UNIT 1 MATHEMATICAL MODELLING – AN OVERVIEW

Structure **Page No**

1.1 INTRODUCTION

Mathematics is a very effective tool in solving real world problems. The critical step in the use of mathematics for solving real world problems is the building of a suitable mathematical model. A mathematical model is a conversion of a real world problem into an abstract mathematical problem involving mathematical concepts such as constants, variables, functions, equations, inequalities, etc. The process by which a real world problem is represented and interpreted in terms of a mathematical model is called **mathematical modelling**. Our main aim in this unit is to introduce you to the basic concepts of mathematical modelling and discuss the process of development of a mathematical model.

Translation into a mathematical model is one of the many approaches to solving real-world problems. Other approaches include experimentations with physical models or schematic models or with the real world directly. The mathematical approach, which is our main concern throughout this course, has a number of advantages which we shall be illustrating through examples. However, to give you a broader idea about various models, we shall start the unit by discussing model classifications in Sec. 1.2. In Sec. 1.3 we shall discuss the process of translating a real world problem into a mathematical problem. The need for mathematical modelling is also illustrated in this section through various examples. Sec. 1.4 introduces you to different types of modelling.

Objectives

After studying this unit, you should be able to

- define a mathematical model;
- realise the need for developing a mathematical model;
- translate a real world problem into its equivalent mathematical formulation;
- identify the type of modelling to be used for a given problem.

1.2 MODEL CLASSIFICATIONS

A model is an abstraction of reality or a representation of a real object or situation. It could be a simple drawing of office plans or a complicated functional representation of a complex machinery part. A model airplane may be assembled from children's kit, or it may actually contain an engine and a rotating propeller that enable it to fly like a real plane.

Some models are replicas of the physical properties (relative shape, form, and weight) of the object they represent. Some are physical models but do not have the same physical appearance as the object of their representation. Other type of models deal with symbols, expressions, mathematical equations and inequalities. Each of these models can be classified into four main categories: physical models, schematic models, verbal models and mathematical models.

Physical Models

Physical models are prototypes models that look like the objects they represent. They are more or less the exact replicas of the object being modelled. Scale models of Taj Mahal, airplanes, buses, ships, office complexes, shopping centres, homes, etc. look exactly like their counterparts but in much smaller scale. The advantage in having a scaled model is that one can tell exactly what the object under study looks like, in three dimensions, before making a major investment. Also, some of these models can even perform as their counterparts would and this allows you to conduct the study on the model to see how it might perform under actual operating conditions. Scaled models of airplanes can be tested in wind tunnels to determine aerodynamic properties and the effects of air turbulence on their outer surfaces. Models of bridges and dams can be subjected to multiple levels of stress from wind, heat, cold and other factors to test their effects as endurance and safety. Scaled models that behave in a manner similar to the real objects are less expensive to create and test than their actual counterparts.

Schematic Models

Schematic models are more abstract than physical models. They do not look like the physical reality they represent. Graphs and charts are schematic models that provide pictorial representations of mathematical relationships. Mathematical linear relationship between two variables may be indicated by plotting a line on a graph. Pie charts, bar charts and histograms can all model some real situations but do not bear any physical resemblance to them. Diagrams, drawings, blueprints and a flow chart describing a computer program are all examples of schematic models.

Verbal Models

Verbal models use words to represent some object, situation or problem that exists, or could exist, in reality. This could be a simple word presentation of scenery described in a book to a complex business problem (described in words and numbers). Verbal models provide all relevant and necessary information to solve the problem, make recommendations and suggest alternatives. The case studies which you must have studied from management text books are examples of verbal models that expose you to the workings of a business

without having to visit the firm's actual premises. Often these verbal models provide enough information to be converted into mathematical models.

Mathematical Models

Mathematical models are the most abstract of the four classifications. These models do not look like their real-life counterparts at all. Mathematical models are built using numbers, symbols, variables, empirical laws related together by means of equations or equalities. Mathematical models can take many forms like statistical models, optimization models, algebraic/differential equations or game theoretic models, etc. In this unit and units to follow, we shall concentrate on mathematical models. We shall be discussing in detail the process and need of mathematical modelling, types of mathematical models and we shall formulate some models with the contexts taken from biology, physics, economics, finance, medicine, etc. Before we discuss basic concepts of mathematical modelling in this unit, you may try the following exercise.

E1) Give two examples each of the physical, schematic, verbal and mathematical models.

Let us now discuss the process of mathematical modelling.

1.3 MATHEMATICAL MODELLING – WHAT AND WHY?

Mathematics is a rich and interesting discipline. It provides a set of ideas and tools effective in solving problems which arise in other fields like history, philosophy, sciences, sociology, political science, life sciences, medical sciences, engineering, etc. It also provides concepts useful for theoretical approach in other fields. Mathematics may be applied to specific problems already posed in mathematical form, or it may be used to formulate such problems. In theory construction, mathematics provides abstract structures which may be used as tools in understanding problems arising in other fields. Problem formulation and theory construction involve a process known as **mathematical modelling** or, mathematical model building. A mathematical model, as we have already mentioned, is an abstract model that uses mathematical language to describe a system. It can be viewed as a representation of the essential aspects of an existing system (or a system to be constructed) which presents knowledge of that system in useable form.

A 'system' is a collection of one or more related objects i.e., physical entities with specific characteristics or attributes.

Mathematical modelling usually begins with a situation in the real world. These situations may arise in different disciplines like engineering, physics, physiology, psychology, ecology, wildlife management, chemistry, economics, sports etc. A psychologist, for example, observes certain types of behaviour in rats running in a maze, a wildlife ecologist notes the number of eggs laid by endangered sea turtles, or an economist records the volume of international trade under a specific tariff policy. Each tries to observe and predict future behaviour. Their efforts may be based completely on intuition, but often they are the result of detailed study, experience, and observing the similarities between the current situation and other situations which are better understood. This study of the system, the accumulation and organisation of information and stating the problem to be studied is in fact the **first step** in model building.

The **next step** is to make the problem simpler by making certain assumptions and approximations. It requires to identify and select the concepts/information to be retained in the problem. For example, the psychologist studying rats in a maze may decide that it makes no difference that all the rats are black or that the maze is constructed of wood. On the other hand, it may be significant for her that all the rats are siblings and the maze is divided into compartments. He may also assume that a rat is always in exactly one compartment and never half in one and half in another compartment.

The **third step** in modelling is to replace the real quantities and processes by mathematical symbols, a set of variables and a set of equations/inequalities that establish relationships between these variables. The values of the variables can be practically anything; real or integer numbers, Boolean values or strings, for example. The variables represent some properties of the system, for example, we may measure system outputs often in the form of signals, timing data, event occurrence (yes/no). The actual model is the set of functions that describe the relations between the different variables.

After the problem is formulated, the **fourth step** is the study of the resulting mathematical system using appropriate mathematical tools and techniques. This may involve a calculation, solving an equation, proving a theorem, etc. The motivation is to produce new information about the problem being studied. It is likely that new information can be obtained by using well-known mathematical concepts and techniques. If not, we may need to develop new techniques or adopt tested methods from other disciplines.

The **final step** in the model-building process is the evaluation of a mathematical model i.e., comparison of the results predicted on the basis of the mathematical analysis with the real world. It is important to know whether or not our model gives reasonable answers i.e., Does our model reflect all the important aspects of the real world problem? If a model is not accurate enough, then, we need to refine our model. We may need a new formulation, a new mathematical analysis and hence a new evaluation. It usually happens that the model-building process proceeds through several iterations, each a refinement of the preceding, until, finally, an acceptable one is found. Broadly, we can divide the modelling process as follows:

- **Formulation** which involves the following three steps:

  i) **Studying the problem:** Accumulating and organising the information about the real world problem. Describing the context of the problem and stating the problem within this context.

  ii) **Identifying relevant information:** Identifying and selecting the concepts/information which are significant and to be retained in the problem.

  iii) **Mathematical representation:** Replacing real quantities and processes by mathematical symbols, a set of variables and a set of equations/inequalities that establish relationship between variables.

- **Mathematical Analysis** which studies the formulated problem using appropriate mathematical tools and techniques.

- **Evaluation** which decides whether the model and its analysis explain the phenomenon/problem we are interested in. If it is not a good model then we need to refine it. There is no unique or correct model; but there are good models and bad models. The skill of modelling lies in being able to judge which is which.

Pictorially, we can represent the process of mathematical modelling as shown in Fig. 1 below.

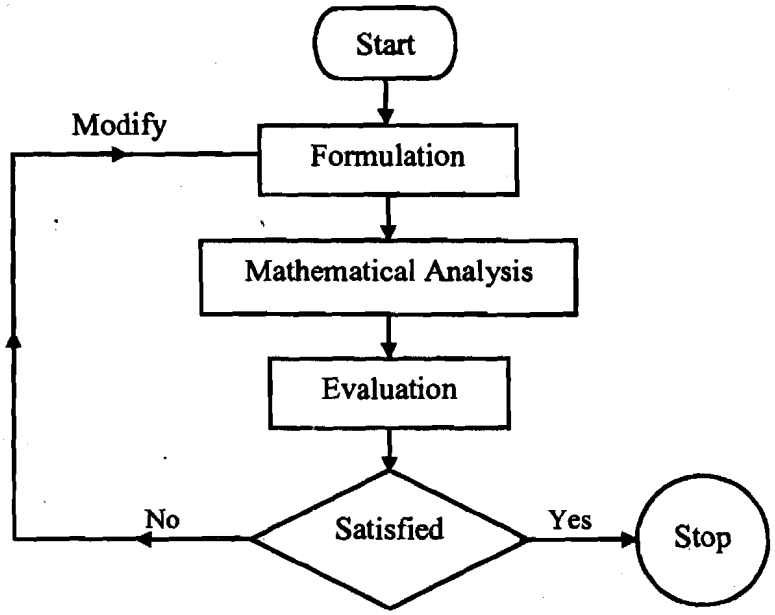


Fig. 1

After a model is evaluated and found good, there are several uses of the model: i) it helps in our better understanding of the real physical system, ii) it can serve as a tool in the prediction of the future state of the system which is currently unknown, iii) it can help in doing trial experiments by changing the parameter values or in perturbing the system to produce a desirable condition, i.e., many unknown parameter values can be estimated.

