices. Thus, in traditional theory, there is no conflict between the owner entrepreneur and the firm, and all the demands of the owners of factors of production are satisfied by money payments.

Side payments- policy commitments

Policy commitments absorb part of the resources of the firm and are in this sense payments to the factors of production. For example, the top management, in order to keep a good scientist in its research department, apart from paying him his salary must allocate certain funds for the development and conduct of the research plans of the scientist.

'Slack' payments

Slack may be 'earned' by all groups of the coalition. For example, workers may be paid wages higher than what is required to keep them in the firm; managers may be paid higher salaries or have other perquisites (luxurious offices, limousines, expense accounts); shareholders may be paid higher dividends than the minimum required to satisfy their demands, customers may be given discounts which are not necessary to keep them tied to the firm.

UNCERTAINTY AND THE ENVIRONMENT OF THE FIRM

Cyert and March distinguish two types of uncertainty: market uncertainty and uncertainty of competitors' reactions. Market uncertainty refers to possible changes in customers' preferences or changes in the techniques of production. This form of uncertainty is inherent in any market structure. It can partly be avoided by search activity and information-gathering, but it cannot be avoided completely. Given the market uncertainty the managerial firm avoids long-term planning and works within a short time-horizon. The behavioral theory postulates that the firm considers

only the short-run and chooses to ignore the long-run consequences of short-run decisions.

A COMPARISON WITH THE TRADITIONAL THEORY

The behavioral theory differs in almost all its aspects from the traditional theory of the firm. The firm in the behavioral theory is conceived as a coalition of groups with largely conflicting interests. There is a dichotomy between ownership and management. There is also a dichotomy between individual members and the firm-organization. The consequence of these dichotomies is conflict between the different members of the coalition. According to the behavioral theory “organisations do not have objectives, only people have objectives” The firm does not exist - it is a set of shifting coalitions of individuals

The firm of the traditional theory has a single goal, that of profit maxirnisation. The behavioral theory recognizes that the modern corporate business has a multiplicity of goals. The goals are ultimately set by the top management through a continuous process of bargaining. These goals take the form of aspiration levels rather than strict maxirnising constraints. Attainment of the aspiration level 'satisfices' the firm: the contemporary firm's behavior is satisficing rather than maximizing. The firm seeks levels of profits, sales, rate of growth (and similar magnitudes) that are 'satisfactory', not maxima.

The behavioral theory is the only theory that postulates satisficing behavior as opposed to the maximizing behavior of other theories. Satisficing is considered as rational, given the limited information, time, and computational abilities of the top management. Thus the behavioral theory redefines rationality: it introduces the concept of 'bounded' or 'limited' rationality, as opposed to the 'global' rationality of the traditional theory of the firm.

In the behavioral theory the instruments which the firm uses in the decision-making process are the same as those of the traditional theory: output, price, and sales strategy (the latter including all activities of non-price competition, such as advertising, salesmanship, service, quality). The difference lies in the way by which the values of these policy variables are determined. In the traditional theory the firm chooses such values of the policy variables which will result in the maximization of the long-run profits. In the behavioral theory the policies adopted should lead to the 'satisficing' level of sales, profits, growth and so on. Cyert and March postulate that the firm is an adaptive organization: it learns from its experience. It is not from the beginning a rational institution in the traditional sense of 'global' rationality.

BEHAVIORALTHEORY: **CONTRIBUTIONS**

The behavioral theory has contributed to the development of the theory of the firm in several respects. Its main contributions are:

- Insight into the process of goal-formation and the internal resource allocation: an aspect neglected in traditional theory.
- Cyert and March's definition of 'slack' shows that this concept is equivalent to the 'economic rent' of factors of production of the traditional theory of the firm. The contribution of the behavioral school lies in the analysis of the stabilizing role of 'slack' on the activity of the firm.

THE BEHAVIORAL THEORY: SERIOUS SHORTCOMINGS.

- The behavioral theories provide a simulation approach to the complexity of the mechanism of the modem multi-goal, multi-product corporation. Simulation, however, is a predictive technique. It does not explain the behavior of the firm; it predicts the behavior without providing an explanation of any particular action of the firm.

- The behavioral theories do not deal with industry equilibrium. They do not explain the interdependence and interaction of firms, nor the way in which the interrelationship of firms leads to equilibrium of output and price at the industry level. Thus the conditions for the attainment of a stable equilibrium in the industry are not determined.

- No account is given of conditions of entry or of the effects on the behavior of established firms of a threat by potential entrants.

- The behavioral theory, although dealing realistically with the search activity of the firm cannot explain the dynamic aspects of invention and innovation.

- The behavioral theory implies a short-sighted behavior of firms. Surely the uncertainty of the market cannot be avoided by short-term planning. Most decisions require long-term view of the environment.

- The behavioral theory resolves the core problem of oligopolistic interdependence by accepting tacit collusion of the firms in the industry. This solution is unstable, especially when entry takes place. This situation is completely ignored by the behavioral theorists.

- Cyert and March based their theory on four actual case studies and two experiment studies conducted with hypothetical firms. It is obvious that the theory is founded on too few case studies so that it cannot become a generalized theory of the firm.

- This model has no predictive power whatsoever.

WHICH APPROACH IS MOST USEFUL?

Behavioural approach is a more accurate description of what happens INSIDE the firm. BUT it tells us almost nothing about how the firm will respond to changes in the environment. To use it to make predictions about how the firm will react to changes in the environment we need to know everything about the individual firm. However, if shareholders are a powerful group and their aspiration level requires making maximum profit the firm will again behave in the same way as a profit-maximizer.

Alternative and broader theories of the firm stress some relevant aspects of the operation of the modern corporation; they do not provide a satisfactory alternative to the traditional theory of the firm. Stiff competition prevailing in markets as well as managerial talent today forces managers to pay close attention to profits, otherwise firm go out of business or management is replaced. The behavioural approach is a useful complement to the profit-maximizing and managerial approaches, not a substitute for them. Traditional theory of profit maximization still holds the ground firmly.

Lesson 40

RISK ANALYSIS

RISK AND UNCERTAINTY IN MANAGERIAL DECISION MAKING

In many managerial decisions the manager often does not know the exact outcome of each possible course of action. For example, the return on a long-run investment depends on economic conditions in the future, the degree of future competition, consumer tastes, technological advances, the political climate, and many other such factors about which the firm has only imperfect knowledge. In such cases, we say that the firm faces "risk" or" uncertainty". Most strategic long run investment decisions of the firm are of this type.

Managerial decisions are made under conditions of certainty, risk, or uncertainty. **Certainty** refers to the situation where there is only one possible outcome to a decision and this outcome is known precisely. For example, investing in Defense Savings Certificate Scheme offered by the Government of Pakistan:

The average compound rate of return on maturity presently works to 12.60% p.a.

Investment	**Return**
Rs 100,000	Rs 108,000 (after 1 year)
Rs 100,000	Rs 327,630 (after 10 year)

At maturity these certificates lead to only one outcome (the amount of the yield), and this is known with certainty. The reason is that there is virtually no chance that the federal government will fail to redeem these certificates at maturity or that it will default on interest payments. On the other hand, when there is more than one possible outcome to a decision, risk or uncertainty is present.

RISK refers to a situation in which there is more than one possible outcome to a decision and the probability of each specific outcome is known or can be estimated. Thus, risk requires that the decision maker knows all the possible outcomes of the decision and have some idea of the probability of each outcome's occurrence. For example, in tossing a coin, we can get either a head or a tail, and each has an equal (i,e., a 50-50) chance of occurring provided the coin is balanced. Similarly, investing in a stock or introducing a new product can lead to one of a set of possible outcomes, and the probability of each possible outcome can be estimated from past experience or from market studies. In, general, the greater the variability of possible outcomes, the greater is the risk associated with the decision or action.

UNCERTAINTY is the case when there is more than one possible outcome to a decision and where the probability of each specific outcome occurring is not known or even meaningful. This may be due to insufficient past information or instability in the structure of the variables. In extreme forms of uncertainty not even the outcomes themselves are known. For example, drilling for oil in an unproven field carries with it uncertainty if the investor does not know either the possible oil outputs or their probability of occurrence.

In the analysis of managerial decision making involving risk, we will use such concepts as strategy, states of nature, and payoff matrix. A **strategy** refers to one of several alternative courses, of action that a decision maker can take to achieve a goal. For example, a manager may have to decide on the strategy of building a large or a small plant in order to maximize

profits or the value of the firm. **States of nature** refer to conditions in the future that will have a significant effect on the degree of success or failure of any strategy, but over which the decision maker has little or no control. For example, the economy may be booming, normal, or in a recession in the future. The decision maker has no control over the states of nature that will exist in the future. Finally, a **payoff matrix** is a table that shows the possible outcomes, or results of each strategy under each state of nature. For example, a payoff matrix may show the level of profit that would result if the firm builds a large or a small plant.

MEASURING RISK WITH PROBABILITY DISTRIBUTIONS

PROBABILITY DISTRIBUTIONS

The probability of an event is the Chance that the event will occur. For example, if we say that the probability of booming conditions in the economy next year is 0.25, or 25 percent, this means that there is 1 chance in 4 for this condition to occur. By listing all the possible outcomes of an event and the probability attached to each, we get a probability distribution. For example, if only three states of the economy are possible (boom, normal, or recession) and the probability of each occurring is specified, we have a probability distribution such as the one shown in Table 1. We might note that the sum of the probabilities is 1, or 100 percent, since one of the three possible states of the economy must occur with certainty.

Table 1

Probability Distribution of States of the Economy

State of the Economy	Probability of Occurrence
Boom	0.25
Normal	0.50
Recession	0.25
Total	1.00

The concept of probability distributions is essential in evaluating and comparing investment projects. In general, the outcome or profit of an investment project is highest when the economy is booming and smallest when the economy is in a recession. If we multiply each possible outcome or profit of an investment by its probability of occurrence and add these products, we get the expected value or profit of the project. That is,

$$E(\pi) = \overline{\pi} = \sum_{i=1}^{n} \pi_i \cdot P_i$$

Where Π**i** is the profit level associated with outcome i, P_i is the probability that outcome i will

occur, and i = 1 to n refers to the number of possible outcomes or states of nature. Thus, the expected profit of an investment is the weighted average of all possible, profit levels that can result from the investment under the various states of the economy, with the probability of those outcomes or profits used as weights. The expected profit of an investment is a very important consideration in deciding whether to undertake the project or which of two or more projects is preferable.

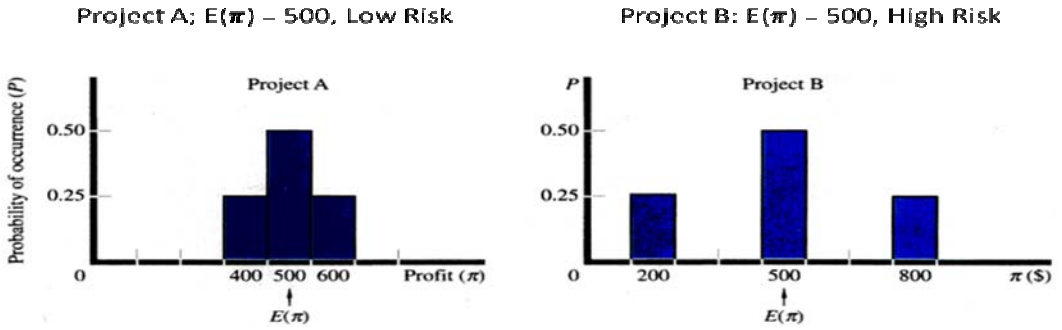
Table 2

Project	State of Economy	Probability (P)	Outcome (π)	Expected Value
A	Boom	0.25	$600	$150
	Normal	0.50	500	250
	Recession	0.25	400	100
	Expected profit from Project A			$500
B	Boom	0.25	$800	$200
	Normal	0.50	500	250
	Recession	0.25	200	50
	Expected profit from Project B			$500

For example, Table 2 presents the payoff matrix of project A and project B and shows how the expected value of each project is determined. In this case the expected value of each of the two projects is $500, but the range of outcomes for project A (from $400 in recession to $600 in boom) is much smaller than for project B (from $200 in recession to $800 in boom). Thus, project A is less risky than B and, therefore, preferable to project B.

Figure 1

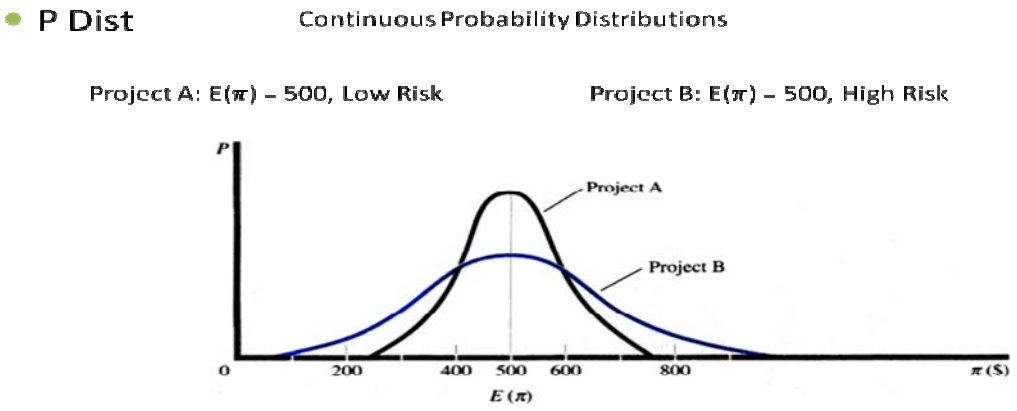
Discrete Probability Distribution



The expected profit and the variability in the outcomes of project A and project B are shown in Figure 1, where the height of each bar measures the probability that a particular outcome will occur. The relationship between the state of the economy and profits is much' tighter (i.e., less dispersed) for project A than for project B. Thus, project A is less risky than project B. Since both projects have the same expected profit, project A is preferable to project B if the manager is risk averse.

So far, we have identified only three possible states of the economy and obtained a step like discrete probability distribution of profits. As we specify more and more different states of nature, each bar becomes thinner and thinner and approaches a vertical line in the limit. We then approach the continuous probability distributions shown in Figure 2. The probability distribution for project A is again tighter or less dispersed from its expected value than the probability distribution of project B and that it reflects the smaller risk associated with project A than with project B.

Figure 2



ABSOLUTE RISK MEASUREMENT

Standard deviation, shown as σ (sigma), is a popular and useful measure of absolute risk.

Absolute risk is the overall dispersion of possible payoffs. The smaller the standard deviation, the tighter the probability distribution and the lower the risk in absolute terms. We have seen that the tighter or the less dispersed is a probability distribution, the smaller is the risk of a particular strategy or decision. The reason is that there is a smaller probability that the actual outcome will deviate significantly from the expected value. We can measure the tightness or the degree of dispersion of a probability distribution by the standard deviation, which is indicated by the symbol σ (sigma). Thus, the standard deviation (σ) measures the dispersion of possible outcomes from the expected value. The smaller the value of σ, the tighter or less dispersed is the distribution, and the lower the risk.

$$\sigma = \sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \cdot P_i}$$

The standard deviation of the probability distribution of profits for project A and project B of Table 2. The expected value or mean (Π) was found to be $500 for each project. The standard deviation of the probability distribution of profits for project A is $70.71, while that for project B is $212.13. These values provide a numerical measure of the absolute dispersion of profits from the mean for each project and confirm the greater dispersion of profits and risk for project B than for project as shown graphically in Figures 1 and 2.

Calculation of the Standard Deviation
Project A

$$\sigma = \sqrt{(600-500)^2(0.25)+(500-500)^2(0.50)+(400-500)^2(0.25)}$$

$$\sigma = \sqrt{5,000} = \$70.71$$

Calculation of the Standard Deviation

Project B

$$\sigma = \sqrt{(800-500)^2(0.25)+(500-500)^2(0.50)+(200-500)^2(0.25)}$$

$$\sigma = \sqrt{45,000} = \$212.13$$

A RELATIVE MEASURE OF RISK: THE COEFFICIENT OF VARIATION

To measure relative dispersion, we use the coefficient of variation (v). This is equal to the standard deviation of a distribution divided by its expected value or mean. That is,

$$v = \frac{\sigma}{\bar{\pi}}$$

Project A

$$v_A = \frac{70.71}{500} = 0.14$$

Project B

$$v_B = \frac{212.13}{500} = 0.42$$

The coefficient of variation, thus, measures the standard deviation per dollar of expected value or mean. As such, it is dimension-free, or, in other words, it is a pure number that can be used to compare the relative risk of two or more projects. The project with the largest coefficient of variation will be the most risky. The coefficient of variation (v) as a measure of relative dispersion or risk would still be smaller for project A than for project B. Thus, project A would have less dispersion relative to its mean (i.e., it would be less risky) than project B.

MEASURING PROBABILITIES WITH THE NORMAL DISTRIBUTION

The relation among risk, standard deviation, and the coefficient of variation can be clarified by examining the characteristics of a normal distribution, as shown in Figure 3. A **normal distribution** has a symmetrical dispersion about the mean or expected value. If a probability distribution is normal, the actual outcome will lie within ±1 standard deviation of the mean roughly 68 percent of the time; the probability that the actual outcome will be within ±2 standard deviations of the expected outcome is approximately 95 percent; and there is a greater than 99 percent probability that the actual outcome will occur within ±3 standard deviations of the mean. The smaller the standard deviation, the tighter the distribution about the expected value and the smaller the probability of an outcome that is very different from the expected value.

A normal distribution is a symmetrical distribution about the mean.

- Actual outcomes lie within ±1σ (68%).
- Actual outcomes lie within ±2σ (95%).

- Actual outcomes lie within ±3σ (99%).

STANDARDIZED VARIABLES

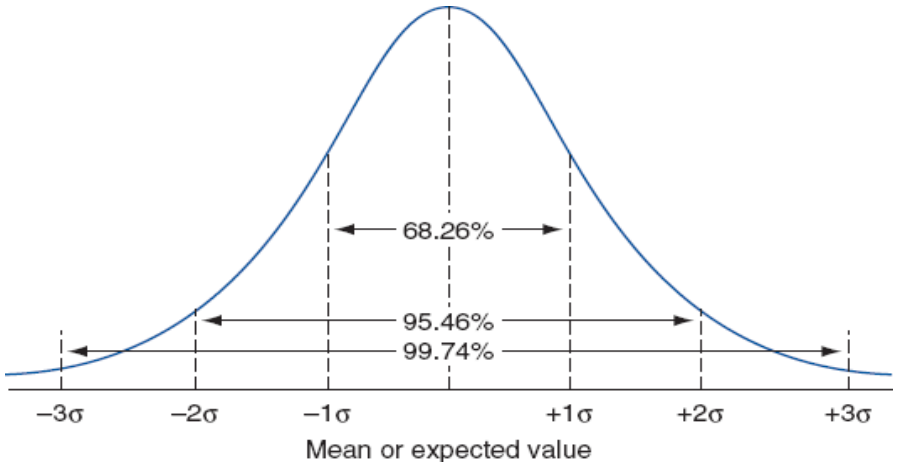
- Standardized variables have a mean of zero and a standard deviation of one. They are measured in units of σ.
- Z = (x-μ)/σ, where z is a standardized variable, x is a point of interest, μ is the mean, and σ is standard deviation.

Distribution of costs or revenues can be transformed or standardized. A **standardized variable** has a mean of 0 and a standard deviation equal to 1. Any distribution of revenue, cost, or profit data can be standardized with the following formula:

$$Z = (x-\mu)/\sigma$$

where z is a standardized variable, x is a point of interest, μ is the mean, and σ is standard deviation. . If the point of interest is 1 σ away from the mean, then $x - \mu = \sigma$, so $z = \sigma/\sigma = 1.0$. When $z = 1.0$, the point of interest is 1 σ away from the mean; when $z = 2$, the value is 2σ away from the mean; and so on. Although the standard normal distribution theoretically runs from minus infinity to plus infinity, the probability of occurrences beyond 3 standard deviations is very near zero.

Figure 3



Lesson 41

RISK ANALYSIS (CONTINUED 1)

CONCEPTS OF RISK AND UNCERTAINTY

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 the possible return on their investment. For example, assume a firm doubles its expenditure on advertisements of its product. Whether sales will increase proportionately cannot be forecasted with a high degree of certainty. There are two approaches to estimate probabilities of outcomes of a business decisions:

- (i) a priori approach: the approach based on deductive logic
- (ii) Statistical probability approach: assumes that the probability of an event in the past will hold in future also. The probability of a decision can be estimated by making use of absolute or relative measure of dispersion i.e. the standard deviation or co-efficient of variation.

UTILITY THEORY AND RISK ANALYSIS

i. In order to make effective investment decisions, one must understand the many faces of risk.
ii. The assumption of risk aversion is basic to many decision models in managerial economics. This is the most crucial assumption.
iii. At the heart of risk aversion is the notion of diminishing marginal utility for money.

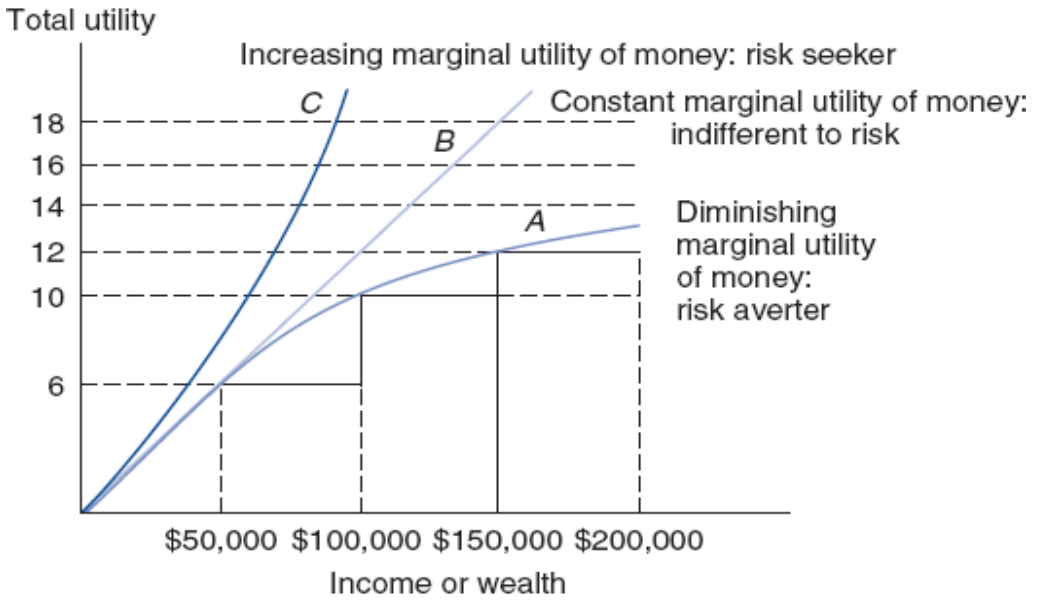
POSSIBLE RISK ATTITUDES

There are three possible attitudes toward risk: aversion to risk, indifference to risk, and preference for risk. **Risk aversion** characterizes individuals who seek to avoid or minimize risk. **Risk neutrality** characterizes decision makers who focus on expected returns and disregard the dispersion of returns. **Risk seeking** characterizes decision makers who prefer risk. Given a choice between more risky and less risky investments with identical expected monetary returns, a risk averter selects the less risky investment and a risk seeker selects the riskier investment. Faced with the same choice, the risk-neutral investor is indifferent between the two investment projects.

RELATION BETWEEN MONEY AND ITS UTILITY

At the heart of risk aversion is the notion of diminishing marginal utility for money. If someone with no money receives $5,000, it can satisfy his or her most immediate needs. If such a person then receives a second $5,000, it will obviously be useful, but the second $5,000 is not quite so necessary as the first $5,000. Thus, the value, or utility, of the second, or marginal, $5,000 is less than the utility of the first $5,000, and so on. Diminishing marginal utility of money implies that the marginal utility of money diminishes for additional increments of money. Figure 1 shows the relation between money and its utility, or value. In Figure 1, money income or wealth is measured along the horizontal axis while the utility or satisfaction of money is plotted along the vertical axis.

Figure 1



For risk averters, money has diminishing marginal utility. If such an individual's wealth were to double suddenly, he or she would experience an increase in satisfaction, but the new level of well-being would not be twice the previous level. In cases of diminishing marginal utility, a less than proportional relation holds between total utility and money. Accordingly, the utility of a doubled quantity of money is less than twice the utility of the original level. In contrast, those who are indifferent to risk identify a strictly proportional relationship between total utility and money. Such a relation implies a constant marginal utility of money, and the utility of a doubled quantity of money is exactly twice the utility of the original level. Risk seekers identify a more than proportional relation between total utility and money. In this case, the marginal utility of money increases. With increasing marginal utility of money, the utility of doubled wealth is more than twice the utility of the original amount.

Even though total utility increases with increased money for risk averters, risk seekers, and those who are indifferent to risk, the relation between total utility and money is quite different for each group. These differences lead to dissimilar risk attitudes. Because individuals with a diminishing marginal utility for money suffer more pain from a dollar lost than the pleasure derived from a dollar gained, they seek to avoid risk. Risk averters require a very high return on any investment that is subject to much risk.

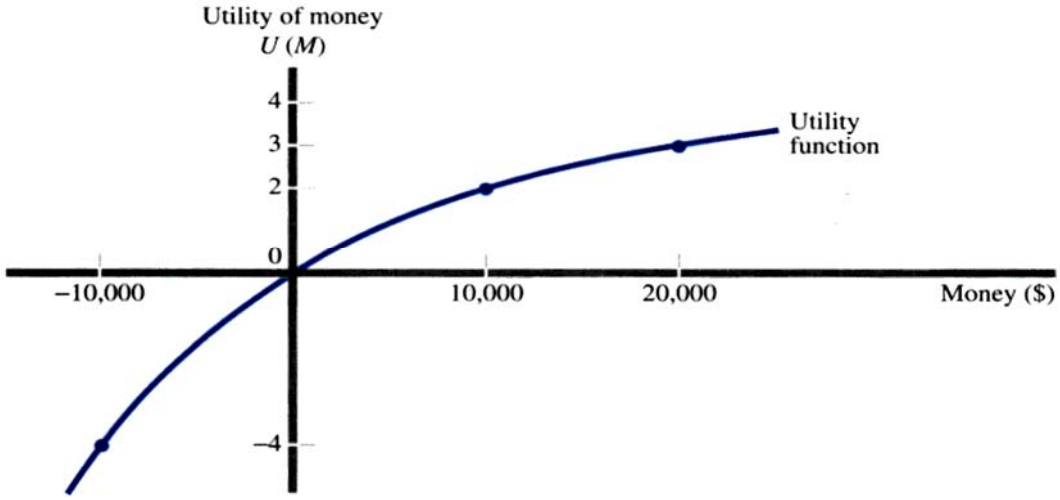
Possible Risk Attitudes

- Risk aversion is desire to avoid risk.
- Risk neutrality is to disregard risk.
- Risk seeking is preference for risk.

Relation between Money and its Utility

- Risk aversion implies diminishing marginal utility (DMU) for money.
- Risk neutrality implies constant marginal utility (CMU) for money.
- Risk seeking implies increasing marginal utility (IMU) for money.

Figure 2



Most individuals are risk averters because their marginal utility of money diminishes i.e., they face a total utility curve that is concave or faces down. For example, suppose that a manager has to decide whether or not to introduce a new product that has 40 percent probability of providing a net return of $20,000, and a 60 percent probability of a loss of $10,000. Since the expected monetary return of such a project is positive, a risk-neutral or a risk-lover manager would undertake the project. However, if the manager is risk-averse and his utility function is as shown in Figure 2, the manager would not undertake the project because the expected utility from the project is negative. (See Table 1 and Table 2).

Table 1

State of Nature	Probability	Monetary Outcome	Expected Return
Success	0.40	$20,000	$8,000
Failure	0.60	- 10,000	- 6,000
Expected Return			$2,000

Table 2

State of Nature	(1) Probability	(2) Monetary Outcome	(3) Associated Utility	(4) E(U) (1) * (3)
Success	0.40	$20,000	3	1.2
Failure	0.60	- 10,000	-4	-2.4

E(U)				-1.2

ADJUSTING THE VALUATION MODEL FOR RISK

BASIC VALUATION MODEL

The basic valuation model developed for the firm:

$$NPV = \sum_{t=1}^{n} \frac{\pi_t}{(1+r)^t}$$

This model states that the value of the firm is equal to the discounted present worth of future profits. Under conditions of certainty, the numerator is profit, and the denominator is a time value adjustment using the risk-free rate of return i. After time-value adjustment, the profits to be earned from various projects are strictly and completely comparable.

An appropriate ranking and selection of projects is possible only if each respective investment project can be adjusted for considerations of both time value of money and risk. There are two popular methods. In the first, the interest rate used in the denominator of the valuation model is increased to reflect risk considerations. In the second, expected profits are adjusted to account for risk. Each method can be used to ensure that value-maximizing decisions are made.

RISK-ADJUSTED DISCOUNT RATE

Another way to incorporate risk in managerial decision making is to adjust the discount rate or denominator of the basic valuation model.

$$k = r + Risk\ Premium$$

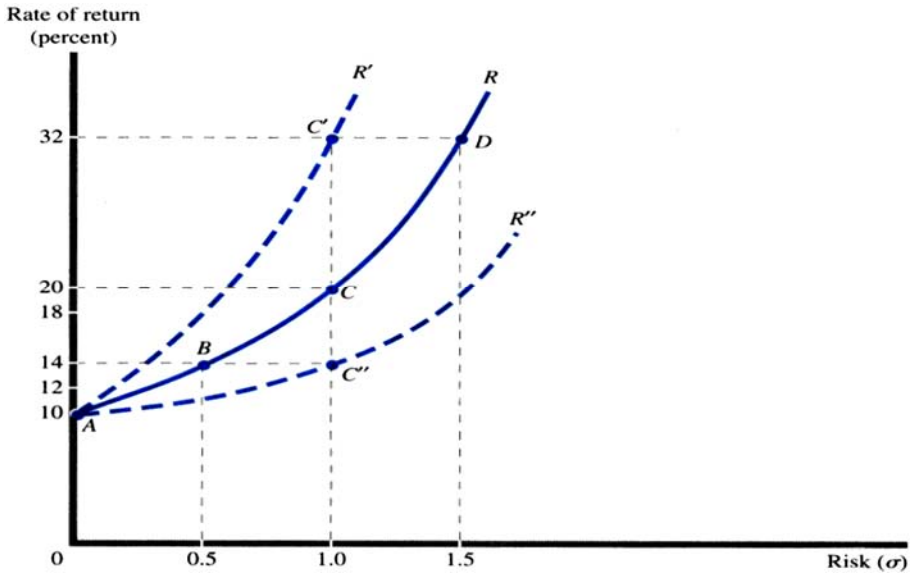
$$NPV = \sum_{t=1}^{n} \frac{\pi_t}{(1+k)^t}$$

The risk-adjusted discount rate k is the sum of the risk-free rate of return, R_F, plus the required risk premium, R_P:

$$k = R_F + R_P$$

In this method we use risk-adjusted discount rates. These reflect the manager's trade-off between risk and return, as shown by the risk return trade-off functions of Figure 3. In the figure, risk, measured by the standard deviation of profit or returns, is plotted along the horizontal axis while the rate of return on investment is plotted along the vertical axis. The risk return trade-off function or indifference curve labeled R (the middle curve in the figure) shows that the manager is indifferent among a 10 percent rate of return on a risk less asset with σ = 0 (point A), a 20 percent rate of return on an investment with σ = 1.0 (point C), and a rate of return of 32 percent for a very risky asset with σ = 1.5 (point D).