Let us now look at some simple mathematical models. We shall not go into the details of their formulations here since you must be already familiar with them.

Example 1: *Lotka-Volterra Equations*

The simplest model of interacting populations was proposed independently by Lotka in 1925 in the United States and by Volterra in 1926 in Italy. If x and y are respectively, the predator and prey populations, then their model is in terms of the following equations:

$$\frac{dx}{dt} = \alpha xy - \beta x$$

$$\frac{dy}{dt} = \gamma y - \delta xy,$$

α, β, γ, δ are positive constants.

γ and β represent constant specific birth rate of the prey and mortality rate of the predator respectively.

The predator's specific growth rate αy depends on availability of the prey as food source, whereas, the prey death rate δx depends on the number of predators. These equations have oscillatory solutions but have the unsatisfactory feature of forming a conservative system, and oscillations of any magnitude are possible. More realistic versions of this model remove this degeneracy. This model is discussed in detail in Unit 9, Block-3 of our undergraduate course MTE-14 on 'mathematical modelling'.

***

Example 2: *Model of a Particle in a Box*

In physics, the **particle in a box** (also known as the **infinite potential well** or the **infinite square well**) is a problem consisting of a single particle inside an infinitely deep potential well, from which it cannot escape, and which loses no energy when it collides with the walls of the box. In one dimensional case, the problem can be described as a single point particle enclosed in a box inside which it experiences no force i.e., it is at zero potential energy. At the walls of the box, the potential rises to infinity, forming an impenetrable wall.

In classical mechanics, the problem can be modelled using Newton's laws of motion and the solution to the problem is trivial: The particle moves in a straight line, always at the same speed, until it reflects from a wall. A quantum-mechanical solution of the problem becomes very interesting and reveals some decidedly quantum behaviour of the particle that agrees with observations but contrasts sharply with the predictions of classical mechanics. According to quantum theory, particle has no definite position or velocity. Rather, a probabilistic interpretation is given to the state of the particle in terms of a time-independent wave function $\psi(x)$. The square of the wave function $|\psi|^2$, is a probability density.

The problem of a particle situated in a 1-dimensional infinite square well with momentum only in the direction of quantum confinement (the x-direction) is described by the **schrödinger equation**, a basic equation of quantum mechanics, given by

$$\frac{-h^2}{2m}\frac{d^2\psi(x)}{dx^2} + V(x)\,\psi(x) = E\,\psi(x)$$

where h is the reduced planck constant,
m is the mass of the particle,
$\psi(x)$ is the complex-valued stationary time-independent wave function,
$V(x)$ is the spatially varying potential and
E is the energy, a real number.

For the case of the particle in a 1-dimensional box of length L, the potential is zero inside the box, but rises to infinity at $x = 0$ and $x = L$. Thus, the equation reduces to

$$\frac{-h^2}{2m}\frac{d^2\psi(x)}{dx^2} = E\,\psi(x),\ 0 < x < L.$$

the solution to which can be found easily under appropriate boundary conditions. One of the possible choices is

$$\psi(0) = \psi(L) = 0.$$

The motivation for this choice is that the particle is unlikely to be found at a location with a high potential (the potential repulses the particle), thus the probability of finding the particle, $|\psi(x)|^2$, must be small in these regions and decreases with increasing potential. For the case of an infinite potential, $|\psi(x)|^2$ must be infinitesimally small or 0, thus $\psi(x)$ must be zero in this region. Thus, we get $\psi(0) = \psi(L) = 0$.

***

Example 3: *Model of Rational Behaviour for a Consumer*

In this model, we assume that a consumer faces a choice of n commodities labelled 1, 2, ..., n each with a market price $p_1, p_2, \ldots, p_n$. The consumer is assumed to have a cardinal utility function U (cardinal in the sense that it assigns numerical values to utility), depending on the amounts of commodities $x_1, x_2, \ldots, x_n$ consumed. The model further assumes that the consumer has a budget M which she uses to purchase $x_1, x_2, \ldots, x_n$ in such a way as to maximise $U(x_1, x_2, \ldots, x_n)$. The problem of rational behaviour in this model then becomes an optimization problem, that is:

$$\max U(x_1, x_2, \ldots, x_n)$$

$$\text{subject to: } \sum_{i=1}^{n} p_i x_i \leq M,\ x_i \geq 0\ \forall\ i = (1, 2, \ldots, n).$$

This model is used in general equilibrium theory, particularly to show existence and optimality of economic equilibria. However, the fact that this particular formulation assigns numerical values to levels of satisfaction is the source of criticism. However, it is not an essential ingredient of the theory, only an assumption.

***

And now an exercise for you.

E2) Give at least two examples of the formulas you are already familiar with as the mathematical models of the real situations.

Why formulate a Mathematical Model?

As we mentioned earlier, the use of mathematics is one of many approaches for understanding and solving a real world problem. Others include experimentation with scaled physical models in a laboratory on a smaller scale simulating all the conditions of the real situation, or with the real world directly. But these may be highly risky and costly as they may involve the use of explosive and expensive chemicals or materials. On the other hand, mathematical approach has many advantages. Mathematical modelling is very inexpensive and provides systematic approach to problem solving. Once developed properly, a great deal can be learnt about the real-life situation by manipulating a model's variables and analysing the results. Mathematical modelling requires users to accumulate and organize information and, in the process, to indicate areas where additional information is needed, and hence increase the understanding of the problem.

Let us look at some of the problems which illustrate the use of mathematical modelling.

For meteorological purposes, rockets are launched so that they reach a certain altitude and record atmospheric conditions such as temperature and pressure. In such cases we are confronted with the following problem:

Example 4: What is the rocket thrust level and duration necessary to ensure that the rocket reaches the desired altitude?

For this problem, a solution, based on experimentation involving rocket launches with different thrust-time, is unacceptable due to the cost and the uncertainty of success. Further, the solution to this problem cannot be obtained using scale-model experiments. For this kind of study, a mathematical approach is preferred.

***

In large supermarkets you must have seen a number of checkout counters. Customers prefer to stand in a queue which is the shortest. The problem to be sorted out by the management is the following:

Example 5: What is the optimum number of check out counters that the supermarket should have in order to maximize the expected profit?

Information of this kind is frequently needed for planning purposes. Fewer counters imply long queue lengths and can affect customer goodwill. On the other hand, if there are too many checkout counters, manning them reduces the total profit. There is really no scientific alternative to a mathematical treatment for problems of this kind.

***

You may now think of more situations like this where mathematical treatments of the problem become necessary.

E3) Give two situations from the field of management where mathematical treatment of the problem is necessary to get the required solution.

Mathematical models can be classified into different categories depending on the mathematical structure of the underlying formulation. We shall discuss these classifications in the next section.

1.4 CLASSIFYING MATHEMATICAL MODELS

According to the structure of the models we can classify mathematical models into the following four types:

(i) **Linear vs. Nonlinear**

Mathematical models are usually composed by variables, which are abstractions of quantities of interest in the described systems, and operators that act on these variables, which can be algebraic operators, functions,

differential operators, etc. If all the operators in a mathematical model present linearity, the resulting mathematical model is defined as **linear**. A model is considered to be **nonlinear** otherwise.

For example, consider the equation

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial x}\left(D\frac{\partial C}{\partial x}\right). \qquad (1)$$

Eqn. (1) is a one-dimensional diffusion equation (also known as Fick's II law) where $C(x, t)$ is the concentration of diffusing substance, x is the space coordinate and D is the diffusion coefficient. This equation represents the diffusion of a substance in the x direction due to a concentration gradient in that direction. In some cases, e.g., diffusion in dilute solutions, D can be approximated by a constant, while in other cases, e.g., diffusion in high polymers, D depends on concentration C.

For the case when D is a constant, Eqn. (1) becomes

$$\frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial x^2}. \qquad (2)$$

which is a linear differential equation and hence is said to be a **linear model** of the problem. You may note that Eqn. (1) is a second order linear partial differential equation with C as dependent variable and t and x as independent variables. In general Eqn. (2) along with appropriate initial and boundary conditions is solvable. You must have also come across this equation and solved it in your differential equation course at the undergraduate level (Ref. Unit 17, Block-4, MTE-08).

If, on the other hand, D is not a constant but depends on C, say, for example, if $D = D_0 \exp[\beta (C - C_s)]$ where D_0, β, C_s are constants then Eqn. (1) reduces to

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial x}\left[D_0 \exp[\beta(C - C_s)]\frac{\partial C}{\partial x}\right]. \qquad (3)$$

Eqn. (3) is a non-linear PDE and hence the model obtained is a **non-linear** one.

You may **notice** that if the model tries to replicate more closely the real situation (i.e., D is a function of the concentration in this case), the corresponding mathematics involved is more challenging. (solving a non-linear PDE, in this case).

(ii) **Static vs. Dynamic**

A **static** model does not account for the element of time and hence the variables and relationships describing the system are time-independent. Consider, for instance, the transportation problem generally associated with industries:

Suppose there are m origins O_i, $i = 1, 2, \ldots, m$ in an industry, where various amount of a commodity are produced or stored for transportation to n destinations D_j, $j = 1, 2, \ldots, n$. It is then required to transport all units of the commodity from all O_i, exhaustively, to all D_j, exactly satisfying all their

requirements in such a way that the total transportation cost becomes minimum. The following assumptions are made:

i) The i^{th} origin can supply exactly a_i unit of the commodity.

ii) The j^{th} destination can accept exactly b_j unit.

iii) The cost of transportation of one unit from O_i to D_j is C_{ij}.