Figure 3



The difference between the expected or required rate of return on a risky investment and the rate of return on a risk less asset is called the risk premium on the risky investment. For example, the middle risk-return trade-off function labeled R in Figure shows that a risk premium of 4 percent is required to compensate for the level of risk given by σ =0.5 (the 14 percent required on the risky investment with σ = 0.5 minus the 10 percent rate on the risk less asset). A 10 percent risk premium is required for an investment with risk given by σ = 1.0, and a 22 percent risk premium for an investment with σ = 1.5. The risk-return trade-off curve would be steeper (R' in Figure) for a more risk-averse manager or investor, and less steep (R" in Figure) for a less risk-averse manager or investor. Thus, the more risk averse manager facing curve R' would require a risk premium of 22 percent (point C') for an investment with risk given by σ = 1.0, while a less risk-averse investor with curve R" would require a risk premium of only 4 percent for the same investment.

For example, suppose that a firm is considering undertaking an investment project that is expected to generate a net cash flow or-return of $45,000 for the next five years and costs initially $100,000. If the risk-adjusted discount rate of the firm for this investment project is 20 percent, we have:

$$NPV = 45{,}000 / (1.20)^5 - 100{,}000$$

$$= 45{,}000(2.9906) - 100{,}000$$
$$= \$34,\ 57$$

If the firm supposed this investment project as much more risky and used the risk adjusted discount rate of 32 percent to adjust for the greater risk, the NPV of the in vestment project would be instead:

$NPV = 45{,}000 / (1.32)^5 - 100{,}000$

$= 45{,}000(2.3452) – 100{,}000$
$= \$5{,}534$

The terms

$1 / (1.20)^5 = 2.9906$

And $1 / (1.32)^5 = 2.3452$

Are present-value-of-an-annuity interest factors. We need to consult tables.

With the risk-adjusted discount rate of 32 percent, the investment project is still acceptable, but the NPV of the project is much lower than if the firm perceived the project as -less risky and used the risk-adjusted discount rate of 20 percent. A risk adjusted discount rate of 20 percent may be appropriate for the firm for the expansion of a given line of business, while the high rate of 32 percent might be required to reflect the much higher risk involved in moving into a totally new line of business.

This method, however, has the serious shortcoming~ that risk-adjusted discount rates are subjectively assigned by managers and investors, and variations in net cash flows or returns are not explicitly considered. This approach is most useful for the evaluation of relatively small and repetitive investment projects. A better method for adjusting the valuation model for risk is the certainty-equivalent approach.

CERTAINTY-EQUIVALENT APPROACH

The certainty-equivalent approach uses a risk-free discount rate in the denominator and in-corporates risk by modifying the numerator of the valuation model, as follows:

$$NPV = \sum_{t=1}^{n} \frac{\alpha R_t}{(1 + r)^t}$$

Where R_t is the risky net cash flow or return from the investment, r is the risk-free discount rate, and α is the certainty-equivalent coefficient. α is the certain sum (i.e., the sum received with certainty that is equivalent to the expected risky sum or return on the project) divided by the expected risky sum. That is,

$$\alpha = \frac{equivalent\ certain\ sum}{expected\ risky\ sum} = \frac{R_t^*}{R_t}$$

Specifically, the manager must specify the certain sum that yields to him the same utility or satisfaction of (i.e., that is equivalent to) the expected risky sum or return from the investment. The value of α ranges from 0 to 1 for a risk-averse, decision maker and reflects his attitude toward risk. A value of 0 for α means that the project is viewed as too risky by the decision maker to offer any effective return. On the other hand, a value of I for α means that the project is viewed as risk free by the decision maker. Thus, the smaller the value of α, the greater is 'the risk perceived by the manager for the project.

For example, if the manager or investor regarded the sum of $36,000 with certainty as equivalent to the expected (risky) net cash flow or return of $45,000 per year for the next five years (on the investment project discussed in the previous example and costing initially $100,000), the value of α is:

α = 36,000 / 45,000 = 0.8
Using the risk-free discount rate of 10 percent, we can then find the net present value of the investment project, as follows:

$NPV = (0.8)\ (45{,}000) / (1.10)^5 - 100{,}000$

$= 36{,}000[1/ (1.10)^5] -100{,}000$

$= 36{,}000(3.7908) – 100{,}000$
$= 36{,}468.80$

This is close to the result obtained by using the risk-adjusted discount rate of 20 percent. If on the other hand, the firm perceived the project as much more risky and applied the certainty-equivalent coefficient of 0.62, we would have:

$NPV = (0.62)\ (45{,}000) / (1.10)^5 - 100{,}000$

$= 27{,}900\ [1/ (1.10)^5] -100{,}000$

$= 27{,}900(3.7908) – 100{,}000$
$= \$5{,}763.32$

This is close to the result obtained by using the risk-adjusted discount rate of 32 percent.

Table 3

If	Then	Implies
Equivalent certain sum < Expected risky sum	α < 1	Risk aversion
Equivalent certain sum = Expected risky sum	α = 1	Risk indifference
Equivalent certain sum > Expected risky sum	α > 1	Risk preference

The appropriate α value for a given managerial decision varies according to the level of risk and degree of the decision maker's risk aversion. Table 3 provides general relations that enable the decision-maker to use the certainty-equivalent coefficient to analyze the risk attitudes.

Lesson 42

RISK ANALYSIS (CONTINUED 2)

DECISION TREES AND COMPUTER SIMULATION

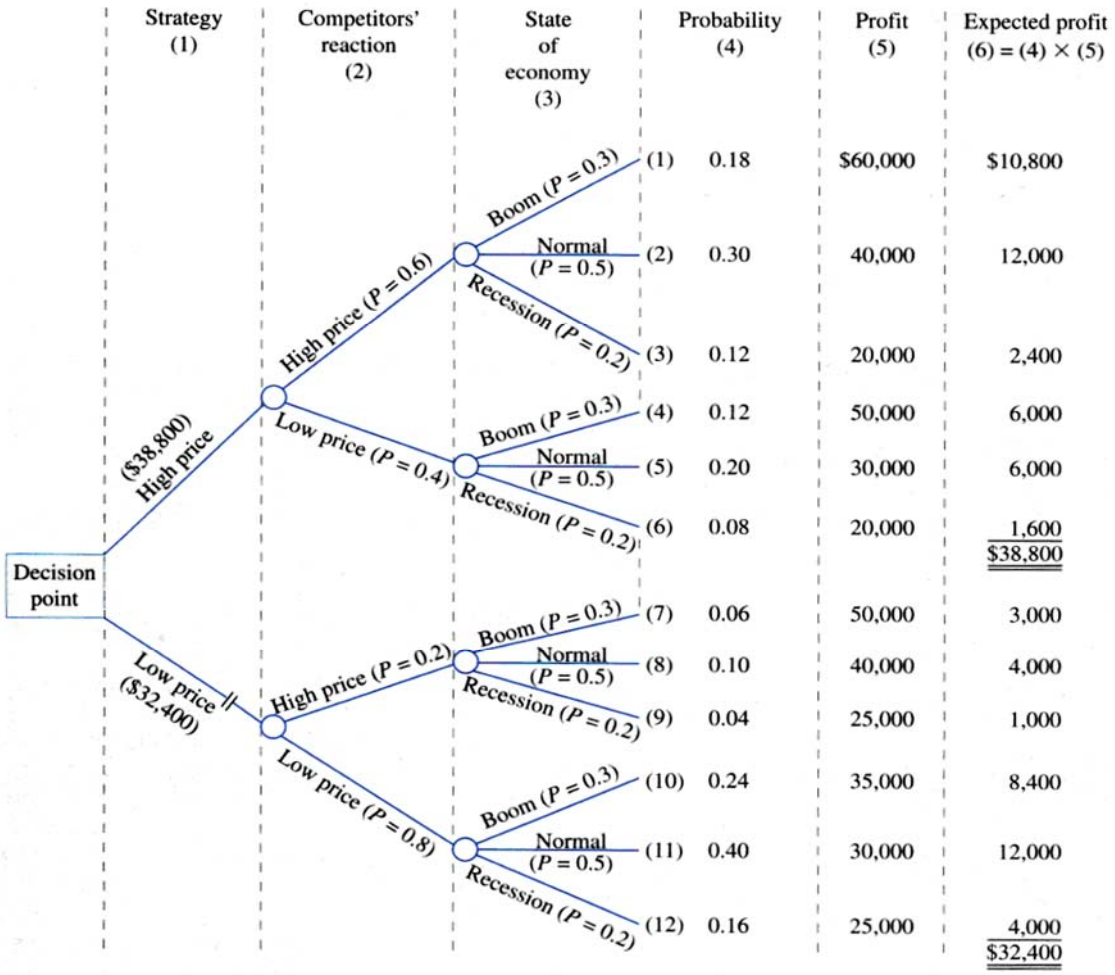
Decision trees are the sequential nature of the decision-making process. They provide a logical framework for decision analysis under conditions of uncertainty. When a high degree of uncertainty exists and data are not readily available, computer simulation often provides the basis for reasonable inference. Application of these methods today, with the use of new computer software fully automates the process of decision tree analysis and computer simulation**.**

DECISION TREES

A **decision tree** is a sequential decision-making process. Decision trees are designed for analyzing decision problems that involve a series of choice alternatives. They illustrate the complete range of future possibilities and their associated probabilities in terms of a logical progression from an initial **decision point**. Decision points are instances where management must select among several choice alternatives. **Chance events** are possible outcomes following each decision point. Decision trees are extensively employed because many important decisions are made in stages. Managerial decisions involving risk are often made in stages, with successive decisions and events depending on the outcome of earlier decisions and events. A decision tree shows the sequence of possible managerial decisions and their expected outcome under each set of circumstances or states of-nature.

Since the sequence of decisions and events is represented graphically as the branches of a tree, this technique has been named decision tree. The construction of decision trees begins with the earliest decision and moves forward in time through a series of subsequent events and decisions. At every point that a decision must be made or a different event can take place, the tree branches out until all the possible outcomes have been depicted. In the construction of decision trees, boxes are used to show decision points, while circles show states of nature. For example, Figure 1 shows a decision tree that a firm can use. to determine whether to adopt a high-price or a low-price strategy). Since the firm has control over this strategy (i.e., whether to charge a high or a low price), no probabilities are attached to these branches.

Figure 1



SIMULATION

Another method for analyzing complex, real-world decision-making situations involving risk is simulation. The first step in simulation is the construction of a mathematical model of the managerial decision-making situation that we look for to simulate. For example, the firm might construct a model for the strategy of expanding the output of a commodity. The model would specify in mathematical form the relationship between the output of the commodity and its price, output, input prices, and costs of production; output and depreciation; output, selling costs, and revenue; output, revenues, and taxes. The manager could then substitute likely values or best estimates for each variable into the model and estimate the firm profit. By then varying the value of each variable substituted into the model, the firm can get an estimate of the effect of the model or profit of the firm. This simplest type of simulation is often referred to as sensitivity analysis. This technique of **sensitivity analysis** is less expensive and less time-consuming

than full-scale computer simulation, but it still provides valuable insight for decision-making purposes.

Full-scale simulation models are very expensive and are generally used only for large projects when the decision-making process is too complex to be analyzed by decision trees. The simulation techniques are very powerful and useful, however, because they explicitly and simultaneously consider all the interactions among the variables of the model. For the evaluation of alternative business strategies involving risk where millions of dollars are involved, computer simulation is becoming more and more widely used today.

COMPUTER SIMULATION EXAMPLE

	Project X	**Project Y**
Cost	$20 million	$20 million
Average Return	15 %	20 %
The Range	-10 to 45 %	5 to 25 %
Standard Deviation	4	12
C.V.	0.267	0.60

Figure 2

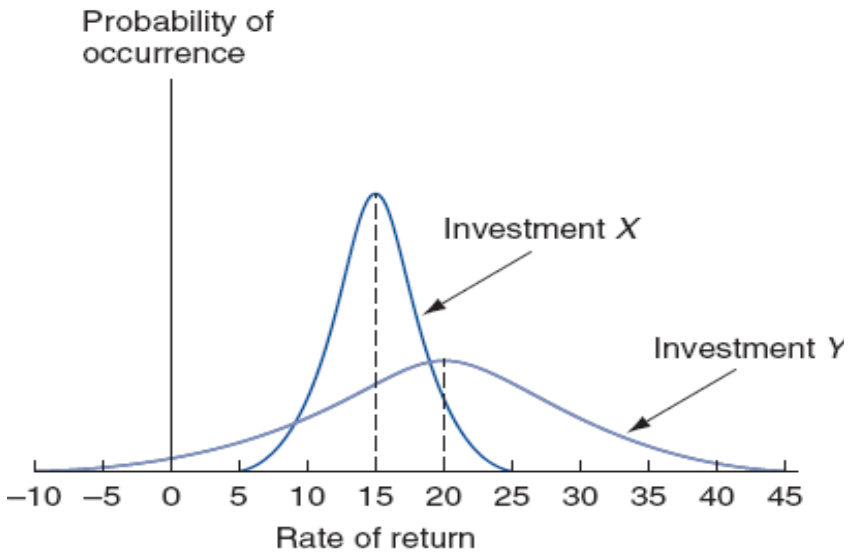


Figure 2 illustrates the frequency distribution of rates of return generated by such a simulation for two alternative projects, X and Y, each with an expected cost of $20 million. The expected rate of return on investment X is 15 percent, and 20 percent on investment Y. However, these are only average rates of return derived by the computer simulation. The range of simulated returns is from –10 percent to 45 percent for investment Y, and from 5 percent to 25 percent for investment X. The standard deviation for X is only 4 percent; that for Y is 12 percent. Based on this information, the coefficient of variation is 0.267 for investment X and 0.60 for investment Y. Investment Y is clearly riskier than investment X. A decision about which alternative to choose

can be made on the basis of expected utility, or on the basis of a present value determination that incorporates either certainty equivalents or risk-adjusted discount rates. Figure 2 clearly shows that Project Y is riskier than Project X.

USES OF GAME THEORY IN RISK ANALYSIS

In an uncertain economic environment, value maximization is achieved using the risk-adjusted valuation models, decision trees and computer simulation. Under certain circumstances, however, when extreme uncertainty exists, game theory decision criteria may be appropriate. Uncertainty is defined as the case where there is more than one possible outcome to a decision and the probability of each specific outcome occurring is not known or even meaningful. As a result, decision making under uncertainty is necessarily subjective, some specific decision rules are available, however, if the decision maker can identify the possible states of nature and estimate the payoff for each strategy. Two specific decision rules applicable under uncertainty are the maximin criterion and the minimax regret criterion.