This transportation problem, therefore, is essentially a minimization problem. Mathematically, if x_{ij} = number of units transported from O_i to D_j, then we want to

$$\text{minimize } Z = \sum_{i=1}^{m} \sum_{j=1}^{n} C_{ij}\, x_{ij} \tag{4}$$

$$\text{subject to} \quad \sum_{j=1}^{n} x_{ij} = a_i,\ i = 1, 2, \ldots, m \tag{5}$$

$$\sum_{i=1}^{m} x_{ij} = b_j,\ j = 1, 2, \ldots, n \tag{6}$$

$$\text{where } x_{ij} \geq 0\ \forall\ i, j \text{ and } a_i, b_j > 0\ \forall\ i, j. \tag{7}$$

Eqns.(4)-(7) define the transportation problem where all variables are independent of time. Such a system is a **static** system. We shall discuss this model in detail in Unit 5 of this course.

A system is open/closed if the objects in the system do/do not interact with objects of the super-system which do not belong to the system.

In contrast to the static system, in **dynamic** system, time plays a very important role. The relationship between the variables describing the system changes with time. If you look at the problem of rocket launch then it can be described in terms of a closed system consisting of two objects-the rocket and the earth. The variables describing the rocket are its position and velocity relative to some fixed point on the earth, and the interaction between the two objects is given by the theory of dynamics. In this description we may as well include the influence of other planets in the solar system by treating the planets as belonging to the environment of the system and the system being open. In either case, the variables- i.e. position and velocity of the rocket-change continuously with time. Hence the system is a dynamic system.

(iii) Discrete vs. Continuous

Mathematical model may be discrete or continuous according as the variables involved are discrete or continuous. In models involving changes over time, the changes may take place continuously with time and the variables of the system are described for all time instants over the interval of interest. On the other hand, the changes may occur at discrete instants of time and the variables described only for the relevant time instants. Thus, one should be clear whether to treat the time element as continuous or discrete. In certain instances, the data available are such that a discrete treatment of time is more appropriate as opposed to continuous. For instance, if a soft drink manufacturing company is interested in estimating the weekly demand of the soft drink, then the time element is a week and hence it has to be treated as a discrete variable.

Models involving continuous variables are often expressed through differential equations-ordinary or partial, whereas, those involving discrete characterisation result in difference equations. Consider for instance, the mathematical model, referred to as the classical Malthusian population scheme for population growth given by Thomas R. Malthus (1766-1834). The model is based on the idea that the population size for one generation depends on the size of the previous generation. This is expressed mathematically by the following equation

$$p_{t+1} = r \times p_t \tag{8}$$

where:

t : represents the time period (which could be minutes, hours, weeks, years, etc.) depending on the species being considered

p_t : represents the population size at time t. The units of time could be hours, days, years, etc.

p_{t+1} : represents the population size at the next time period. Again it could be the next hour, next day, next year, etc.; and

r : referred to as the Malthusian factor, is the multiple that determines the growth rate.

The model given by Eqn. (8) which is a difference equation, is a **discrete model**. It allows you to find out the value of p at different discrete time intervals say, at years 3, 4, 5, etc. You couldn't use this model to find out the size of p when $t = 3.578$ because $t = 3.578$ does not represent a prescribed discrete time. The value of r in Eqn. (8) has a strong impact on how fast the population will grow. We shall be discussing this model and other discrete and continuous population models in detail in Unit 4. Analogous **continuous** model of the population growth resulting in differential equation is given by the equation

$$\frac{dx(t)}{dt} = r\,x(t),\ x(0) = x_0. \tag{9}$$

where $x(t)$ (> 0) is the size of the population at time t, x_0 is the size of the population at the initial time and r represents the net growth rate. For details of this model refer to Unit 8, Block-3, MTE-14.

(iv) Deterministic vs. Stochastic

A system is said to be deterministic if the values assumed by the variables or the changes in the variables are known with certainty. Consider for instance, the problem of rocket launch considered in Example 4. The variables of the system are the position and velocity of the rocket. The laws of classical dynamics can be used to describe the motion fairly accurately and the changes in position and velocity can be predicted with a high degree of certainty. Hence, in this case we can view the system as being **deterministic**. On the other hand, in a **stochastic** model, randomness is taken into account and variables are described by probability distributions instead of unique values. For example, a manufacturer, having trouble deciding whether to build a large or small facility knows that the solution to this capacity problem depends upon the volume of demand. High demand would require a large facility while low

demand would require a small facility. While the manufacturer has no way of knowing with certainty what the exact demand will be, he can, at least, determine the probability of the occurrence of each (high or low). For example, if the manufacturer estimates that the probability of the occurrence of high demand is 70 percent and the occurrence of low demand is 30 percent, he can use this information along with the monetary value (expected pay-off) of each situation to construct mathematical models such as **pay-off matrices** to find an optimal decision (see Table-1)

Table-1

	High demand (70%)	**Low demand (30%)**
Large facility	Rs.240, 000	–Rs.60,000
Small facility	Rs.120, 000	Rs.100,000

This type of model is said to be a **stochastic optimization model**. We shall be discussing some optimization models in Unit 5. Some models of this type are also discussed in Unit 12, Block-4 of MTE-14.

Once again look at the problem of supermarket operation considered in Example 5. The problem is to determine the optimum number of checkout counters in a supermarket to maximize the expected profit. The model can be successful once a relationship between some measure of queue (e.g. average queue length or average time period for which the customer has to wait before being served) and the number of checkout counters is obtained. The variables characterizing the system are the number of customers, their arrival rate, their departure rate, service time, peak period, etc. Here the arrival of customers, departure of customers and their service time are all random. They cannot be determined uniquely, rather they are given by certain probability distributions. Stochastic models based on fitting these probability distributions to the arrival, departure and service time can be obtained. For the details of some of these models you can refer to Unit 14, Block-4 of Mathematical Modelling course (MTE-14).

In real life, there is always uncertainty. If the uncertainty is insignificant then it can be ignored and the system can be treated as a deterministic system. This is a process of simplification. If the uncertainty is significant then it cannot be ignored and must be taken into account while characterising the system.

As mentioned earlier, most of the discrete and stochastic models lead to difference/algebraic equations whereas the knowledge of algebraic/differential equations is required while dealing with linear/nonlinear, static/dynamic, and continuous models. The skills you have in algebra, calculus, analysis, differential equations, optimization and probability theory will be useful for successfully dealing with mathematical modelling. The type of mathematics required depends on the type of model formulated.

You may now try the following exercises.

E4) State the types of modelling you will use for the following problems. Also give reasons in support of your answer.

a) Human motion (e.g. walking, lifting, jumping) involves dynamic action at one or more joints in the body. If the joints function normally they cause no discomfort otherwise they cause severe pain when in motion. One way of reducing the pain and discomfort is to replace the defective joint with an artificial one. For the artificial joint to be effective it must be capable of executing all the motions of a normal joint. The problem is to **build a model** to describe the functioning of a joint, say, hip joint, so that it is useful to bioengineers when designing hip replacements.

b) Research and development (R&D) is an important activity in any modern industrial organization. The R&D manager is faced with the difficult problem of allocating limited resources (money, manpower, space, etc.) between a number of competing projects. The problem is to **build a model** to help the R&D manager to make right decisions.

c) In a chemical process, the output quality and yield depend critically on the levels assigned to relevant factors (temperature, pressure, concentration, etc.) It is important to optimally select these levels as well as to monitor and control the system to ensure that they stay at the desired levels. The problem is to **build a model** to help achieve this.

d) The formation of sand dunes and their encroachment into de-forested lands near deserts has become a serious problem in many parts of the world. It is needed to predict the spread of desert, as well as to devise policies to control the spread. The problem is to **build a model** to describe the movement of sand dunes.

e) Advertising is a means by which a manufacturer can promote the product and improve sales and hence, the revenue generated. However, advertising costs money and is worthwhile only if the cost of advertising is less than the increase in the sale revenue. The problem is to evaluate the effectiveness of different advertising policies and the selection of the optimal policy. The problem is to **build a model** to help solve this problem.

1.5 LIMITATIONS OF A MATHEMATICAL MODEL

Mathematical modelling is a multistage activity involving various concepts and techniques. Ideally, a mathematical model ends by returning to its origin. We need to check whether the model and its analysis explain the phenomenon we are interested in. The whole art of mathematical modelling lies in its self-consistency. A mathematical model is an adequate mathematical model if it captures the salient features of the system associated with the problem and is capable of yielding a meaningful solution to the original problem.

It is rare that we obtain an adequate mathematical model at the first attempt for the problem under consideration. In general, an iterative procedure is needed

where improvements are progressively made until an adequate mathematical model is obtained. The adequacy of a model is established by checking the validity of the assumptions made in building the model and by the closeness of the agreement between the behaviour of the model and the system under consideration. For example, while developing a model to describe the motion of a simple pendulum if the time interval of study is sufficiently small, so that the energy loss due to frictional drag is very small, then the assumption that the frictional drag is negligible is valid. However, if the time interval of study is large then the assumption of negligible frictional drag is not valid, since the effect of drag on the pendulum motion is cumulative.

In a rocket launch model, the assumption that the thrust generated by the rocket is an impulse, is valid if the thrust lasts for a very small fraction of the total flight time, something that is not known initially. Only after a first solution, can this be checked and if the need be, the model has to be modified. This emphasizes the iterative nature of modelling.

To end the unit we now give the summary of what we have covered in it.

1.6 SUMMARY

In this unit we have covered the following points:

1) Mathematical model uses mathematical language to describe a real world problem.
2) Mathematical models are built using numbers, symbols, variables related together by means of functions, formulas, empirical laws, equations or equalities and can take many forms like statistical models, optimization model, algebraic, differential equations or game theoretic models, etc.
3) The process of mathematical modelling involves three main steps – formulation, mathematical analysis and evaluation.
4) Depending on the mathematical structure of the underlying formulation, mathematical models can be classified into linear/nonlinear, static/dynamic, discrete/continuous and deterministic/stochastic models.
5) The skills in subjects like algebra, calculus, analysis, differential equations, statistics optimization, matrix theory and probability theory are useful for dealing with mathematical modelling. The type of mathematics required depends on the type of model formulated.
6) A mathematical model is adequate if it captures the salient features of the system associated with the problem and is capable of yielding a meaningful solution to the original problem.
7) Mathematical modelling is an iterative process where improvements are progressively made until an adequate mathematical model is obtained.