MAXIMIN DECISION RULE

One decision that is sometimes applicable for decision making under uncertainty is the maximin criterion. This criterion states that the decision maker should select the alternative that provides the best of the worst possible outcomes. This is done by finding the worst possible (minimum) outcome for each decision alternative and then choosing the option whose worst outcome provides the highest (maximum) payoff. This criterion instructs one to maximize the minimum possible outcome. The maximin criterion postulates that the decision maker should determine the worst possible outcome of each strategy and then pick the strategy that provides the best of the worst possible outcomes. The maximin criterion can be illustrated by applying it to the example in Table 1, where the firm could follow the strategy of introducing a new product that would provide a return of $20,000 if it succeeded or lead to a, loss of $10,000 if it failed, or choose not to invest in the business enterprise, with zero possible return or loss. This matrix is shown in Table 1. We might notice that no probabilities are given in Table 1 because we are now dealing with uncertainty. That is, we now assume that the manager does not know and cannot estimate the probability of success and failure of investing in the new product. Therefore, he or she cannot calculate expected payoff or return and risk of the investment.

Table 1

The payoff matrix below shows the payoffs from two states of nature and two strategies

	State of Nature		
Strategy	Success	Failure	Maximin
Invest	20,000	-10,000	-10,000
Do Not Invest	0	0	0

To apply the maximin criterion to this investment, the manager first determines the worst possible outcome of each strategy (row). This is $10,000 in the case of failure for the investment strategy and 0 for the strategy of not investing. These worst possible outcomes are recorded in the last or maximin column of the Table 1. Then, he picks the strategy that provides the best (maximum) of the worst (minimum) possible outcomes (i.e., maximin), This is the strategy of not investing. Thus, the maximin criterion picks the strategy of not investing, which has the

maximum of the minimum payoffs. The payoff matrix 2 shows the payoffs from two states of nature and two strategies. For the strategy "Invest" the worst outcome is a loss of 10,000. For the strategy "Do Not Invest" the worst outcome is 0. The maximin strategy is the best of the two worst outcomes - Do Not Invest. Which is indicated by the red circle to its zero return or loss in the last column of the Table 2 (compared with the loss of $10,000 in the case of failure with the introduction of the new product).

Table 2

	State of Nature		
Strategy	Success	Failure	Maximin
Invest	20,000	-10,000	-10,000
Do Not Invest	0	0	0

This criterion is appropriate when the firm has a very strong aversion to risk, as, for example, when the survival of a small firm depends on avoiding losses. The maximin criterion is also appropriate in the case of oligopoly, where the actions of one firm affect the others. Then, if one firm lowers its price, it can expect the others to soon lower theirs, thus reducing the profits of all.

MINIMAX REGRET DECISION RULE

A second useful decision criterion focuses on the opportunity loss associated with a decision rather than on its worst possible outcome. This decision rule, known as the **minimax regret criterion**, states that the decision maker should minimize the maximum possible regret (opportunity loss) associated with a wrong decision. This criterion instructs one to minimize the difference between possible outcomes and the best outcome for each state of nature. **Opportunity loss** is defined as the difference between a given payoff and the highest possible payoff for the resulting state of nature. Opportunity loss is always a positive figure or zero, because each alternative payoff is subtracted from the largest payoff possible in a given state of nature.

The minimax regret criterion postulates that the decision maker should select the strategy that minimizes the maximum regret or opportunity cost of the wrong decision, whatever the state of nature that actually occurs. From the opportunity loss or regret matrix, the **cost of uncertainty** is measured by the minimum expected opportunity loss. From the payoff matrix, the cost of uncertainty is measured by the difference between the expected payoff associated with choosing the correct alternative under each state of nature (which will be known only after the fact) and the highest expected payoff available from among the decision alternatives. The cost of uncertainty is the unavoidable economic loss that is due to chance.

Regret or opportunity cost is measured by the difference between the payoff of a given strategy and the payoff of the best strategy under the same state of nature. The rationale for measuring regret this way is that if we have chosen the best strategy (i.e., the one with the largest payoff) for the particular state of nature that has actually occurred, then we have no regret or zero regret. But if we have chosen any other strategy, the regret is the difference between the payoff of the best strategy under the specific state of nature that has occurred and the payoff of the strategy chosen. After determining the maximum regret for each strategy under each state of nature, the decision maker then chooses the strategy with the minimum regret value.

To apply the minimax regret criterion, the decision maker must first construct a regret matrix from the payoff matrix. For example, Table 3 presents the payoff and regret matrices for the investment problem of Table 1. The regret matrix is constructed by determining the maximum payoffs for each state of nature (column) and then subtracting each payoff in the same column from that figure. These differences are the measures of regrets.

Table 3

	State of Nature		Regret Matrix	
Strategy	Success	Failure	Success	Failure
Invest	20,000	-10,000	0	10,000
Do Not Invest	0	0	20,000	0

For example, on one hand, if the manager chooses to invest in the product and the state of nature that occurs is the one of success, he has no regret because this is the correct strategy. Thus, the regret value of zero is appropriately entered at the top of the first column in the regret matrix in Table 3. On the other hand, if the firm had chosen not to invest, so that it had a zero payoff under the same state of nature of success, the regret is $20,000. This regret value is entered at the bottom of the first column of the regret matrix. Moving to the state of failure column in the payoff matrix, we see that the best strategy (i.e., the one with the largest payoff) is not to invest. Compare 0 and – 10,000, remember 0 > -10,000. This has a payoff of zero. Thus, the regret value of this strategy is zero (the bottom of the second column in the regret matrix). If the firm undertook the investment under the state of nature of failure, it would incur a loss and a regret of $10,000 (the top of the second column of the regret matrix). Note that the regret value for the best strategy under each state of nature is always zero and that the other regret values in the regret matrix must necessarily be positive since we are always ,subtracting smaller payoffs from the largest payoff under each state of nature (column) i.e. 0 – (-10,000) = 10,000.

After constructing the regret matrix (with the maximum regret for each strategy under each state of nature), the decision maker then chooses the strategy with the minimum regret value. In our example, this is the strategy of investing, which has the minimum regret value of $10,000 (indicated by the green rectangle in the maximum regret column of Table 4). This compares with the maximum regret of $20,000 resulting from the strategy of not investing. Thus, while the best strategy for the firm according to the maximin criterion is not to invest, the best strategy according to the minimax regret criterion is to invest. The choice as to which of these two decision rules the firm might apply under conditions of uncertainty depends on the firm's objectives and on the particular investment decision that it faces.

Table 4

	State of Nature		Regret Matrix		Maximum
Strategy	Success	Failure	Success	Failure	Regret
Invest	20,000	-10,000	0	10,000	10,000
Do Not Invest	0	0	20,000	0	20,000

INFORMATION AND RISK

Risk often results from lack of information or insufficient information. The relationship between information and risk can be analyzed by-examining asymmetric information, adverse selection, and moral hazard.

ASYMMETRIC INFORMATION AND THE MARKET FOR USED CARS

Asymmetric Information: Situation in which one party to a transaction has less information than the other with regard to the quality of the product or service.

ADVERSE SELECTION

- Problem that arises from asymmetric information
- Low-quality goods drive high-quality goods out of the market

A classic example of this is the market for “lemons" (i.e., a defective product, such as a used car, that will require a great deal of costly repairs and is not worth its price). Buyers have incentive to buy inferior goods when the quality of goods is difficult to identify. When the buyer is less informed about the quality than the seller, adverse selection may result. Akerlof's paper used the market for used cars as an example of the problem of quality uncertainty. It concludes that owners of good cars will not place their cars on the used car market. This is sometimes summarized as "the bad driving out the good" in the market. Lemons vs Cherries. The buyers, wary of being cheated by lemons, have lower willingness to pay for the goods than if they have full information. Meanwhile, with a lower price acceptable by the buyers, sellers of high quality good opt to quit, with only lemons left on the market, further shrinking the trade.

Specifically, sellers of used cars know exactly the quality of the cars that they are selling but prospective buyers do not. As a result, the market price for used cars will depend on the quality of the average used car available for sale. The owners of lemons would then tend to receive a higher price than their cars are worth, while the owners of high-quality used cars would tend to get a lower price than their cars are worth. The owners of high-quality used cars would therefore withdraw their cars from the market. The process continues until only the lowest-quality cars are sold in the market at the appropriate very low price. Thus, the end result is that low-quality cars drive high-quality cars out of the market. This is known as adverse selection.

The problem of adverse selection arises not only in the market for used cars, but in, any market characterized by asymmetric information, such as the market for individual health insurance. Here the individual knows much more about the state of her health than an insurance company can ever find out, even with a medical examination. As a result, when an insurance company sets the insurance premium for the average individual (i.e., an individual of average health), unhealthy people are more likely' to purchase insurance than healthy people. Because of this adverse selection problem, the insurance company is forced to raise the insurance premium, thus making it less advantageous for healthy people to purchase insurance.

THE PROBLEM OF MORAL HAZARD

Moral Hazard

- Tendency for the probability of loss to increase when the loss is insured

Methods of Reducing Moral Hazard

- Specifying precautions as a condition for obtaining insurance
- Coinsurance

Another problem that arises in the insurance market is that of moral hazard. This refers to the increase in the probability of an illness, fire, or other accident when an individual is insured than when he is not. With insurance, the loss from an illness, fire, or other accident is shifted from the individual to the insurance company. Therefore, the individual will take fewer precautions to

avoid the illness, fire, or other accident, and when a loss does occur, he will tend to inflate the amount of the loss. With auto insurance, an individual may drive more recklessly (thus increasing the probability of a car accident) and then is likely to exaggerate the injury and inflate the property damage that he suffers if he does get into an accident.

CONCLUSION

Uncertainty is an important factor in Investment decision but there is no unique method of dealing with uncertainty. None of the methods leads to flawless decision, but they do add some degree of certainty to decision-making.

Decision making under conditions of uncertainty is greatly facilitated by use of the tools and techniques that we have discussed. Although uncertainty can never be eliminated, it can be assessed and dealt with to minimize its harmful consequences.

Lesson 43

CAPITAL BUDGETING

CAPITAL BUDGETING DEFINED

The term Capital refers to the funds employed to finance business; a budget is a detailed plan of projected inflows and outflows over future periods. Capital Budgeting is planning expenditures that generate cash flows expected to stretch beyond one year. Process of planning expenditures that give rise to revenues or returns over a number of years.

CATEGORIES OF INVESTMENT

- Replacement Investments
- Cost Reduction Investments
- Output Expansion to Accommodate Demand Increases
- Output Expansion for New Products
- Government Regulation

In general, firms classify investment projects into the following categories:
1. Replacement: Investments to replace equipment that is worn out in the production process.
2. Cost reduction. Investments to replace working but obsolete equipment with new and more efficient equipment, expenditures for training programs aimed at reducing labor costs, and expenditures to move production facilities to areas where labor and other inputs are cheaper.
3. Output expansion of traditional products and markets: Investments to expand production facilities in response to increased demand for the firm's traditional products in traditional or existing markets.
4. Expansion into new products and/or markets: Investments to develop, produce, and sell new products and/or enter new markets.
5. Government regulation. Investments made to fulfill government regulations. These include investment projects required to meet government health and safety regulations, pollution control, and to satisfy other legal requirements.

In general, investment decisions to replace worn-out equipment are the easiest to make since management is familiar with the specifications and operating and maintenance costs of existing equipment and with the time when it needs to be replaced. Investment projects to reduce costs and expand output in traditional products and markets are generally more complex and usually require more detailed analysis and approval by higher-level management.
Investment projects to produce new products and move into new markets are very complex because of the much greater risk involved. They are also the most vital and financially rewarding in the long run, since a firm's product line tends to become obsolete over time and its traditional market may shrink or even disappear, for example the market for type writers, floppy disk and VCRs.

It is clear that the generation of ideas and proposals for new investment projects is central to the future profitability of the firm. Most large firms have a research and development divisions. Such a division is staffed by experts in product development, marketing research, industrial engineering and so on.

Although the final decision to undertake or not to undertake a major investment project is made by the firm's top management, the capital budgeting process involves most of the firm's divisions. The marketing division will need to forecast the demand for the new products that the firm plans to sell; the production, engineering, personnel, and purchasing departments provides feasibility studies and estimates of the cost of the investment project; and the financing

department determines how the required investment funds are to be raised and their cost to the firm. Thus, capital budgeting can be said to join together the operations of all the major divisions of the firm.

THE CAPITAL BUDGETING PROCESS
PROJECTING CASH FLOWS
One of the most important and difficult aspects of capital budgeting is the estimation of the net cash flow from a project. This is the difference between cash receipts and cash expenditures over the life of a project. Since cash receipts and expenditures occur in the future, a great deal of uncertainty is involved in their estimation.
A typical project involves making an initial investment and generates a series of net cash flows over the life of the project. The initial investment to add a new product line may include the cost of purchasing and installing new equipment.

For example, suppose that a firm estimates that it needs to make an initial investment of $1 million in order to introduce a new product. The marketing division of the firm expects the life of the product to be five years. Incremental sales revenues are estimated to be $1 million during the first year of operation and to rise by 10 percent per year until the fifth year, when the product will be replaced. The production department projects that the incremental variable costs of producing the product will be 50 percent of incremental sales revenues and that the firm would also incur additional fixed costs of $150,000 per year. The finance department anticipates a marginal tax rate of 40 percent for the firm. The finance department of the firm would use the straight-line depreciation method so that the annual depreciation charge would be $200,000 per year for five years. The salvage value of the initial equipment is estimated to be $250,000, and the firm also expects to recover $100,000 of its working capital at the end of the fifth year. The cash flows from this project are given in Table 1.

TABLE 1
Example: Calculation of Net Cash Flow

Sales	$1,000,000
Less: Variable costs	500,000
Fixed costs	150,000
Depreciation	200,000
Profit before taxes	$150,000
Less: Income tax	60,000
Profit after taxes	$90,000
Plus: Depreciation	200,000
Net cash flow	$290,000

TIME VALUE OF MONEY
The decision, whether the firm should undertake the investment or not, firm must compare the net cash flow over the five years of the project given in Table 1, to the initial $1 million cost of the project. Since capital budgeting involves cash flows occurring at various times in future, we must make them equivalent at a particular point in time . This involves the use of time value of money. All this term means is that a dollar today is worth more than a dollar tomorrow. As long as there is an opportunity to earn positive return on funds, a $ today and a $, a year from now are not equivalent. Since $1 received in future years is worth less than $1 spent today. Thus to put cash flows originating at different times on an equal basis, we must apply an interest rate to each of the flows so that they are expressed in terms of the same point in time.

METHODS OF CAPITAL PROJECT EVALUATION

Methods that **discount** cash flows to a present value

- internal rate of return (IRR)
- net present value (NPV)
- profitability index (PI)

NON-DISCOUNTED PAYBACK MODELS

- **Payback**: time period (years) necessary to recover the original investment
- **Accounting rate of return**: percentage resulting from dividing average annual profits by average investment

PAYBACK PERIOD

Payback Period is the amount of time required for the firm to recover its initial cost in a project, as calculated from a *cash inflow.* In the case of an annuity, the payback period can be found by dividing the initial investment by the annual cash inflow. For a mixed stream of cash inflows, the yearly cash inflows must be accumulated until the initial investment is recovered. Although popular, the payback period is generally viewed as an unsophisticated capital budgeting technique, because it does not explicitly consider the time value of money. When the payback period is used to make accept-reject decisions, the following decision criteria apply:

The decision criteria

- If the payback period < the maximum acceptable payback period, accept the project.
- If the payback period > the maximum acceptable payback period, reject the project.

In equation form, the payback period is

Payback Period = Number of Years to Recover Investment

The payback period can be thought of as a breakeven time period. The shorter the payback period, the more desirable the investment project. The longer the payback period, the less desirable the investment project. Payback period is a useful but rough measure of liquidity and project risk.