1.7 SOLUTIONS/ANSWERS

E1) Physical model – models of comic book super-heroes which look exactly like their counterparts but in much smaller scale.

Schematic model – a diagram showing the sequence of activities that must be maintained in an assembly-line balancing.

Verbal model – word problems given in any mathematics book. Similar examples of various types may be given.

Other similar examples from your surroundings or real life experience may be given.

E2) You may give examples of familiar mathematical models from your own experience.

E3) A firm that assembles computers and computer equipment is to start the production of two new types of computers. Each type will require assembly time, inspection time and storage space. The amounts of each of these resources that can be devoted to the production of the computers is limited. The manager of the firm likes to determine the quantity of each computer to be produced in order to maximize the profit generated by their scale. For this kind of study a mathematical approach is preferred.

Examples of other similar situations may be given.

E4) a) Dynamic, continuous, deterministic.
b) Static, discrete, deterministic.
c) Dynamic, continuous, deterministic.
d) Dynamic, discrete, probabilistic.
e) Static, discrete, probabilistic.

— x —

UNIT 2 MODEL FORMULATION

Structure **Page No**

2.1 INTRODUCTION

In Unit 1 we introduced you to the need of modelling a real life situation and the role of mathematics in it. Various steps involved in modelling a real life situation were discussed there. We classified the mathematical models into various types and illustrated them through examples from population dynamics, optimization problem, queueing system, etc. In this unit, we shall proceed with the next step in modelling i.e., given a real world problem, how do you convert it to model abstraction leading to a mathematical equation? We shall draw your attention to the various modelling aspects which need to be taken into account while formulating a mathematical model.

In Sec. 2.2, we shall discuss the method of identifying the essentials of a problem which need to be incorporated into the model. In Sec. 2.3, we shall focus on an example from the field of finance, formulate a model and estimate the financial implications and risks associated with reality. Whenever you decide to make some investment in the market, your concern is to choose an investment strategy which maximizes your profit and minimizes the risk involved. But how to select such a strategy? Solution to this problem as provided by Markowitz will be discussed in Sec. 2.3. Sharpe's model which simplifies the rigours involved in various computations using Markowitz's model will also be discussed here.

Objectives

After studying this unit, you should be able to:

- identify the essentials of a given real world problem;
- formulate a model incorporating all the essentials required for the problem;
- obtain a solution of the formulated problem and interpret the result.

In addition to the above objectives, using the discussed model from the field of finance, you should also be able to

i) explain the meaning and calculation of expected return and risk measures for an individual security;

ii) explain the portfolio selection problem of investments;

iii) explain portfolio return and risk measures as formulated by Markowitz;

iv) explain what the efficient frontier is and how important it is to investment analysis.

2.2 ESSENTIALS OF A PROBLEM

Mathematical modelling involves transforming some idealised form of the real-world situation into mathematical terms. It is, therefore, an activity which requires more than the ability to just solve complex sets of equations. Though there is no hard and fast approach to develop a model, you still need to broadly follow the following steps in the beginning.

i) **Establish a main purpose:** Models rarely replicate a system and also they can mean different things to different people. For example, a business man and a biologist have entirely different points of view of looking at the camel as depicted in Fig. 1.

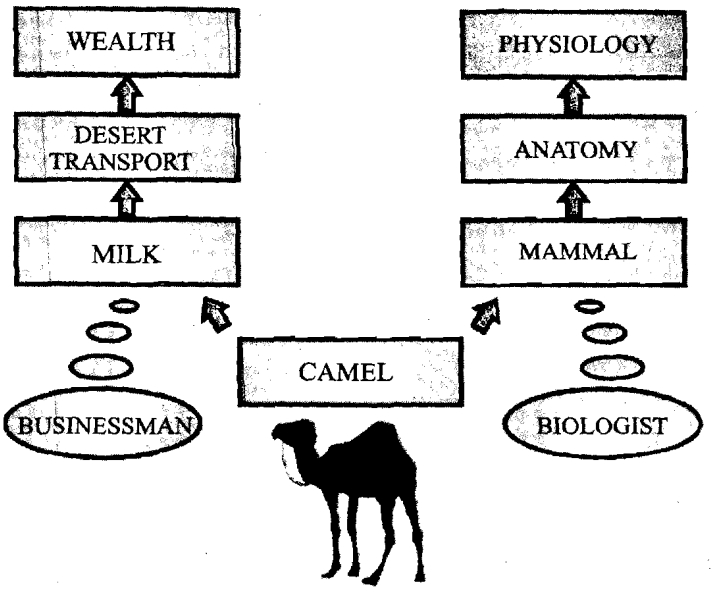


Fig. 1: Different conceptual views.

Their conceptual views of the same system or object are rather different since they are both heavily influenced by their own environment, background and objectives. For example, a biologist would be interested in the anatomy and physiology of a camel whereas, a business man's concern would be the profit he can earn from it. The same is true when we come to the mathematical modelling of any system or process. Although the motivation for building the model is usually to find the means to answering a particular question, the form of that question influences the way in which we build the mathematical model.

In the case of the familiar problem of the motion of a simple pendulum, what is our main purpose? It is to find the period of oscillation, or, angular velocity of the bob, or, we may want to know the tension in the string. Thus, before developing a model we must be clear about the purpose of doing it.

ii) **Sifting essentials from the non-essentials:** Real situations are quite complex. Developing a model which accounts for all the aspects of a real world phenomenon is likely to be difficult, mathematically complex and unmanageable. On the other hand, it may also be possible that a model not accounting for all the aspects of the phenomenon under consideration, is easier to develop and handle and it may still serve our purpose. It is, therefore, necessary to keep a balance between the complexity of the problem and the objectives of our study. For example, while studying the motion of a simple pendulum, if you assume the oscillations to be small and restrict the range of the values of the amplitude θ, to say $|\theta| \leq 30^\circ$, then you only need to study a linear model and hence solve a linear differential equation $\frac{d^2\theta}{dt^2} + \frac{g}{\ell}\theta = 0$, where ℓ is the length of the pendulum and g is acceleration due to gravity. On the other hand, if you have a problem in which you cannot assume the oscillations to be small then you need to study a non-linear model resulting in a non-linear differential equation $\ell\frac{d^2\theta}{dt^2} + g\sin\theta = 0$ (ref Unit 3, Block-1 of MTE-14).

Similarly, a realistic study for understanding the population dynamics of an infectious disease is one which differentiates the population by age, gender, geographical location, social stratification and takes into consideration factors like immigration, emigration, birth, death, incubation period of the infection and removal of infectives from circulation. This study will be definitely superior but more complex. The model developed would involve more dependent variables and hence more number of differential equations to be solved as compared to the model where, the population considered is assumed to be closed, homogeneously mixed, with no incubation period of infection and no removal of infectives from circulation.

The search for essentials of the problem is related to the main purpose of the model. For instance, the assumption of no removal of infectives in general, is highly restrictive in the context of a human population. However, this might be a reasonable approximation to the early stages of some upper respiratory infections where a long time may elapse before an infective is removed from circulation. A mild cold infection in a classroom may also be considered as an example of no removal in a human population. On the other hand, the assumption of no removal is fairly applicable to epidemics in insect and plant populations. In the case of a common fungal disease of flies, the diseased or dead fly remains attached to a leaf or a blade of grass, creating a situation roughly analogous to no removal. The dead or diseased plants in a forest are rarely removed, and if dead plants are infectious, the assumption of no removal is fulfilled. If you are interested in knowing more about modelling the spread of infectious diseases, you may refer to Unit 10, Block-3 of MTE-14.

iii) **Significance of essentials:** Once the essentials of a problem are identified, they determine the number and type of the variables in the formulation and hence the complexities involved in problem solving. The variables involved may be discrete, continuous or piece-wise

continuous, deterministic, stochastic, fuzzy or any combination of these. For example, biological population of a given species of large sized animals is an example of deterministic and discrete quantity. Flow of money in a market through organized financial institutions is deterministic and continuous. The sale of long sized items like trucks, car, aeroplanes per year is discrete and stochastic. Flow of water in a canal may be piecewise continuous when the up stream gates are opened or closed according to rainfall in the catchment area. The response level in a human neural network is a fuzzy quantity.

As we have already mentioned in Unit 1, the types of variables involved in a formulation determine the type of model formulated and hence the type of mathematics required to solve the problem. We shall be illustrating this point in the next section through an example. But, before that, you may try the following exercises.

E1) State the type of modelling you will use for the following problems giving reasons for your answers. List the essentials and non-essentials in the problems.

i) The economic viability of an insurance company depends critically on its ability to assess risks and decide on the premium charged to cover risks. If the premium are low, then payouts can exceed revenue collected and the company can go bankrupt. On the other hand, if they are high, the number of customers will go down, thus affecting profitability. The problem is to develop a model to help the insurance company decided the premium it should charge for different risks to ensure economic viability and maximise its profits.

ii) A company manufacturing soft drinks is thinking of expanding its plant capacity so as to meet future demand. The monthly sales for the past 5 years are available. The problem is to develop a model to obtain good estimates for future demand so as to help the company make the right decisions.

E2) Give examples of at least two real life situations where mathematical treatment of the problem is the only approach to find the solution of the problem. Why do you think that there is no other scientific alternative for the treatment of these problems? List the essentials and non-essentials in these problems.

E3) Give at least two examples each of physical situations in which the variables involved are i) continuous and stochastic; ii) piece-wise continuous and stochastic; iii) fuzzy.

Let us now illustrate the mathematical formulation of real world situations from the field of finance.

2.3 MODELS FROM FINANCE

All of us make some or the other kind of investments to secure our future, old age, for our children's education, their marriage etc. People invest in shares, stocks, bonds, mutual funds, etc. using an investment strategy (called **security**)

related to the expected market scenario in the future. Our aim is to select a security or a group of securities (i.e., a **portfolio**) which maximize the expected return and minimize the uncertainty (i.e., **risk**). But then the problem is how to choose such a security or a portfolio? Markowitz's model, named after Harry Max Markowitz, an economist, who is best know for his pioneering work in Mordern Portfolio theory, provides a solution to this problem. He studied the effects of asset risk, correlation and diversification on expected investment portfolio return. We shall discuss this model here, but before that let us define some of the terms which are the **essentials** for the model under consideration.

Securities

When you borrow some money or take a loan from a broker then you have to leave some item of values as security with the broker or sign a piece of paper promising repayment with interest. Failure to repay the loan (plus interest) means that the broker can sell your item to recover the amount of the loan (plus interest) and perhaps make a profit. The terms of agreement are recorded when the deal is made. This piece of paper serving as evidence is called a **security**. Similarly, you may have some spare money which you would like to lend to earn some profit out of it. You then think of investing your money in Government Bonds, saving certificates, shares, mutual funds, etc. These investment strategies are called the securities. Broadly speaking, a security helps us to save our funds in the event of default (and that is why the name). It may be a simple promissory note, share of the common stock, a bond, etc.

Return

Once the investment has been made, you are interested in knowing how good was your investment strategy. For this, you need to calculate the rate of return on the investment strategy. What is the rate of return? It gives a relation between the initial input and final output of an investment and is calculated as follow.

$$\text{Return} = \frac{\text{End of period value} - \text{Beginning of period value}}{\text{Beginning of period value}}. \qquad (1)$$

Risk

Risk denotes the probability of specific eventualities which may have both beneficial and adverse consequences. However, in general usage the convention is to focus only on potential negative impact of the investment strategy. Often, it is described as a situation which would lead to negative consequences.

Uncertainty

It is the lack of certainty. You may have limited knowledge and it is impossible for you to exactly describe the existing state or future outcome of an investment strategy. For example, there is always an uncertainty while investing in the stock market or mutual funds.

Portfolio

Suppose you want to invest an amount W_0 in n securities (say). Let w_i be the proportion of W_0 that you invest in the i^{th} security $(1 \le i \le n)$. Then the ordered n-tuple $\mathbf{P} = (w_1, w_2, \ldots, w_n)$, where $\sum_{i=1}^{n} w_i = 1$, is called a **portfolio of n securities** and w_i is its i^{th} **portfolio value** (or **weight**).

For example, $\mathbf{P} = \left(0, \frac{1}{3}, \frac{2}{3}\right)$ is a portfolio of three securities where no amount is invested in the first security, $\frac{1}{3}$rd of the total funds are invested in the second and $\frac{2}{3}$rd in the third security. By changing the value of w_i in $\mathbf{P}$, subjected to the condition that $\sum_{i=1}^{n} w_i = 1$, all the portfolios of given n securities i.e., a feasible set can be obtained. Formally, we give the following definition.

Definition: The set of all the possible portfolios which can be constructed from a given set of securities is called the **feasible set** or the **opportunity set.**

Given a portfolio $\mathbf{P} = (w_1, w_2, \ldots, w_n)$ of n securities, our main purpose is to see the effect of portfolio values w_i on the terminal value of the return on the portfolio $\mathbf{P}$. But each w_i is a certain proportion of the initial funds that are invested in i^{th} security of the portfolio $\mathbf{P}$. Thus, for quantifying the return and risk of the portfolio $\mathbf{P}$, we have to calculate the return and risk of its constituting n securities. It is therefore, important to select the proportions w_i of the initial funds in such a way that the portfolio $\mathbf{P} = (w_1, w_2, \ldots, w_n)$ is **optimally** good according to our investment objectives. Such a portfolio which provides an investor the maximum level of satisfaction is called an **Optimal Portfolio** and the problem of selecting such a portfolio is referred to as portfolio selection problem. Thus, as a **first step** towards modelling we state the **formulated** problem as follows:

Portfolio Selection Problem: From a set of feasible portfolios of risky securities, how can an investor choose a portfolio that gives him/her the maximum level of investment satisfaction (in terms of return and risk)?

Before we take up the problem you may try the following exercises.

E4) At the end of year 2005, Mohan decided to invest Rs.30,000 in a portfolio of stocks and bonds. Rs.10,000 were put into common stocks and Rs.20,000 into corporate bonds. At the end of 2006, Mohan's stock and bond holdings were worth Rs.13,000 and Rs.16,000, respectively. During 2006, Rs.500 in cash dividends was received on the stocks and Rs.1000 interest payments was received on the bonds. What was the percentage return on

i) Mohan's stock portfolio during 2006?

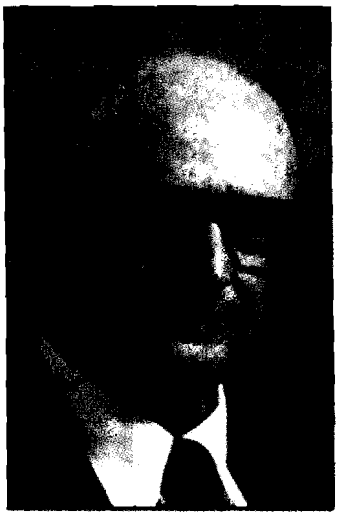
ii) Mohan's bond portfolio during 2006?

iii) Mohan's total portfolio during 2006?

E5) Given a set of eight securities with portfolio values $w_i's = 0, \frac{1}{3}, \frac{1}{4}, \frac{1}{2}, \frac{2}{3}, \frac{1}{2}, \frac{3}{4}, \frac{1}{4}$. Find a feasible set of portfolio of these securities.

Let us now discuss the Markowitz's approach to solving the portfolio selection problem.

2.3.1 Markowitz Model



Markowitz
(1927 – till date)

Markowitz's approach to portfolio selection begins by assuming that an investor has a given sum of money to invest at the present time. This money will be invested for a particular length of time, known as the investor's **holding period**. Markowitz's approach is a single period approach where at the beginning of the period say $t = 0$, the investor must make decision on what particular securities to purchase and hold until the end of the period say $t = 1$. Since a portfolio is a collection of securities, this decision is equivalent to selecting an optimal portfolio from a set of possible portfolios. Typically it is seen that an investor wants the returns to be high as well as, as certain as possible. This means that the investor, looking forward to maximize expected return and minimize risk has two conflicting objectives that must be balanced against each other when making the purchase decision at $t = 0$. The Markowitz approach for how the investor should go about making this decision gives full consideration to both these objectives. He used certain concepts from probability theory to solve this problem. According to Markowitz, the investor should view the rate of return associated with any of these portfolios to be what is known in statistics as a random variable. Expected rate of **return** can be viewed as a measure of the potential reward associated with any portfolio and standard deviations can be viewed as a measure of **risk** associated with any portfolio. Thus, once each portfolio has been examined in terms of its potential rewards and risks, the investor is in a position to identify the portfolio that appears most desirable to him or her.

Thus, as a **second step** in the model building process, we can say that the Markowitz model is based on the following assumptions:

- Investor invest money for a particular length of time, called the holding period.
- Investors are rational and behave in a manner as to maximize their utility with a given level of income or money.
- Investors are risk averse and try to minimize the risk and maximize return.
- Investors prefer higher returns to lower returns for a given level of risk.
- The evaluation of portfolios is carried out in terms of returns and the risk associated with the constituting securities, over a given holding period.

The obvious question which must be occurring in your mind is – How to quantify the expected return and risk of a security or a portfolio? Let us try to do that as a **next step** in the modelling process.

Expected Return for a Security

Some investments, like bank certificates of deposit, have guaranteed rates of return. Investments like stocks are a bit more complicated. If you ask an investor how much she expects to earn on the stock of a pharmaceutical company or high-tech firm, she will answer in terms of a range like "I expect to earn somewhere between 10 and 20 percent on these shares". Whenever an investor describes future returns in terms of a range like this, you can be sure that there is risk involved. In the investment context, risk is the uncertainty that a given investment will earn its anticipated rate of return.

To calculate the expected rate of return, you have to first enumerate all the possible rates of return that an investment could have. For simplicity's sake, let's imagine an investment with four possible rates of return, $-10\%, -5\%, 10\%$ and 20%.

The first two rates of return indicate a loss while the last two indicate a gain. The next step is to assign probabilities to each rate of return. How do you assign these probabilities? It requires to make some educated guesses based on past performance of the investment itself, and the demonstrated performance of similar investments. General market and economic factors should also be taken into account, while assigning these probabilities. Let the probabilities $0.1, 0.1, 0.5$ and 0.3 are assigned respectively to the above four rates of return.

Then the **expected rate of return** or simply **expected return E(R)** of the given investment (security) is calculated as follows:

$$E(R) = \sum_{i=1}^{n} (\text{Possible Return} \times \text{Probability}) \qquad (2)$$

$$= [(-0.10)\,(0.10) + (-0.05)\,(0.10) + (0.10)\,(0.50) + (0.20)\,(0.30)]$$

$$= 0.095.$$

The expected return is 0.095. This means that the investor can expect a return of 9.5% on her investment.

Thus, for any security, the **expected return** is the **weighted average all possible outcomes**, where each outcome is weighted by its respective probability of occurrence. It is calculated as

$$E(R) = \sum_{j=1}^{n} R_j\, p_{ij} \qquad (3)$$

where

$E(R)$ = the expected return on a security,

R_j = the j^{th} possible return,

p_{ij} = the probability of the i^{th} return R_j, and

n = the number of possible returns or outcomes.

You may now use the above result to do the following exercise.

E6) Let the return distribution on a security A be given as follows:

Possible rate of return (in percent) (R_j)	−8	−6	9	10	12	8
Associated probabilities (p_{ij})	0.04	0.06	0.2	0.3	0.25	0.15

Find the expected return of the security A.

Risk for a Security

A security with more standard deviation is more risky than a security with less standard deviation.

In Markowitz sense, the variability or risk of a security is measured by its standard deviation. To calculate the risk associated with the expected return, the variance or standard deviation is used. As you already know standard deviation is a measure of the spread or dispersion in the probability distribution, that is, it measures the dispersion of a random variable around its mean. The larger this dispersion, the larger the variance or standard deviation and hence larger is the risk for a security. Knowing the returns associated with the securities the **variance of the returns** is calculated as

$$\sigma^2 = \sum_{j=1}^{n} [R_j - E(R)]^2 \, p_{ij} \qquad (4)$$

where

σ^2 = the variance of returns,

E(R) = the expected return on a security,

R_j = the j^{th} possible return,

p_{ij} = the probability of the i^{th} return R_j, and

n = the number of possible returns.