The length of maximum acceptable payback period is determined by management. This value is set subjectively on the basis of:

- the type of the project (expansion, replacement, renewal)
- the perceived risk of the project
- the perceived relationship between the payback period and the share value

TABLE 2
Payback Period Example
Firm XYZ

Project A	**Project B**
Initial Cost	$42,000	$45,000
Year	Operating Cash Inflows	
1	14,000	28,000
2	14,000	12,000
3	14,000	10,000
4	14,000	10,000
5	14,000	10,000

In Table 2, for the firm XYZ, project A, which is an annuity, the payback period is 3 years ($42,000 / $14,000). Because project B generates a mixed stream of cash inflows, the calculation of its payback period is not as clear cut. In year 1, the firm will recover $28,000 of its initial investment of $45,000. By the end of year 2, $40,000 ($28,000 from year 1 + $12,000 from year 2) will have been recovered. Only 50% of the year 3 cash-inflow of $10,000 is needed to complete the initial investment of $45,000. The payback period for project B is therefore 2.5 years (2 years + 50% of year 3). If for the firm XYZ, maximum acceptable payback period were 2.75 years, project A would be rejected and project B would be accepted. If the maximum acceptable payback period were 2 years, both projects would be rejected. If the projects were being ranked, B would be preferred over A, because it has a shorter payback period.

The payback period is frequently used by large firms to evaluate small projects and by small firms to evaluate most projects. Its popularity is due to its computational simplicity. However, it has some **drawbacks**:

- The major weakness of the payback period is that the appropriate payback period is only a subjectively determined number by the managers.
- It cannot be specified in light of the wealth maximization goal because it is not based on discounting cash flows.
- The appropriate payback period is simply the maximum acceptable period of time over which management decides that a project's cash flows must break even i.e. just equal the initial investment.
- Another weakness is that this approach fails to take fully into account the time factor in the value of money.
- Yet another one is this technique's failure to recognize cash flows that occur after the payback period.

NET PRESENT-VALUE ANALYSIS (NPV)

One method of deciding whether or not a firm should accept an investment project is to determine the net present value of the project. The most commonly employed method for long-term investment project evaluation is called **net present-value (NPV)** analysis. NPV analysis is the difference between the marginal revenues and marginal costs for individual investment projects, when both revenues and costs are expressed in present value terms. NPV analysis meets all of the criteria for an effective capital budgeting decision rule. As a result, it is the most commonly applied capital budgeting decision rule. The net present value (NPV) of a project is equal to the present value of the expected stream of net cash flows from the project, discounted at the firm's cost of capital, minus the initial cost of the project. As pointed out in the following equation, the net present value (NPV) of a project is given by:

$$NPV = \sum_{t=1}^{n} \frac{R_t}{(1+k)^t} - C_0$$

Where

R_t = Return (net cash flow)
k = Risk-adjusted discount rate
C_0 = Initial cost of project

The value of the firm will increase if the NPV of the project is positive and decline if the NPV of the project is negative.

The decision criteria

- the firm should undertake the project if its NPV is positive
- and should not undertake it if the NPV of the project is negative

For example, if k (the risk-adjusted discount rate) or cost of capital of the project to the firm is 12 percent, the net present value of the project with the estimated net cash flow given in Table 1, and the initial cost of$1 million:

NPV = 1,454,852 – 1,000,000
= $454,852 k =12%
NPV = $543,012 k = 10%
NPV = $169,078 k = 20%

This project would thus add $454,852 to the value of the firm, and the firm should undertake it. If on the other hand, the firm had used the risk-adjusted discount rates of 10 percent and 20 percent, respectively, the net present value of the project would have been $543,012 and $169,078. Thus, even if k = 20 percent, the firm should undertake the project.

CAPITAL RATIONING

Capital rationing: the practice of restricting capital expenditures to a certain amount due to:

- reluctance to incur increasing levels of debt
- due to limits on external financing
- Management may not want to (add to equity) sell stocks in fear of losing control of the firm.
- Undertaking all the projects with positive NPV may involve such rapid expansion as to strain the managerial and other resources of the firm.

PROFITABILITY INDEX (PI) OR THE BENEFIT/COST RATIO

In cases of capital rationing (i.e., when the firm cannot undertake all the projects with positive NPV), the firm should rank projects according to their profitability index and choose the projects with the highest profitability indexes rather than those with the highest NPVs. A variant of NPV analysis that is often used in complex capital budgeting situations is called the **profitability index (PI), or the benefit/cost ratio method.** The profitability index is calculated as follows:

PI = PV of Cash Inflows / PV of Cash Outflows

PI or cost/benefit ratio shows the relative profitability of any project, or the present value of benefits per dollar.

$$PI = \frac{\sum_{t=1}^{n} \frac{R_t}{(1+k)^t}}{C_0}$$

R_t = Return (net cash flow)
k = Risk-adjusted discount rate
C_0 = Initial cost of project

PI & NPV accept/reject decision criteria

In PI analysis, a project with PI > 1 should be accepted and a project with PI < 1 should be rejected. Projects will be accepted provided that they return more than a dollar of discounted benefits for each dollar of cost. Thus, the PI and NPV methods always indicate the same accept/ reject decisions for independent projects, because PI > 1 implies NPV > 0 and PI < 1 implies NPV < 0. However, for alternative projects of unequal size, PI and NPV criteria can give different project rankings. This can sometimes cause problems when mutually exclusive projects are being evaluated.

TABLE 3

Comparison of NPV and PI Rankings of Projects with Unequal Costs

	Project A	Project B	Project C
PVNCF	2,600,000	1,400,000	1,400,000
C_0	2,000,000	1,000,000	1,000,000

$NPV = PVNCF - C_0$

$$PI = \frac{PVNCF}{C_0}$$

	Project A	Project B	Project C
	$ 600,000	$ 400,000	$ 400,000
PI	1.3	1.4	1.4

For example, the data in Table 3 show that while project A has a higher NPV, than either project B or C and would, therefore, be the only project undertaken according to the NPV investment rule if the firm could invest only $2 million, the profitability indexes for projects Band C are greater than for project A, and the firm should undertake both of these projects instead of project A. That is, jointly, projects B and C increase the value of the firm by more than project A, but they would not be undertaken if the firm followed the NPV rule and could invest only $2 million. This example shows that with capital rationing, the profitability index or relative NPV rule may lead to a different ranking or order in which projects are to be undertaken. Obviously, in the absence of capital rationing, the firm will undertake all projects with a positive NPV or profitability index larger than 1.

INTERNAL RATE OF RETURN ANALYSIS (IRR)

The **internal rate of return (IRR)** is the interest or discount rate that equates the present value of the future receipts of a project to the initial cost. The equation for calculating the internal rate of return is simply the NPV formula set equal to zero:

NPVi = 0

Here the equation is solved for the discount rate, *ki**, which produces a zero net present value or causes the sum of the discounted future receipts to equal the initial cost. That discount rate is the internal rate of return earned by the project. It is given by:

$$\sum_{t=1}^{n} \frac{R_t}{(1 + k^*)^t} = C_0$$

R_t = Return (net cash flow)

k* = IRR

C_0 = Initial cost of project

Because the net present-value equation is complex, it is difficult to solve for the actual internal rate of return on an investment without a computer or sophisticated calculator. For this reason,

trial and error method is sometimes used. In general, internal rate of return analysis suggests that projects should be accepted when the *IRR* > *k* and rejected when the *IRR* < *k*. When the *IRR* > *k*, the marginal rate of return earned on the project exceeds the marginal cost of capital. As in the case of projects with an *NPV* > 0 and *PI* > 1, the acceptance of all investment projects with *IRR* > *k* will lead management to maximize the value of the firm. In instances in which capital is scarce and only a limited number of desirable projects can be undertaken at one point in time, the IRR can be used to derive a rank ordering of projects from most desirable to least desirable. Like a rank ordering of all *NPV* > 0 projects from highest to lowest PIs, a rank ordering of potential investment projects from highest to lowest IRRs allows managers to effectively employ scarce funds.

Capital Budgeting Decision Rules

Net Present-value Analysis

- If NPV > 0, the project should be accepted.
- If NPV < 0, the project should be rejected.

Profitability Index or Benefit/cost Ratio Analysis

- PI > 1 indicates a desirable investment.
- PI < 1 indicates an undesirable investment.

Internal Rate of Return Analysis

- Accept when IRR > k; reject when IRR < k.

Capital Budgeting in Practice

- NPV is most recommended measure of a project
- A recent study of 392 large firms (with sales > Rs 1 billion) found that about 75% used the IRR and NPV methods " always and almost always."
- Another study covered 232 small companies (with sales < Rs 5 million) found that payback and accounting rate of return were most frequently used methods, while IRR and NPV lagged far behind.

Corporate capital budgeting and cost of capital estimation are among the most important decisions made by the financial manager. Prior studies extended over the past four decades show financial managers prefer methods such as IRR or non-discounted payback models over NPV (the model academics consider superior). This interesting anomaly has long been a puzzle to the academic community. A recent survey (2002) of the Fortune 1000 Chief Financial Officers finds NPV to be the most preferred tool over IRR and all other capital budgeting tools.

Lesson 44

CAPITAL BUDGETING (CONTINUED)

PROJECT SELECTION

Decision Rule Conflict Problem

- NPV analysis has large project bias.
- With scarce capital, PI method can lead to a better project mix.
- IRR can overstate attractiveness if you can't reinvest excess cash flows at the IRR.

Ranking Reversal Problem

- Ranking reversal occurs when a switch in project standing follows an increase in the relevant discount rate.

Crossover discount rate is the interest factor that equates NPV for two or more projects

When Independent projects are being analyzed, both IRR and NPV criteria give consistent results. "Independent" implies that if a firm is considering several projects at the same time, they can all be implemented simultaneously as long as they pass the IRR and NPV tests and as long as funds are not limited. The adoption of one independent project will have no effect on the cash of another. However, proposals may be mutually exclusive projects. This occurs when two solutions of a particular proposals are offered, only one of which can be accepted.

Conflicting results can be caused by a difference in project size. Table 1 shows such a case. Project A involves an original outlay of Rs 1,500 and Project B is less expensive i.e. only Rs.1, 000. Each project has a 4-year life and no salvage value. The cost of capital is 15%. The IRR of Project A > Project B.

To resolve the dilemma, we calculate the NPV and IRR for an "incremental" (or delta) project. That is, we take the differences between two project cash flows and create a delta project. Both criteria indicate that the additional investment of Rs.500 is worthwhile. It follows that the NPV rule, which suggested Project A, was the correct indicator, and that Project A should be chosen over Project B.

Conflicting rankings occurs when:

1. The initial costs of the two projects differ.
2. The shapes of the cash inflow streams differ significantly.

The reason for the differences between NPV and IRR results is the implicit re-investment assumption. In the NPV calculation, inflows are automatically assumed to be re-invested at the cost of capital (the project's *k*). The IRR solution assumes re-investment at the internal rate of return (the project's k^*).

Table 1

Delta Project

Project	t = 0	t = 1	t = 2	t = 3	t =4	t =5			
		A		1500	580	580	580	580	0
	B	1000	400	400	400	400	0		
Cost of CAP			15%						
	Project A IRR	20.1%		NPV		156			
	Project B IRR	21.9		NPV		142			
		Delta Project							
	(A – B)	500	180	180	180	180	0		
	IRR	16.4%			NPV	14			

Even though alternative capital budgeting decision rules consistently lead to the same project accept/reject decision, they involve important differences in terms of project ranking. Projects ranked most favorably using the NPV method may appear less so when analyzed using the PI or IRR methods. Projects ranked most favorably using the PI or IRR methods may appear less so when analyzed using the NPV technique.

Both NPV and PI methods differ from the IRR technique in terms of their underlying assumptions regarding the reinvestment of cash flows during the life of the project. In the NPV and PI methods, excess cash flows generated over the life of the project are "reinvested" at the firm's cost of capital. In the IRR method, excess cash flows are reinvested at the IRR. For especially attractive investment projects that generate an exceptionally high rate of return, the IRR can actually overstate project attractiveness because reinvestment of excess cash flows at a similarly high IRR is not possible. When reinvestment at the project-specific IRR is not possible, the IRR method must be adapted to take into account the lower rate of return that can actually be earned on excess cash flows generated over the life of individual projects. Otherwise, use of the NPV or PI methods is preferable.

RANKING REVERSAL PROBLEM

A more serious conflict can arise between NPV and IRR methods when projects differ significantly in terms of the magnitude and timing of cash flows. When the size or pattern of alternative project cash flows differ greatly, each project's NPV can react quite differently to changes in the discount rate. As a result, changes in the appropriate discount rate can sometimes lead to reversals in project rankings.

Table 1
NPV Profile: Crossover Discount Rate

CAP-Bud Tech	Build New (A)	Remodel Old (B)
0% D Rate		
NPV	38.4 m	42.1 m
15% D Rate		
NPV	7.7 m	8.3 m
25% D Rate		
NPV	0.99 m	0.03 m
Crossover 18.08%	4.7 m	4.7 m

Figure 1 displays the potential conflict between NPV, PI, and IRR project rankings at various interest rates by showing the effect of discount rate changes on the NPV of each alternative investment project. This **net present-value profile** relates the NPV for each project to the discount rate used in the NPV calculation. Using a k = 0 percent discount rate, the NPV for the "build new plant" investment project is $38.4 million, and it is $42.1 million for the "remodel old plant" alternative. These NPV values correspond to the difference between nominal dollar cash inflows and outflows for each project and also coincide with NPV line Y-axis intercepts of $38.4 million for the "build new plant" project and $42.1 million for the "remodel old plant" alternative. The X-axis intercept for each curve occurs at the discount rate where NPV = 0 for each project. Because NPV = 0 when the discount rate is set equal to the IRR, or when $IRR = k$, the X-axis intercept for the "build new plant" alternative is at the IRR = 25.06 percent level, and it is at the IRR = 23.57 percent level for the "remodel old plant" alternative.

Figure 1

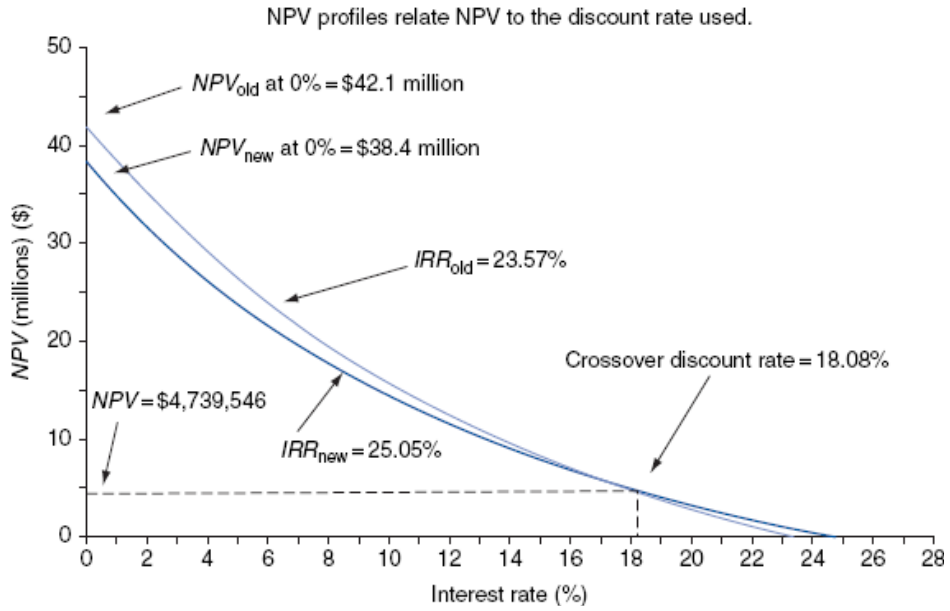


Figure 1 illustrates how ranking reversals can occur at various NPV discount rates. Given higher nominal dollar returns and, therefore, a higher *Y*-axis intercept, the "remodel old plant" alternative is preferred when very low discount rates are used in the NPV calculation. Given a higher IRR and, therefore, a higher *X*-axis intercept, the "build new plant" alternative is preferred when very high discount rates are used in the calculation of NPV. Between very high and low discount rates is an interest rate where NPV is the same for both projects. A reversal of project rankings occurs at the **crossover discount rate**, where NPV is equal for two or more investment alternatives. In this example, the "remodel old plant" alternative is preferred when using the NPV criterion and a discount rate *k* that is less than the crossover discount rate. The "build new plant" alternative is preferred when using the NPV criterion and a discount rate *k* that is greater than the crossover discount rate. This ranking reversal problem is typical of situations in which investment projects differ greatly in terms of their underlying NPV profiles. Hence, a potentially troubling conflict exists between NPV, PI, and IRR methods.