$\sigma = (\sigma^2)^{1/2}$, the positive square root of the variance is the standard deviation of the returns.

Let us consider the following example.

Example 1: Assume that the return distribution of a security is as given follows:

Possible Return	0.01	0.07	0.08	0.1	0.15
Associated Probability	0.2	0.2	0.3	0.1	0.2

Find the standard deviation of the security.

Solution: The calculations for the standard deviation are as shown in Table 1.

Table 1

Possible Return (R_j)	Associated Probability p_{ij}	$R_j \times p_{ij}$	$[R_j - E(R)]^2$	$[R_j - E(R)]^2 p_{ij}$
0.01	0.2	0.002	0.0049	0.00098
0.07	0.2	0.014	0.0001	0.00002
0.08	0.3	0.024	0.0	0.0
0.10	0.1	0.010	0.0004	0.00004
0.15	0.2	0.030	0.0049	0.00098
	$\sum p_{ij} = 1$	$\sum_j R_j p_{ij}$ = 0.08 = E(R)		$\sum_j [R_j - E(R)]^2 p_{ij}$ = 0.00202 = σ^2

$\sigma = (0.00202)^{1/2} = 0.0449 = 4.49\%$.

***

You may now try the following exercises.

E7) Calculate the expected return and risk of a security given the following information

Probabilities	0.15	0.20	0.40	0.10	0.15
Possible returns	0.20	0.16	0.12	0.05	– 0.05

For the sake of convenience, we assume that all **securities in a portfolio have equal number of return outcomes.** Also, it is assumed that in a particular situation, $p_{ij} = p_{kj}$, $\forall$ i and k.

E8) Compare the risk of the two securities 1 and 2 whose return distributions are as given in Table 2 below:

Table 2

Possible rate of return for security		Associated probabilities
1	2	$p_{1j} = p_{2j}$
0.19	0.09	0.13
0.17	0.16	0.15
0.11	0.18	0.42
0.10	0.11	0.30

When analysing the returns and risks on an investment made by an investor we must be concerned with the total portfolio held by an investor. Individual security returns and risks are important, but it is the return and risk to the investor's total portfolio that ultimately matters, because investment opportunities can be enhanced by packaging them together to form portfolios.

Portfolio Expected Return

The expected return on any portfolio is easily calculated as a weighted average of the individual securities' expected returns. For a portfolio $P = (w_1, w_2, \ldots, w_n)$ of n securities, where w_i's are the portfolio values (or weights), the expected return can be calculated as

$$E(R_p) = \sum_{j=1}^{n} w_j E(R_j) \qquad (5)$$

where

$E(R_p)$ = the expected return on the portfolio **P**,

w_j = the portfolio weight for the j^{th} security,

$\sum w_j = 1$,

$E(R_j)$ = the expected return on the j^{th} security, and

n = the number of different securities in a portfolio.

Consider for example a portfolio consisting of three stocks X, Y and Z with expected returns of 12%, 20% and 17% respectively. Assume that 50% of the funds is invested in security X, 30% in Y and 20% in Z. Then the expected return on this portfolio is

$$E(R_p) = 0.5 \times 12\% + 0.3 \times 20\% + 0.2 \times 17\% = 15.4\%.$$

It may be **noted** that regardless of the number of assets (securities) held in a portfolio, or the proportion of total investable funds placed in each asset, the expected return on the portfolio is always a weighted average of the expected returns of individual asset in the portfolio.

Portfolio Risk

Portfolio risk is measured by the variance (or standard deviation) of the portfolio's return, exactly as in the case of each individual security. Although the expected return of a portfolio is a weighted average of its expected returns, **portfolio risk is not a weighted average of the risk of the individual securities in a portfolio**. Mathematically,

$$E(R_p) = \sum_{j=1}^{n} w_j \, E(R_j).$$

But, $$\sigma_p^2 \neq \sum w_j \, \sigma_j^2 . \qquad (6)$$

Portfolio risk is a unique characteristic and not simply the sum of individual security risk. This is because a security may have a large risk if it is held by itself but risk may be small when held in a portfolio of securities. But then the question is how a portfolio of assets can reduce risk and how the risk is measured? Let us analyze the portfolio risk.

Analyzing Risk of a Portfolio

Let us first assume that rates of return on individual securities are statistically independent such that rate of return on any one security is unaffected by the rate of return on another security. In this situation, the standard deviation of the portfolio is given by

$$\sigma_p = \frac{\sigma_j}{n^{1/2}} . \qquad (7)$$

Thus, σ_p decreases as n increases i.e., the risk of the portfolio will be reduced as more securities are added to the portfolio. You do not have to take decision about which security to add, as all of them have identical properties. The only concern is how many securities are to be added? However, in the real world, the assumption of statistically independent returns on stocks is unrealistic.

Most stocks are positively correlated with each other, that is, the movements in their returns are related. For example, a rise in interest rates will adversely affect most of the firms, because most of the firms borrow funds to finance part of their operations. Therefore, there is a need for diversification i.e., putting small fractions of the total funds in as many securities as are found suitable after their evaluation is done.

Diversification

Diversification is the key to the management of portfolio risk. It allows the investors to reduce portfolio risk without affecting returns. Thus, there is a need to quantify diversification. Markowitz was the first one to quantify the concept of diversification. He showed quantitatively why and how portfolio diversification helps to reduce the risk of an investor's portfolio. Markowitz showed that to measure and reduce portfolio risk to its minimum level for any given level of return, we must measure the interrelationships among security returns. The reason being that in any one time period, poor performance by some securities may be offset by strong performance in other securities. Let us now consider how to measure these interrelationships, or co-movements or covariance, among security returns.

Covariance in Security Returns

Covariance is an **absolute measure** of the co-movements between security returns used for calculating portfolio risk. Therefore, in order to calculate portfolio variance or standard deviation we need to calculate covariance between securities in a portfolio. The formula for calculating covariance between two securities A and B is given as follows:

$$\sigma_{AB} = \sum_{j=1}^{n} [R_{Aj} - E(R_A)]\,[R_{Bj} - E(R_B)]\,p_{ij} \qquad (8)$$

where

σ_{AB} = the covariance between securities A and B,

R_{Aj} = j^{th} possible return on security A,

$E(R_A)$ = the expected value of the return on security A, and

n = the number of likely outcomes for a security for the period.

Eqn. (8) gives the covariance as the expected value of the product of deviations from the mean. Covariance can be positive, negative or zero.

- Positive covariance indicates that the returns on the two securities tend to move in the same direction at the same time. When one increases (decreases), the other also do the same.
- Negative covariance indicates that the returns on the two securities move inversely. When one increases (decreases), the other tends to decrease (increase).
- Zero covariance indicates that the returns on two securities are independent and do not move together in the same or opposite directions.

The covariance σ_{AB} when divided by the standard deviations of securities A and B gives

$$\rho_{AB} = \frac{\sigma_{AB}}{\sigma_A \sigma_B} \qquad (9)$$

where ρ_{AB} is called the **correlation coefficient** of securities A and B. From Eqn. (9) we can write

$$\sigma_{AB} = \rho_{AB} \sigma_A \sigma_B . \qquad (10)$$

The correlation coefficient ρ_{AB} is a statistical measure of relative comovements between security returns. It measures the extent to which the returns on any two securities are related and is bounded by $+1.0$ and -1.0.

When $\rho_{AB} = +1.0$, the returns are **perfectly positively correlated**, the risk of a portfolio is simply a weighted average of the individual risks of the securities.

When $\rho_{AB} = -1.0$, the returns are **perfectly negatively correlated**. The returns of the securities have a perfect inverse linear relationship to each other. When the returns of one security is high that of other is low.

With zero correlation, there is **no linear relationship between the returns** on the two securities. The risk of the portfolio reduces when two securities with zero correlation are combined with each other.

Let us now consider an example to understand how ρ_{AB} and σ_{AB} is calculated between two securities A and B.

Example 2: Assume that the return distribution on the two securities X and Y be as given in Table 3 below.

Table 3

Possible rates of returns for security		Associate probabilities
X	Y	$p_{Xj} = p_{Yj}$
.19	.18	0.33
.17	.16	0.25
.11	.11	0.22
.10	.9	0.20

p_{Xj}'s are the associated probabilities of the security X and p_{Yj}'s are the associated probabilities of the security Y.

Find σ_{XY} and ρ_{XY}.

Solution: Calculations for σ_{XY} are shown in Table 4 below:

Table 4

Possible Returns		Associated Probabilities	Expected Returns			
R_{xj}	R_{yj}	$p_{xj} = p_{yj} = p_j$	$R_{xj} \times p_j$	$R_{yj} \times p_j$	$[R_{xj} - E(X)]$	$[R_{yj} - E(Y)]$
0.19	0.18	0.33	0.0627	0.0594	0.0406	0.0384
0.17	0.16	0.25	0.0425	0.0400	0.0206	0.0184
0.11	0.11	0.22	0.0242	0.0242	–0.0394	–0.0316
0.10	0.09	0.20	0.0200	0.0180	–0.0494	–0.0516
			$\sum R_{xj}p_j = 0.1494 = E(X)$	$\sum R_{Yj}p_j = 0.1416 = E(Y)$		

You can check that

$$\sigma_X = \{0.33\times(0.0406)^2 + 0.25\times(0.0206)^2 + 0.22\times(-0.0394)^2 + 0.2(-0.0494)^2\}^{1/2}$$
$$= (0.0014)^{1/2} = 0.038$$

$$\sigma_Y = \{0.33\times(0.0384)^2 + 0.25\times(0.0184)^2 + 0.22\times(-0.0316)^2 + 0.2(-0.0516)^2\}^{1/2}$$
$$= (0.0013)^{1/2} = 0.036$$

then σ_{XY} can be calculated using Formula (1) as

$$\sigma_{XY} = 0.33\times0.0406\times0.0384 + 0.25\times0.0206\times0.0184$$
$$+ 0.22\times(-0.0394)\times(-0.0316) + 0.20\times(-0.0536)\times(-0.0516)$$
$$= 0.0013.$$

Hence, by Formula (2), we have

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} = \frac{0.0013}{0.038\times0.036} = 0.95 < 1.$$

***

You may now try the following exercise.

E9) For the data given in E8), find the covariance σ_{12} and correlation coefficient ρ_{12} between two securities 1 and 2.

Knowing the correlation and covariance that gives the comovement in security returns, we are now ready to calculate portfolio risk. We start with the case of two securities and then generalise it to n securities.

If $\mathbf{P} = (w_1, w_2)$ be a portfolio of two securities 1 and 2 with $w_1 + w_2 = 1$, then the **risk of the portfolio**, as measured by the standard deviations of returns is given by

$$\sigma_P = [w_1^2\ \sigma_1^2 + w_2^2\sigma_2^2 + 2w_1w_2\rho_{12}\sigma_1\sigma_2]^{1/2} \qquad (11)$$

where

σ_P = the standard deviation of the portfolio **P**,

σ_1 = the standard deviation of security 1,

σ_2 = the standard deviation of security 2,

w_1 = the portfolio weight of security 1, and

w_2 = the portfolio weight of security 2.

ρ_{12} = the correlation coefficient between security 1 and 2.

Portfolio risk is effected both by the correlation between assets and by the percentage of funds invested in each asset. We shall now illustrate it through an example.

Example 3: Consider a portfolio **P** of two securities X and Y whose return distributions are as given in Example 2. Let 50% of investable funds is to be placed in each security. Find the risk for the portfolio **P**.

Solution: We have the following data from Example 2.

$$\sigma_X = 0.0038,\ \sigma_Y = 0.036,\ \rho_{XY} = 0.95.$$

Also we have $w_1 = 0.5,\ w_2 = 0.5$.

Then the risk σ_P of the portfolio **P** is obtained as

$$\begin{aligned}\sigma_P &= [(0.5)^2(0.038)^2 + (0.5)^2(0.036)^2 + 2(0.5)(0.5)(0.038)(0.036)(0.95)]^{1/2} \\ &= [0.0007 + 0.0006 + 0.0006 \times 0.95]^{1/2} \\ &= 0.0436\end{aligned}$$

You may note that the risk of the portfolio depends heavily on the value of the correlation coefficient between the returns of the two securities as is evident from the following values

$\rho = +1.0;\ \sigma_P = .0447$

$\rho = +0.5;\ \sigma_P = 0.04$

$\rho = +0.15;\ \sigma_P = 0.0373$

$\rho = 0.0;\ \sigma_P = 0.036$

$\rho = -0.5;\ \sigma_P = 0.0316$

$\rho = -1.0;\ \sigma_P = 0.0265$

The risk of the portfolio steadily decreases from 0.0447 to 0.0265 as the correlation coefficient declines from $+1.0$ to -1.0. In the same way it can be seen by holding the correlation coefficient constant that the size of the portfolio weights assigned to each security has an effect on portfolio risk. You can check it yourself in the above example by assigning different values to w_1 and w_2 instead of assuming $w_1 = w_2 = 0.5$.

***

The two-security case for calculating portfolio risk can be generalised to the n-security case. Portfolio risk in the case of n securities can be calculated as follows:

$$\sigma_P^2 = \sum_{j=1}^{n} w_i^2 \sigma_i^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \sigma_{ij} \qquad (12)$$

where

σ_P^2 = the variance of the return on the portfolio,

σ_i^2 = the variance of return for security i,

σ_{ij} = the covariance between the returns for securities i and j, and

w_i = the portfolio weights or percentage of investable funds invested in security i.

You may now try the following exercises.

E10) Four securities have the following expected returns:
A = 15%, B = 12%, C = 30% and D = 22%
Calculate the expected returns for a portfolio consisting of all four securities under the following conditions.

a) The portfolio weights are 25% each.

b) The portfolio weights are 10% in A and the remaining is equally divided among the other three securities.

c) The portfolio weights are 10% each in A and B and 40% each in C and D.

E11) Let the returns on three securities 1, 2, 3 be 25%, 28% and 14% respectively with $\sigma_1 = 4$, $\sigma_2 = 5$, $\sigma_3 = 6$, $\sigma_{12} = \sigma_{13} = 15$ and $\sigma_{12} = -10$. Find the standard deviation σ_p of the portfolio $\mathbf{P} = (0.5, 0.3, 0.2)$.

You have learnt to evaluate portfolios on the basis of their expected returns and risk as measured by the standard deviation. The evaluation of the risk of a portfolio involves the evaluation of the following three parameters:

i) Standard deviation of each security.

ii) Covariance between pair of securities

iii) Proportion of funds invested in each security.

Once the risk-return opportunities available to an investor is determined it is seen that a large number of possible portfolios exist when the percentage of the investors wealth to be invested in each security is varied.

Doest the investor need to evaluate all these portfolios? The answer to this question is no. The reason being that an investor needs to look at only a subset of the available portfolios meeting the following two conditions.

1. Portfolios that offer maximum expected returns for varying level of risk.
2. Portfolios that offer maximum risk for varying levels of expected returns.

The set of portfolios meeting these two conditions is known as the **efficient set** or **efficient frontier**. As a **next step** in the modelling process, we now try to get new information about the problem being studied. From a large number of possible portfolios, we try to locate the efficient frontier. The efficient set can be located from the feasible set, also know as the opportunity set. We shall now see how this is done.

Efficient Frontier

Let us consider Fig. 2 illustrating the location of the feasible set.

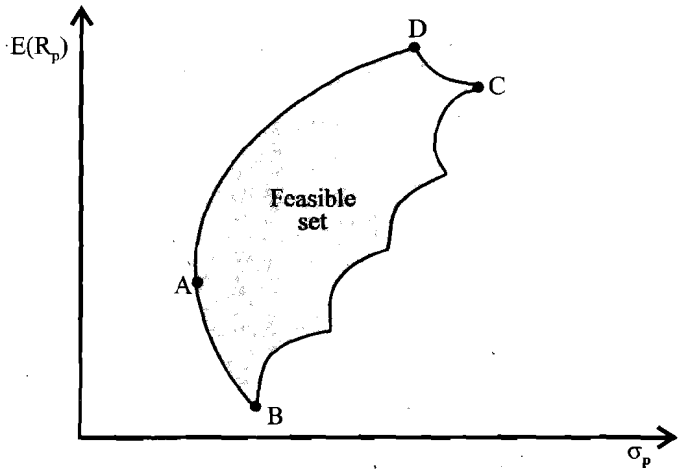


Fig. 2: The feasible set of portfolios.

You know that the feasible set represents the set of all portfolios that could be formed from a group of n-securities. That is, all possible portfolios that could be formed from the n-securities lie either on or within the boundary of the feasible set. The points denoted by A, B, C and D in Fig. 2 are examples of such portfolios. In general this set will have an umbrella type shape similar to the one shown in Fig. 2. The efficient set can then be located by applying the two conditions of the efficient set. Let us first, identify the set of portfolios meeting the first condition.

If you look at Fig. 2, there is no portfolio offering less risk than that of portfolio **A**. This is because if a vertical line is drawn through **A**, there will no point in the feasible set to the left of the line. Also there is no portfolio offering more risk than that of portfolio **C**. Thus, the set of portfolios offering maximum expected return for varying levels of risk is the set of portfolios lying on the 'northern' boundary of the feasible set between points **A** and **C**.

Let us now consider the second condition. The portfolio offering maximum expected return is **D**. If you draw a horizontal line through **D** then there is no point in the feasible set that lies above this line. Similarly, there is no portfolio offering a lower expected return then portfolio **B**. Thus, the set of portfolios offering minimum risk for varying levels of expected return is the set of portfolios lying on the 'western' boundary of the feasible set between points **B** and **D**. Now since both the conditions are to be fulfilled while identifying the efficient set, only the portfolios lying on the northwest boundary between points **A** and **D** need to be considered. Accordingly, these portfolios form the **efficient set**, and the investor will have to find his or her optimal portfolio from this set of efficient portfolios. All other portfolios are inefficient and can be safely ignored.

As a **final step** in the portfolio selection problem, let us now see how an optimal portfolio can be obtained from the set of efficient portfolios.

Selecting an Optimal Portfolio

To select an optimal portfolio of assets using the Markowitz analysis, investor should

1. Identify optimal risk-return combinations available from the set of risky assets being considered by using the Markowitz efficient frontier analysis.
2. Choose the final portfolio from among those in the efficient set based on an investor's preferences.

To select the expected return-risk combination that satisfy an individual investor's personal preferences, **indifference curves** are used. Investors are assume to know their indifference curve. These curves representing an investor's preferences for risk and return can be drawn on a two-dimensional figure, where the horizontal axis indicates risk as measured by standard deviation (σ_p) and vertical axis indicates reward as measured by expected return ($E(R_p)$).

Indifference curves for a hypothetical investor are shown in Fig. 3.

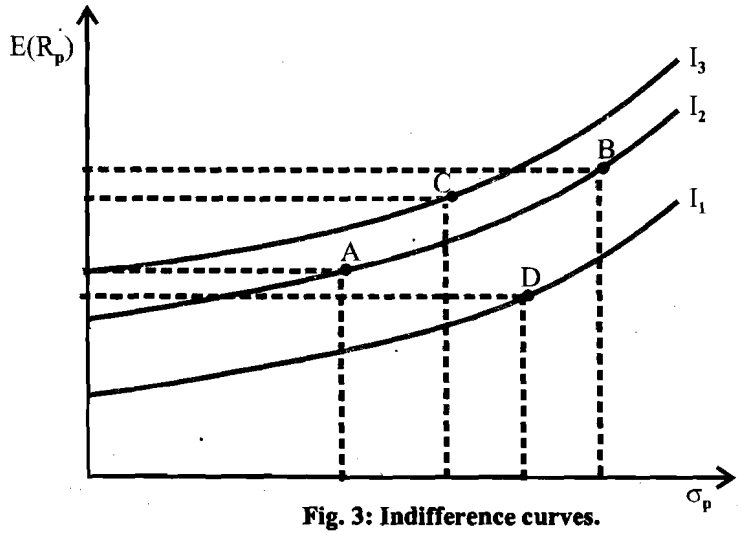


Fig. 3: Indifference curves.