COST OF CAPITAL

The correct discount rate (cost of capital) for each investment project is simply the marginal cost of capital for that project. However, determination of the correct discount rate for individual projects is not an easy task.

THE COST OF DEBT

The cost of debt is easier to explain. It is simply the interest rate that must be paid on the debt. The cost of debt is the return that lenders require to lend their funds to the firm. Since the interest payments made by the firm on borrowed funds are deductible from the firm's taxable income, the after-tax cost of borrowed funds to the firm (k_d) is given by the interest paid (r) multiplied by 1 minus the firm's marginal tax rate, t. That is,

$k_d = r(1 - t)$

For example, if the firm borrows at a 12.5 percent interest rate and faces a 40 percent marginal tax rate on its taxable income, the after-tax cost of debt capital to the firm is

$k_d = 12.5\%(1- 0.40) = 7.5\%$

THE COST OF EQUITY CAPITAL: THE RISK – FREE RATE PLUS PREMIUM

The cost of equity capital is the rate of return that stockholders require to invest in the firm. The cost of raising equity capital externally usually exceeds the cost of raising equity capital internally by the flotation costs (i.e., the cost .of issuing the stock).

One method employed to estimate the cost of equity capital (k_e) is to use the risk free rate (r_f) plus a risk premium (r_p). That is,

$k_e = r_f + r_p$

Since dividends vary with the firm's profits, stocks are more risky than bonds so that their return must include an additional risk premium. If the premiums associated with these two types of risk are labeled P_1 and P_2, we can restate the formula for the cost of equity capital as:

$k_e = r_f + p_1 + p_2$

THE COST OF EQUITY CAPITAL: THE DIVIDEND VALUATION MODEL

The equity cost of capital to a firm can also be estimated by the dividend valuation model. To derive this model, we begin by pointing out that, with perfect information, the value of a share of the common stock of a firm should be equal to the present, value of all future dividends expected to be paid on the stock, discounted at the investor's required rate of return (k_e). If the dividend per share (D) paid to stock-holders is expected to remain constant over time, the present value of a share of the common stock of the firm (P) is then:

$$P = \sum_{t=1}^{\infty} \frac{D}{(1+k_e)^t} = \frac{D}{k_e}$$

$$k_e = \frac{D}{P}$$

Cost of Equity Capital (k_e): Dividend Valuation Model

P = price of a share of stock

D = Constant dividend per share

k_e = Required rate of return

If dividends are assumed to remain constant over time and to be paid indefinitely, Equation is nothing else than an annuity and can be rewritten as:

$P = D / k_e$

If dividends are instead expected to increase over time at the annual rate of g, the price of a share of the common stock of the firm will be greater and is given by:

$$P = \frac{D}{K_e - g}$$

Solving Equation for k_e, we get the following equation to measure the equity cost of capital equation to the firm:

$$k_e = \frac{D}{P} + g$$

P = Price of a share of stock

D = Price of a share of stock

K_e = Required rate of return

g = Growth rate of dividends

For example, if the firm pays a dividend of $20/share and the growth rate of dividend payments is expected to be 5 %/year, the cost of equity capital for this firm is:
k_e = $20 / $200 + 0.05 = 0.10 + 0.05
= 0.15 or 15%

THE COST OF EQUITY CAPITAL: THE CAPITAL ASSET PRICING MODEL (CAPM)
This method takes into consideration not only the risk differential between common stocks and government securities but also the risk differential between the common stock of the firm and the average common stock of all firms or broad-based market portfolio. The risk differential between common stocks and government securities is measured by (k_m - r_f), where' km is the average return on all common stocks and r_f is the return on government securities.

A beta coefficient of 1 means that the variability in the returns on the common slack of the firm is the same as the variability in the returns on all stocks. Thus, investors holding the stock of the firm face the same risk as holding a broad-based market portfolio of all stocks. A beta coefficient of 2 means that the variability in the returns on (i.e., risk of holding) the stock of the firm is twice that of the average stock. On the other hand, holding a stock with a beta coefficient of 0.5 is half as risky as holding the average stock.

The cost of equity capital to the firm estimated by the capital asset pricing model is measured by
$\mathbf{k_e = r_f + \beta (k_m - r_f)}$
Where $\mathbf{k_e}$ is the cost of equity capital to the firm, **rj** is the risk-free rate, **β** is the beta coefficient and $\mathbf{k_m}$ is the average return on the stock of all firms. Thus, the CAPM postulates that the cost of equity capital to the firm is equal to the sum of the risk free rate plus the beta coefficient (**β)** times the risk premium on the average stock (k_m = r_f). Note that multiplying **β** by (k_m - r_f) gives the risk premium on holding the common stock of the particular firm.

For example, suppose that the risk-free rate (r_f) is 8 percent, the average return on common stocks (k_m) is 15 percent, and the beta coefficient (P) for the firm is 1. The cost of equity capital to the firm (k_e) is then
$\mathbf{k_e = 8\% + 1(15\% - 8\%) = 15\%}$

That is, since a beta coefficient of 1 indicates that the stock of this firm is as risky a the average stock of all firms, the equity cost of capital to the firm is 15 percent (the same as the average return on all stocks). If **β** = **1.5** for the firm (so that the risk involved in holding the stock of the firm is 1.5 times larger than the risk on the average stock), the equity cost of capital to the firm would be
$\mathbf{k_e = 8\% + 1.5(15\% - 8\%) = 18.5\%}$
On the other hand, if **β** = 0.5,
$\mathbf{k_e = 8\% + 0.5(15\% - 8\%) = 11.5\%}$
Firms usually use all three methods and then attempt to reconcile the differences and arrive at a consensus equity cost of capital for the firm.

THE WEIGHTED COST OF CAPITAL
In general, a firm is likely to raise capital from undistributed profits, by borrowing, and by the sale of stocks, and so the marginal cost of capital to the firm is a weighted average of the cost of raising the various types of capital.
Since the interest paid on borrowed funds is tax deductible while the dividend paid on stocks are not, the cost of debt is generally less than the cost of equity capital. The risk involved in raising funds by borrowing, however, is greater than the risk on equity capital because the firm must regularly make payments of the interest and principal on borrowed funds before paying

dividends on stocks. Thus, firms do not generally raise funds only by borrowing but also by selling stock (as well as from undistributed profits).

Firms often try to maintain or achieve a particular long-term capital structure of debt to equity. For example, public utility companies may prefer a capital structure involving 60 percent debt and 40 percent equity, while auto manufacturers may prefer 30 percent debt and 70 percent equity. The particular debt/equity ratio that a firm prefers reflects the risk preference of its managers and stockholders and the nature of the firm's business. Public utilities accept the higher risk involved in a higher debt/equity ratio because of their more stable flow of earnings than automobile manufacturers. When, a firm needs to raise investment capital, it borrows and it sells stocks so as to maintain or achieve a desired debt/equity ratio.

The composite cost of capital to the firm (k_c) is then a weighted average of the cost of debt capital (k_d) and equity capital (k_e) as given by:

$$k_c = w_d k_d + w_e k_e$$

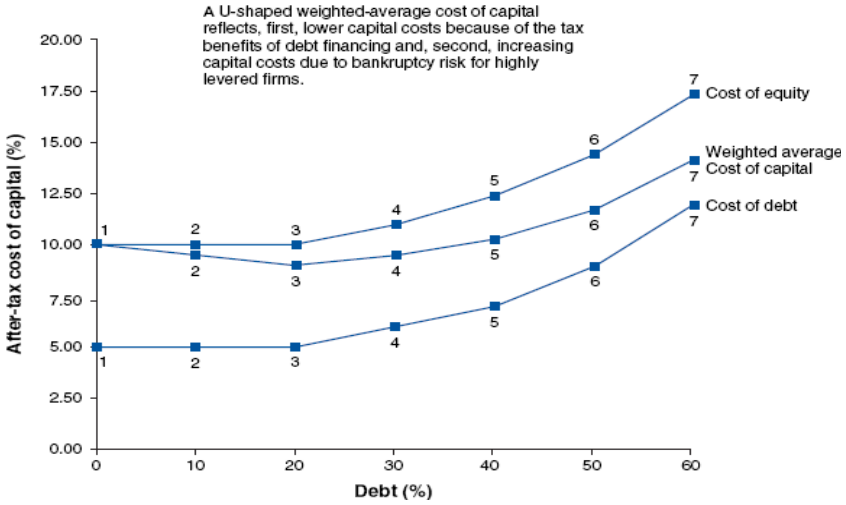
where Wd and We are, respectively, the proportion of debt and equity capital in the firm's capital structure. For example, if the (after-tax) cost of debt is 7.5 percent, the cost of equity capital is 15 percent, and the firm wants to have a debt/equity ratio of 40:60, the composite or weighted marginal cost of capital to the firm will be

k_c = (0.40)(7.5%) + (0.60)(15%) = 3% + 9% = 12% .

This is the composite marginal cost of capital that we have used to evaluate all the proposed investment projects.

Figure 2 shows how the cost of capital changes as the debt ratio increases for a hypothetical industry with about average risk. The average cost of capital figures in the graph are calculated in Table 15.6. In the figure, each dot represents one of the firms in the industry. For example, the dot labeled “one” represents firm 1, a company with no debt. Because its projects are financed entirely with 10 percent equity money, firm 1’s average cost of capital is 10 percent.

Figure 2



Firm 2 raises 10 percent of its capital as debt, and it has a 4.5 percent after-tax cost of debt and a 10 percent cost of equity. Firm 3 also has a 4.5 percent after-tax cost of debt and 10 percent cost of equity, even though it uses 20 percent debt. Firm 4 has an 11 percent cost of equity and a 4.8 percent after-tax cost of debt. Because it uses 30 percent debt, a before-tax debt risk premium of 0.5 percent and an equity risk premium of 1 percent have been added to account for the additional risk of financial leverage. Notice that the required return on both debt and equity rises with increasing leverage for firms 5, 6, and 7. Providers of debt and equity capital typically believe that because of the added risk of financial leverage, they should obtain higher yields on

the firm's securities. In this particular industry, the threshold debt ratio that begins to worry creditors is about 20 percent. Below the 20 percent debt level, creditors are unconcerned about any risk induced by debt; above 20 percent, they are aware of higher risks and require compensation in the form of higher expected rates of return.

OPTIMAL CAPITAL BUDGET

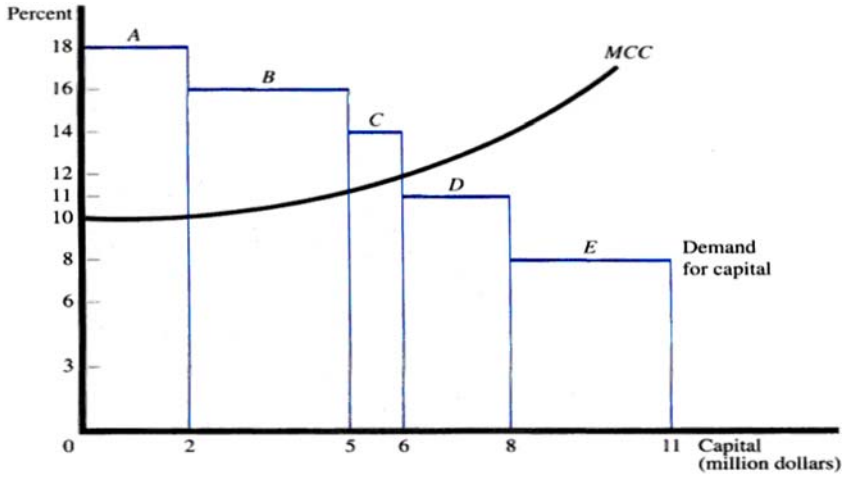
A profit-maximizing firm operates at the point where marginal revenue equals marginal cost. In terms of the capital budgeting process, this implies that the marginal rate of return earned on the last acceptable investment project is just equal to the firm's relevant marginal cost of capital. The **optimal capital budget** is the funding level required to guarantee a value-maximizing level of new investment.

Investment Opportunity Schedule (IOS)

- IOS shows the pattern of returns (IRR) for all potential investment projects.
- Marginal cost of capital is the extra financing cost necessary to fund an additional investment project.
- Optimality requires setting IRR = MCC.

Capital budgeting is essentially an application of the general principle that a firm should produce the output or undertake an activity until the marginal revenue from the output or activity is equal to its marginal cost. In a capital budgeting framework, this principle implies that the firm should undertake additional investment projects until the marginal return from the investment is equal to its marginal cost. The schedule of the various investment projects open to the firm, arranged from the one with the highest to the lowest return, represents the firm's demand for capital. The marginal cost of capital schedule, on the other hand, gives the cost that the firm faces in obtaining additional amounts of capital for investment purposes. The intersection of the demand and marginal cost curves for capital that the firm faces determines how much the firm will invest. This is shown in Figure 3.

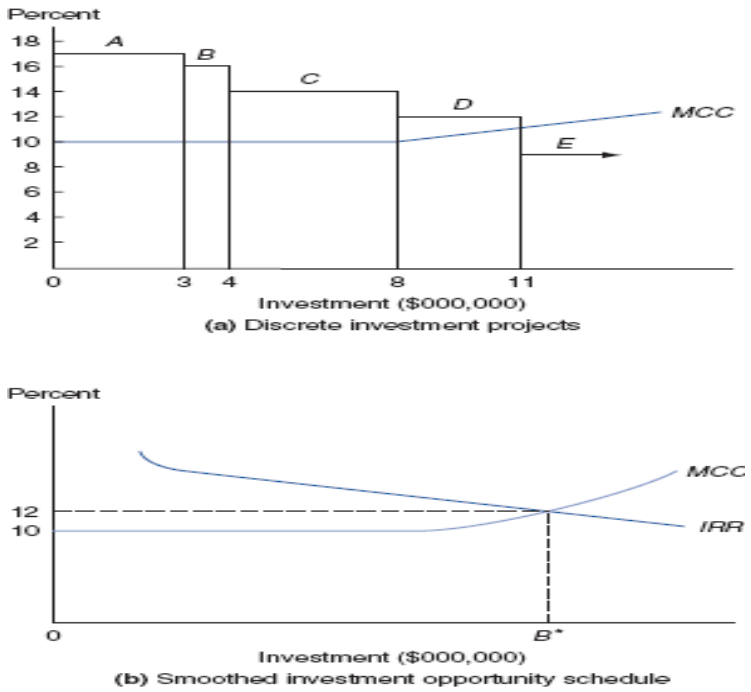
Figure 3



In Figure 3, the various lettered bars indicate the amount of capital required for each investment project that the firm can undertake and the rate of return expected on each investment project. Faced with the demand for capital and marginal' cost of capital curves shown in Figure 3, the firm will undertake projects A, B, and C because the expected rates of return on these projects exceed the cost of capital to make these investments. Specifically, the firm will undertake

project A because it expects a return of 18 percent from the project as compared with its capital cost of only 10 percent. Similarly, the firm undertakes project B because it expects a return of 16 percent as compared with capital cost of between 10 and 11 percent for the project. The firm also undertakes project C because its expected return of 14 percent exceeds its capital cost of nearly 12 percent. On the other hand, the firm will not undertake projects D and E because the expected rate of return on these projects is lower than the cost of raising the capital to make these investments.

Figure 4



The **investment opportunity schedule (IOS)** shows the pattern of returns for all of the firm's potential investment projects. Figure 4(a) shows an investment opportunity schedule for a firm. The horizontal axis measures the dollar amount of investment commitments made during a given year. The vertical axis shows both the rate of return earned on each project and the percentage cost of capital. Each box denotes a given project. Project *A*, for example, calls for an outlay of $3 million and promises a 17 percent rate of return; project *B* requires an outlay of $1 million and promises a 16 percent yield, and so on. The firm's IOS is depicted in Figure 4(b) generalizes the IOS concept to show a smooth pattern of potential returns. The curve labeled *IRR* shows the internal rate of return potential for each project in the portfolio of investment projects available to the firm. It is important to remember that these projects are ranked by IRR from highest to lowest. Therefore, project *A* is more attractive than project *E*, and the IRR schedule is downward sloping from left to right.

The **marginal cost of capital (MCC)** is the extra financing cost necessary to fund an additional investment project, expressed on a percentage basis. Given these IOS and MCC schedules, the firm should accept projects *A* through *D*, obtaining and investing $11 million. Project *E*, the government bond investment alternative, should be rejected. The smooth curves in Figure 4(b) indicate that the firm should invest *B** dollars, the optimal capital budget. At this investment level, the marginal cost of capital is 12 percent, exactly the same as the IRR on the marginal investment project. Whenever the optimal capital budget *B** is determined, the IRR always equals the MCC for the last project undertaken. The condition that must be met for any budget

to be optimal is that *IRR* = *MCC*. This means that the final project accepted for investment is a breakeven project.

Lesson 45

GOVERNMENT IN THE MARKET ECONOMY

OVERVIEW

- **Government Intervention**
- **Market Failure**
- **Market Power**
- **Natural Monopolies**
- **Externalities**
- **Solving Externalities**
- **Public Goods**
- **Asymmetric Information**

WHY GOVERNMENT INTERVENTION?

Throughout this course of Managerial Economics, we have treated the market as a place where consumers and sellers come together to trade goods and services without government Intervention. But as we know, rule and regulations passed and implemented by the government, enter in to almost every decision consumers and sellers make. As a manager, it is important to understand regulations passed by the government, and how they affect optimal managerial decision-making.

Competitive markets can achieve social economic efficiency without government regulation. Competitive markets can do a number of desirable things for society. Under perfect competition, the firms produce the right amount of goods and services, charge the right price. They cannot control the prices because there are many sellers and they sell almost identical products. As a result, prices in competitive markets are determined by the market forces of demand and supply. In the long run consumers enjoy the very lowest prices. But, not all markets are competitive, and even competitive markets can sometimes fail to achieve maximum social surplus.

MARKET FAILURE

One of the main reasons for government Intervention in the market is that free markets do not always result in the socially efficient quantities of goods and services at socially efficient prices. Market failure occurs when a market fails to achieve economic efficiency and as a result, fails to maximize social surplus. And here comes in government to play its role.

In the absence of market failure, however, no efficiency argument can be made for government intervention in competitive markets. Government intervention is justified when it is undertaken to overcome market failures.

FIVE MAIN FORMS OF MARKET FAILURE CAN DAMAGE ECONOMIC EFFICIENCY:

1. Market power
2. Natural monopoly
3. Negative (& positive) externalities
4. Public goods
5. Information problems

MARKET POWER

A firm has market power when it sells output at a price that exceeds its marginal cost of production. Monopoly power or a high degree of market power can arise in three ways:

I. Actual or attempted monopolization
II. Price-fixing cartels

III. Mergers among horizontal competitors

In our lecture # 27, we have already discussed in detail the social cost of monopoly. Just to refresh, we'll only consider the main points here.

SOCIAL COSTS OF MONOPOLY

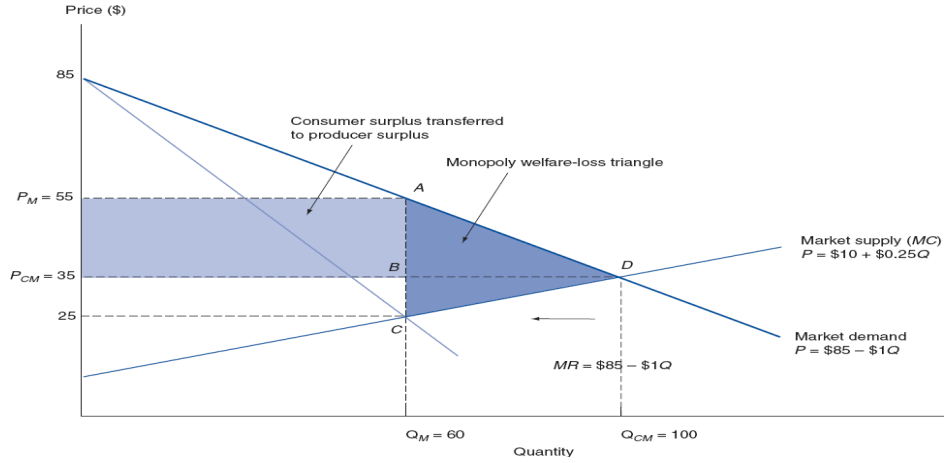
- Monopolists produce too little output.
- Monopolists charge prices that are too high, P > MC.

DEADWEIGHT LOSS FROM MONOPOLY

- Monopoly markets create a loss in social welfare due to the decline in mutually beneficial trade activity.
- There is also a wealth transfer problem associated with monopoly; consumer surplus is transferred to producer surplus

This deadweight loss from monopoly is shown in Figure 1 as a monopoly welfare-loss triangle.

Figure 1



PROMOTING COMPETITION THROUGH ANTI TRUST POLICIES

To reduce the cost of market failure, caused by market power, most countries rely on Anti Trust Policies as they are known in the US and Canada. In some countries they are known as Competition Policies.

In Pakistan, we have various regulatory authorities such as PEMRA (Pakistan Electronic Media regulatory authority), OGRA (Oil and Gas regulatory authority). The Competition Commission of Pakistan (CCP) was established in 2007. Major aim of the CCP is to provide a legal framework to create a business environment based on healthy competition for improving economic efficiency, developing competitiveness and protecting consumers from anti-competitive practices. Prior to the CCP, Pakistan had an anti-monopoly law namely 'Monopolies and Restrictive Trade Practices (Control and Prevention) Ordinance' (MRTPO) 1970.

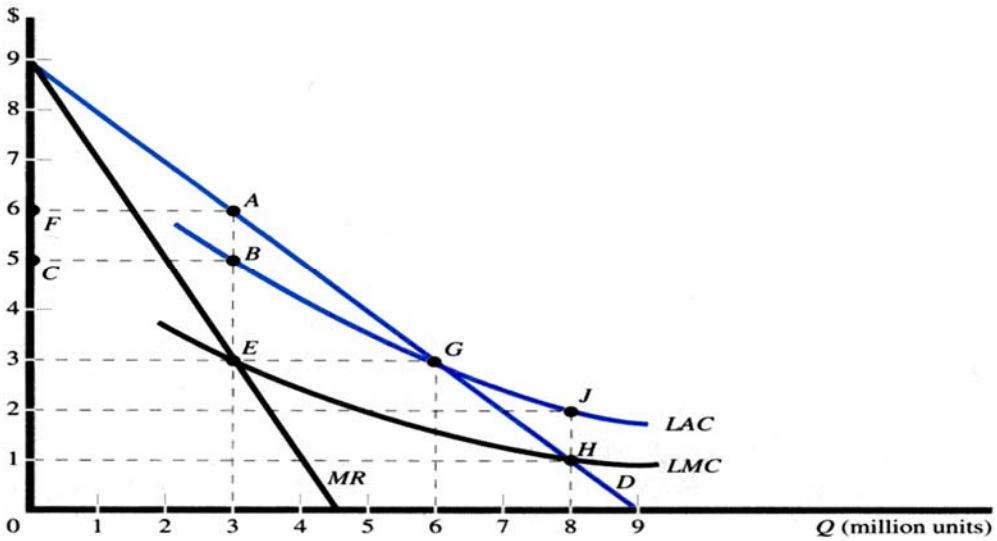
NATURAL MONOPOLY

When a single firm can produce total consumer demand for a good or service at a lower long-run total cost than if two or more firms produce total industry output, it is called a natural monopoly. Breaking up a natural monopoly is undesirable as increasing number of firms drives up total cost & damages productive efficiency. In the presence of large economies of scale, it may be desirable for a single firm to serve the entire market. So that the government may allow

a firm to exist as a monopoly but choose to regulate the price to reduce deadweight loss from the monopoly. This situation is called a natural monopoly and it might cause market failure.

In some industries, economies of scale operate (i.e., the long-run average cost curve may fall) continuously as output expands, so that a single firm could supply the entire market more efficiently than many small firms. Such a large firm supplying the entire market is called a natural monopoly as mentioned earlier. The distinguishing characteristic of a natural monopoly is that the firm's long-run average cost curve is still declining when the firm supplies the entire market. Monopoly in this case is the natural result of a larger firm having lower costs per unit than smaller firms. Examples of natural monopolies are public utilities (electrical, gas, water, telecommunication services and transportation companies). To have more than one such firm in a given market would lead to duplication of supply lines and to much higher costs per unit as these public utilities require extremely costly infrastructure. To avoid this needless duplication of supply lines, governments usually allow a single firm to operate in the market but regulate the price and quality of the services provided, so as to allow the firm only a normal risk-adjusted rate of return on its investment. This is shown in Figure 2.

Figure 2



In Figure 2, the D and MR curves are, respectively, the market demand and marginal revenue curves for the service faced by the natural monopolist, while the LAC and LMC curves are its long-run average and marginal cost curves. If unregulated, the best level of output of the monopolist in the long run would be 3 million units per time period and is given by point E, at which the LMC and MR curves intersect. For Q = 3 mi1lion units, the monopolist would charge the price of $6 (point A on the D curve) and incur at LAC = $5 (point B on the LAC curve), thereby earning a huge profit of $1 (AB) per unit and $3 million (the area of rectangle ABCF) in total. Note that at Q = 3 million units, the LAC curve is still declining. We can also notice that at the output level of 3 million units, P > LMC, so that more of the service is desirable from society's point of view. That is, the marginal cost of the last unit of the service supplied is smaller than the value of the service to society, as reflected by the price of the service. There is, however, no incentive for the unregulated monopolist to expand output beyond Q = 3 million units per time period because its profits are maximized at Q = 3 million.

To ensure that the monopolist earns only a normal rate of return on its investment, the regulatory commission usually sets the price at which P = LAC. In Figure 2, this is given by point

G, at which P = LAC = $3 and output is 6 million units per time period. While the price is lower and the output is greater than at point A, P > LMC at point G. The best level of output from society's point of view would be 8 million units per time period, as shown by point H, at which P = LMC = $1. At Q = 8 million, however, the LAC = $2 (point J on the LAC curve), and the public utility company would incur a loss of $1 (HJ) per unit and $8 million per time period. As a result, the public utility would not supply the service in the long run without a per-unit subsidy of $1 per-unit. In general, regulatory commissions set P = LAC (point G in Figure 2) so that the public utility company breaks even in the long run without a subsidy. Apparently a Public Utility Regulation seems fairly simple, but actually the price (or rates) determination for public utilities is quite a complex task. Since the public utility company supply the service to different classes of customers, each with different price elasticities of demand, many different rate schedules could be used to allow the public utility company to break even. Under natural monopoly, no single price can establish social economic efficiency. To determine a fair price, a regulatory commission must estimate a fair or normal rate of return, given the risk inherent in the enterprise. The commission then approves prices that produce the target rate of return on the required level of investment.

EXTERNALITIES

An ***externality*** is a cost or benefit resulting from some activity or transaction that is imposed or bestowed upon parties outside the activity or transaction. Sometimes called ***spillovers*** or ***neighborhood effects*** or ***side effects*** of economic activities. When actions taken by market participants create either benefits or costs that spill over to other members of society

- *Positive externalities* occur when spillover effects are beneficial to society
- *Negative externalities* occur when spillover effects are costly to society

NEGATIVE EXTERNALITIES

External Diseconomies of Production or Consumption or Uncompensated costs. If economic activity among producers and consumers harms the well-being of a third party who is not compensated for any resulting damages, a negative externality is said to exist. Environmental pollution is one well-known negative externality. Air pollution, water pollution and noise pollution are the most familiar types of negative externalities. Sometimes we do not even think about polluting the water, for example, washing a car in our driveway. Similarly many firms are doing the same but not un-knowingly. Paper making require poisonous solvents and chlorine compounds to bleach paper products. A firm that produces textiles usually creates waste products that contain dioxin, a cancer-causing chemical. When a textiles manufacturing firm can dispose of this waste “for free” by dumping into a nearby river, this can create great problems for the community, birds, fish and other water animals of the river. While the firm benefits from dumping waste into the river, the waste reduces the oxygen content of the water.

Externalities undermine allocative efficiency because market participants rationally choose to ignore the benefits & costs of their actions that spill over to others. Managers rationally ignore external costs when making profit-maximizing production decisions. Competitive market prices do not capture social benefits or costs that spill over to society. As a result, externalities can cause market failures. Negative externalities cause overproduction because sellers do not consider all social costs. In short, externalities lead to a difference between private and social costs and benefits Producers that generate negative externalities do not pay the full costs of production and tend to over utilize social resources.

POSITIVE EXTERNALITIES

If economic activity among producers and consumers helps the well-being of a third party who is does not pay for any resulting benefits, a Positive externality is said to exist. Positive

externalities in production can result when a firm trains employees who later apply their knowledge in work for other firms. Education generates positive externalities. Positive externalities cause underproduction because sellers cannot reflect the full social value of production in the prices charged.

Figure 3

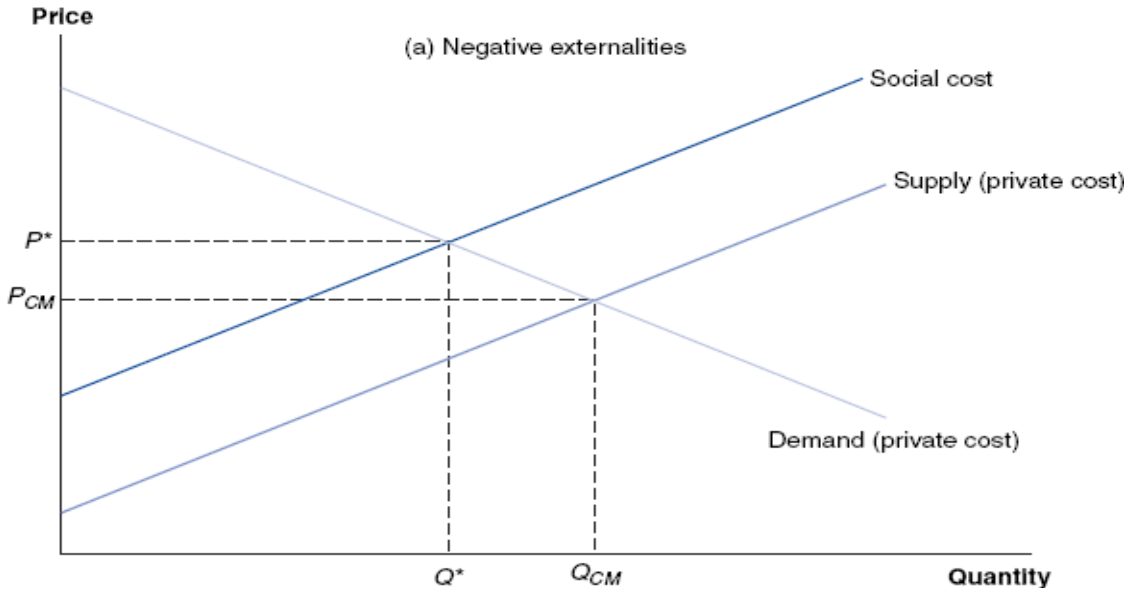
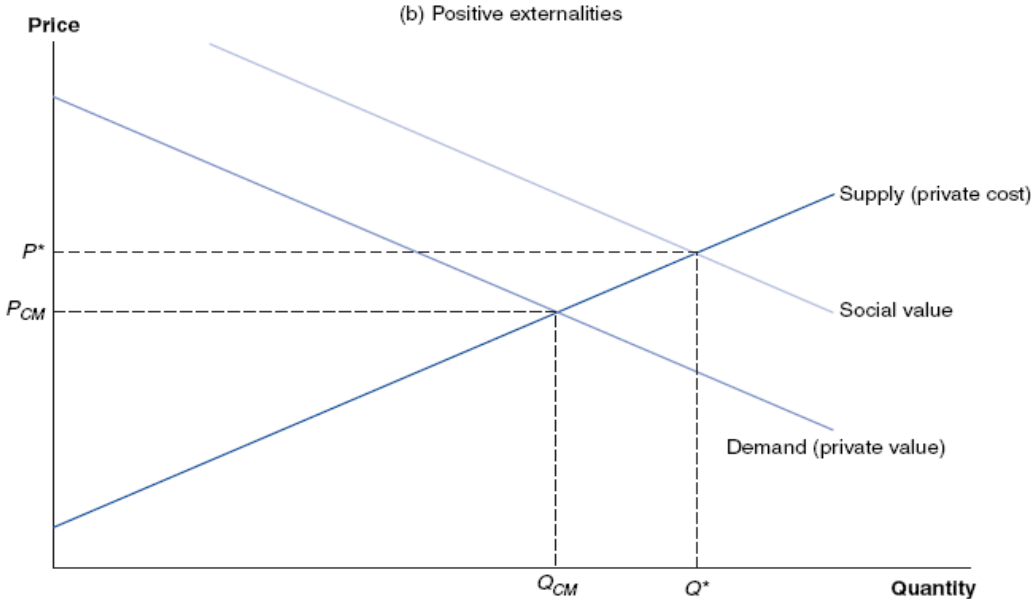


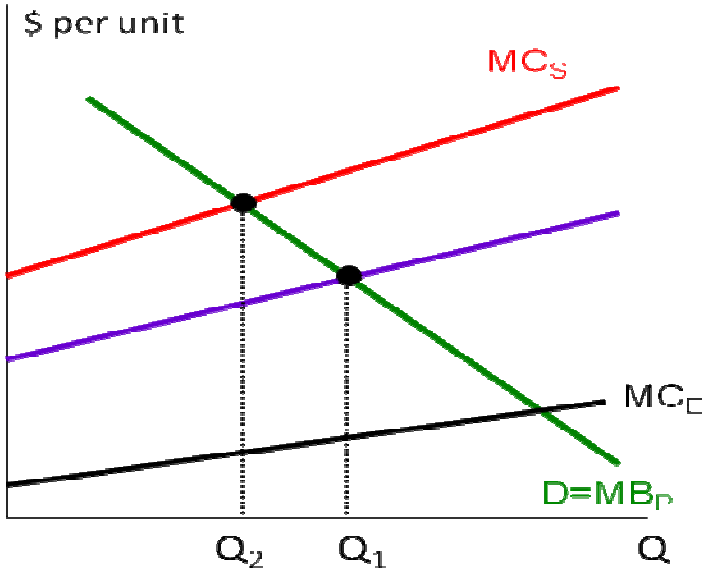
Figure 4



From a social welfare perspective, the problem created by negative externalities and positive externalities are presented in Figure 3 and 4 respectively. In a competitive market without externalities, the demand curve reflects the value to buyers and the supply curve reflects costs

to sellers. In the absence of externalities, the competitive market equilibrium maximizes consumer surplus. In the absence of externalities, the competitive market equilibrium is efficient, however this all changes in the face of negative externalities. ***Marginal social cost* (*MSC*)** is the total cost to society of producing an additional unit of a good or service. *MSC* is equal to the sum of the marginal costs of producing the product and the correctly measured damage costs involved in the process of production.

Figure 5



If a steel producing firm emits pollutants into the water as a by-product, a cost, or negative externality is borne by the society in the form of dirtier water shown by MC_P.
Suppose that MC_E in Figure 5 is the marginal external cost of producing steel, its production generates pollution. MC_P is the marginal private cost of producing steel. The industry Supply curve that is the sum of MC curves of the firms. The vertical sum of MC_E and MC_P is the marginal social cost MC_S. D is the demand for steel. When the pollution is ignored, the amount of the steel produced is Q_1, where the D (MBp or marginal private benefit) equals the MC_P. However, the efficient amount from the viewpoint of society is only Q_2, where MC_S (S) equals D i.e. MBp. Since the firms are dumping pollution into the water for free, the cost of pollution is not internalized by those who buy and sell steel. If the firms internalize the cost of pollution, the sum of their marginal cost curves would be the vertical sum of the MC_E and MC_P (purple curve), shown as MC_S in Figure 5. The socially efficient level of output is Q_2.

SOLVING EXTERNALITIES
GOVERNMENT SOLUTIONS

- Internalizing the Externality: Altering incentives so that people take account of the external effects of their actions by defining the property rights clearly.
- Taxes are often used to correct negative externalities.
- Government sometimes controls the effects of externalities by command and control regulation.

The basic reason for the “market failure” is the absence of well-defined property rights. The steel firms believe that they have the right to use the river to dump waste, and environmentalists

believe that they have the right to a river. So the government may force firms to internalize the externalities.

Taxes are often used to correct negative externalities. Incentive-based solutions to the externality problem originated with Pigou, who suggested that the most direct solution would be to tax the externality-creating entity. The amount of such taxes must be the estimated precisely equal to the amount of social costs as shown in Figure 6 and 7.

FIGURE 6

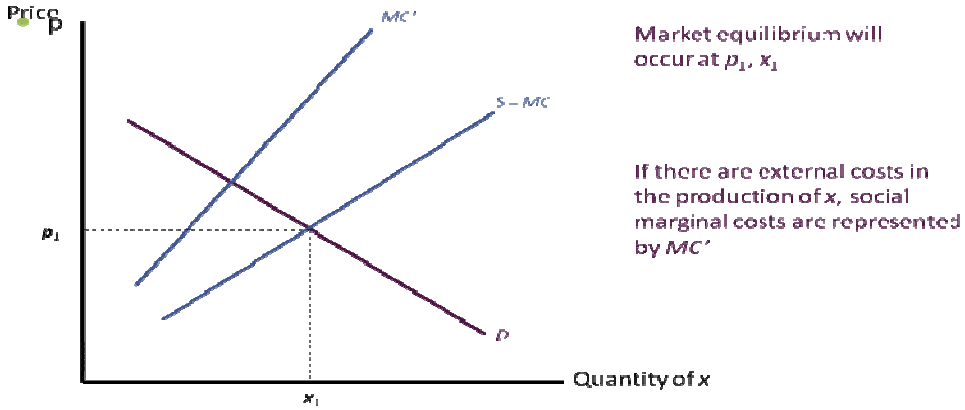
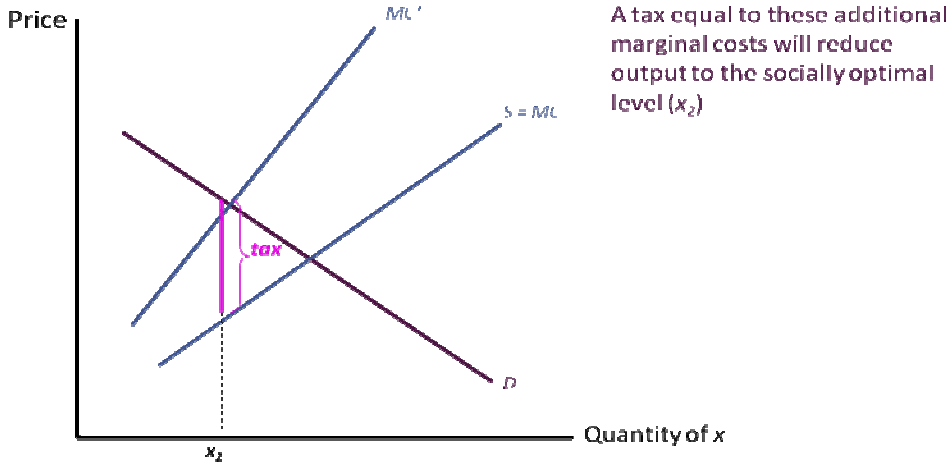


Figure 7



THE COASE THEOREM

Government need not be involved in every case of externality. Private bargains and negotiations are likely to lead to an efficient solution in many social damage cases without any government involvement at all. This argument is referred to as the *Coase Theorem*. According to Coase Theorem, an acceptable solution to an externality will be found if:

- ownership of property is clearly defined,
- the number of people involved is small,
- The costs of bargaining are negligible.

PUBLIC GOODS

Public Goods (***social*** or ***collective goods***) A good that is *non rival and non exclusionary in* consumption.

- Public goods are non rival in consumption. Use by certain individuals does not reduce availability for others.
- Public goods tend to be non exclusionary. It is often impossible to confine benefits to paying customers.

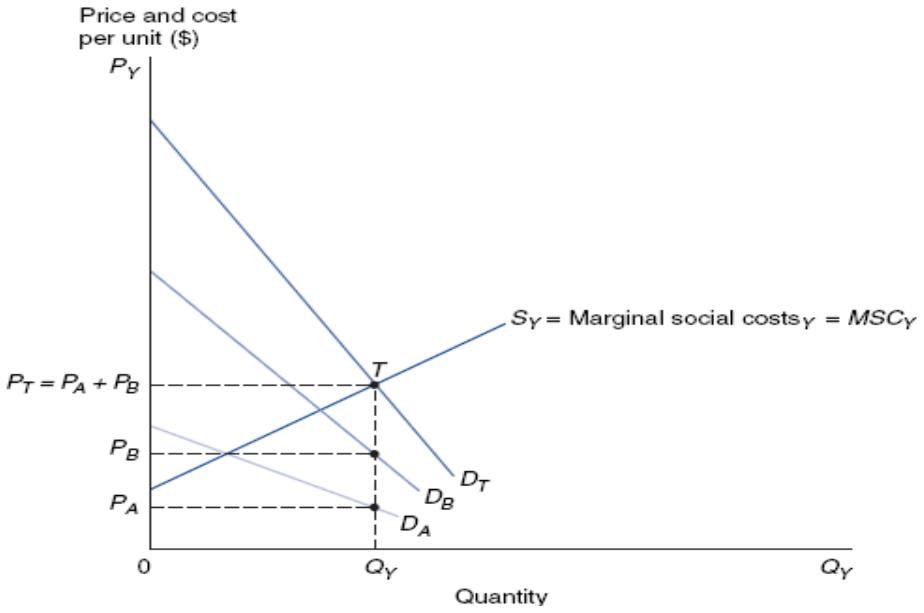
Non rival: A good which when consumed by one person does not preclude other people from also consuming the good. Example: radio signals, national defense, dams, airports, highways, trash collection. National defense – everyone is protected by the same defense system. Dams – everyone in the community is protected from flooding by the dam.

Non exclusionary: No one is excluded from consuming the good once it is provided. Example: clean air, Government regulation and antitrust policy are often used to protect consumers, workers, and the environment; to discourage and regulate monopoly; and to overcome problems posed by externalities such as pollution. Another important function of government is to provide goods and services that cannot be provided and allocated in optimal quantities by the private sector. Public goods have characteristics that make it difficult for the private sector to produce them profitably (market failure) because of two main problems:

I. **"Free Rider" Problem** Individuals have little incentive to buy a public good because of their non rival & non exclusionary nature.

II. **Hidden preferences Problem** can exist for public goods. There is an incentive not to reveal true valuation of the good or sevice, since if the good is provided, public is going to get it anyway. But if everyone refuses to reveal their true value of the good and so refused to voluntarily pay what it is worth, the good will not be provided. Someone who enjoys the benefit of a good without paying for it is called a free rider.

Figure 8



Because public goods can be enjoyed by more than one consumer at the same point in time, the aggregate or total demand for a public good is determined through the vertical summation of the demand curves of all consuming individuals. As shown in Figure 8, *DA* is the demand curve of consumer A, and *DB* is the demand curve of consumer B for public good *Y*. If consumers A and B are the only two individuals in the market, the aggregate demand curve for public good *Y*, *DT* is obtained by the vertical summation of *DA* and *DB*. This contrasts with the market demand curve for any private good, which is determined by the horizontal summation of individual demand curves. Given market supply curve *SY* for public good *Y* in Figure 8, the optimal amount of *Y* is *QY* units per time period given by the intersection of *DT* and *SY* at point *T*.

$$MSB_Y = MSC_Y$$

At point *T*, the sum of marginal benefits enjoyed by both consumers equals the marginal social cost of producing *QY* units of the public good. That is, *PT* = *PA* + *PB* = *MCY*. Although the optimal quantity is *QY* units in Figure 8, there are two related reasons why less than this amount is likely to be supplied by the private sector. First, because individuals not paying for public good *Y* cannot be excluded from consumption, there is a tendency for consumers to avoid payment responsibility. A **free-rider problem** emerges because each consumer believes that the public good will be provided irrespective of his or her contribution toward covering its costs. Second, a **hidden preferences problem** also emerges in the provision of public goods because individuals have no economic incentive to accurately reveal their true demand.

INCOMPLETE INFORMATION

Participants in a market that have incomplete information about prices, quality, technology, or risks may be inefficient. The Government serves as a provider of information to fight the inefficiencies caused by incomplete and/or asymmetric information. Information is non rival in consumption. When information is very costly for individuals to collect and disperse, it may be cheaper for government to produce it once for everybody.

RENT SEEKING

Government policies can improve the allocation of resources in the economy by alleviating the problems associated with market power, natural monopolies, externalities, public goods and asymmetric information. However, government policies generally benefit some parties at the expense of others. For this purpose, lobbyists spend large sums of money in an attempt to affect these policies. This process is known as *rent-seeking.*

CONCLUSION

Market power, externalities, public goods, and incomplete information create a potential role for government in the marketplace. However, government's presence creates rent seeking incentives, which may damage its ability to improve matters and results in government's failure.

Nobel Laureate Gary S. Becker argues that corruption is a common denominator whenever big government infiltrates all facets of economic life. Bribery and illegal favor-seeking do considerable damage. They always divert resources away from the production of useful goods and services.

One way to discourage corruption is to vote out crooked politicians and punish people in business who illegally influence the political process. Becker argues that the only way to permanently reduce undesirable business influence over the political process is to simplify and standardize needed regulations.