Each cruved line in Fig. 3 indicates one indifference curve for the investor and represents all combinations of portfolios that the investor would find equally desirable. For example, the investor with the indifference curves as shown in Fig. 3, would find portfolios A and B equally desirable, since they lie on the same indifference curve I_2, even though they have different expected returns and standard deviations. Portfolio B has higher standard deviation then portfolio A so it is less desirable on that dimension but on the other hand, it is preferred as it provides higher expected return. Thus, all portfolios that lie on a given indifference curves are equally desirable to the investor. Also, indifference curves cannot intersect, since they represent different levels of desirability.

As an investor, we would always love to have portfolios with more return and less risk, our indifference curves will always show some inclination towards the return line. That is, we choose portfolios in such a way that the corresponding curve heads towards the northwest direction. In other words, farther an indifference curve is from the horizontal axis, the greater is the utility. Let us now see how do we use indifference curves in conjunction with the efficient frontier to find which feasible portfolio is optimal.

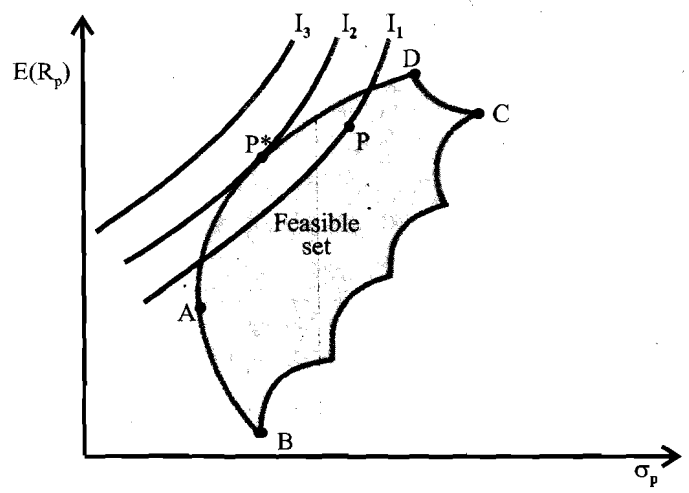


Fig. 4: Selecting an optimal portfolio on the efficient frontier.

To start with, an investor should plot his or her indifference curves on the same figure as the efficient set and then choose the portfolio that is on the indifference curve and is furthest northwest. This portfolio will correspond to the point where an indifference curve is just tangent to the efficient set. In Fig. 4, this is portfolio $\mathbf{P}^*$ on indifference curve I_2.

In Fig. 4, the curve I_3 represents the best level of satisfaction but it is of no use as it does not meet the feasible region at all. In regard to I_1, there are several portfolios that an investor could choose, say for example $\mathbf{P}$, but portfolio $\mathbf{P}^*$ dominates these portfolios, since it is on an indifference curve that is further northwest. Also, you may note that this point of intersection will be unique because of the convexity of the efficient frontier. Hence $\mathbf{P}^*$ is our optimal portfolio. With the selection of $\mathbf{P}^*$ our problem of selecting an optimal portfolio is solved completely. However, the model has certain limitations which we shall list now.

Limitations of Markowitz Model

Markowitz model presents a systematic procedure for the selection of the optimal portfolio of risky securities but is cumbersome to work with. It requires a full set of covariance between the returns of all securities to be considered in order to calculate portfolio variance. This leads to gathering an enormous amount of input data while dealing with the portfolios having large number of securities. Evaluation of a single portfolio of n-securities requires $2n$ quantities for estimating the expected return and standard deviation of these n-securities. There are $\frac{n(n-1)}{2}$ covariances to be estimated for these n-securities. Thus to evaluate a portfolio of n (risky) securities we need to estimate in all a total of $2n + \frac{n(n-1)}{2} = \frac{n(n+3)}{2}$ quantities. A lot of energy and time is needed to carry out the computations for the selection of an optimal portfolio. However, with the availability of computer packages, things can be managed for large values of n.

You may now try the following exercises.

E12) How many portfolios are on an efficient frontier? What is the Markowitz efficient set?

E13) How is an investor's risk aversion indicated in an indifference curve? Are all indifference curves upward sloping?

E14) Why indifference curves of an investor cannot intersect?

The limitations of the Markowitz model raise two issues:

i) Are there simpler methods for computing the efficient frontier?

ii) Can the Markowitz analysis be used to optimize asset classes rather than individual assets?

As mentioned in Unit 1, the model-building process proceeds through several **iterations**, each a refinement of the preceding model.. The Sharpe model developed by William Sharpe has simplified the rigours involved in computing the efficient frontier using the Markowitz approach. We shall now discuss this model in the next sub-section.

2.3.2 Sharpe Model

The Sharpe model commonly known as the single-index model, provides an alternative expression for portfolio variance, which is easier to calculate than in the case of the Markowitz analysis. This alternative approach can be used to solve the portfolio problem as formulated by Markowitz for determining the efficient set of portfolios. The single-index model, relates return on each security to the returns on a **common index**. Generally a broad market index of common stock returns is used for this purpose.

The single-index model can be expressed by the following equation

$$R_j = \alpha_j + \beta_j R_M + e_j \qquad (13)$$

where

R_j = the return on j^{th} security,

R_M = the return on the market index,

α_j = part of j^{th} security return independent of market performance,

β_j = a constant measuring the expected change in the dependent variable, R_i, given a change in the independent variable, R_M, and

e_j = the random residual error.

The single index model divides a security's return into two components α_i and a market related part $\beta_j R_M$. Given these values, the error term is the difference between the return on j^{th} security and the sum of two components of return, i.e.,

$$e_j = R_j - (\alpha_j + \beta_j R_M) \qquad (14)$$

For example, let us assume the return for the market index for period t is 10%, $\alpha_j = 4\%$ and $\beta_j = 1.5$. Then the estimate for stock j is

$$R_j = 4\% + 1.5 R_M + e_j$$

or, $R_j = 4\% + (1.5)\ (10\%) = 19\%$

This shows that if the market index return is 10%, then the likely return for your stock is 19%. Further, if the actual return on j^{th} stock for period t is 16%, the error term is $16\% - 19\% = -3\%$.

Note that R_M and e_j are random variables and the β term, or beta, is important as it measures the sensitivity of a stock to market movements. To use our model, we need to know for each stock we consider, the estimates of beta. Previously available data may be used to estimate future beta or analysts could also give some subjective estimate of beta.

The single-index model uses two simplifying assumptions. The **first assumption** used in the single-index model is that the random error term and the market index are uncorrected, meaning that the outcome of the market index has no bearing on the outcome of the random error term. The **second assumptions** is that the random error terms of any two securities are uncorrelated, meaning that the outcome of the random error term of i^{th} security has no bearing on the outcome of the random error term of j^{th} $(i \neq j)$ security. This can be expressed as $\text{cov}(e_i, e_j) = 0$. In other words, the returns of the two securities will be correlated (i.e., move together) only in their common response to the return on the market. This mean that stocks covary together only because of their common relationship to the market index. If either of these two assumptions are invalid i.e., they are not good description of reality, then model will be an approximation and alternately, more than one index model may be useful in such cases.

In the single-index model, all the covariance terms can be accounted for stocks being related only through common reactions to the market index, that is, covariance depends only on the market risk. Therefore, the covariance between any two securities can be written as

$$\sigma_{ij} = \beta_i \beta_j \sigma_M^2 \qquad (15)$$

where σ_M^2 is the variance of the market index. In the Markowitz model, we need to consider all the covariance terms in the variance-covariance market. In the single-index model, just as the security return is split into two components, the risk of an individual security is also divided into two components. This simplifies the covariance and also the calculations for the total risk for a security and for a portfolio are simplified to a large extent. The total risk of a security, as measured by its variances, consists of two components, market risk, which is common to all companies, and the company specific risk. Thus, we can write

$$\sigma_j^2 = \beta_j^2 [\sigma_M^2] + \sigma_{ej}^2 \qquad (16)$$

(Market risk) + (company specific risk)

This simplification also holds for portfolios. It provides an alternative expression for finding the minimum variance set of portfolios in the form

$$\sigma_P^2 = \beta_P^2 [\sigma_M^2] + \sigma_{ep}^2 \qquad (17)$$

$$\begin{pmatrix}\text{Total}\\ \text{portfolio}\\ \text{variance}\end{pmatrix} = \begin{pmatrix}\text{Portfolio}\\ \text{market}\\ \text{risk}\end{pmatrix} + \begin{pmatrix}\text{Portfolio}\\ \text{residual}\\ \text{variance}\end{pmatrix}$$

The single-index model greatly simplifies the calculation of the portfolio variance and therefore, the calculation of efficient portfolios. In the case of the Markowitz analysis, 200 stock require 19, 900 covariances and 200 variances. Whereas, we would require $3n + 2$ or 602 estimates for 200 securities while using single index model.

As in the case of Markowitz analysis, in order to determine the composition of the tangency portfolio, the investor needs to estimate all the expected returns, variances and covariances. In the case of single-index model, this can be done by estimating α_j, β_j and σ_{ej} for each of the n risky securities. Also needed

are the values of R_M and its variance σ_M^2. One way to estimate the parameters is with time series regression. With these estimates Eqns. (13), (16) and (17) can be used to calculate the returns, variances and covariances for the securities. Using these values, the curved efficient set of portfolios can be derived, from which the tangency portfolio can be determined.

Limitations of the Sharpe Model

The single-index model is a valuable simplification of the full variance-covariance matrix needed for the Markowitz model. It reduces by a large amount the number of estimates needed for a portfolio of securities. However, the model makes an assumption – the residuals for different securities are uncorrelated. The accuracy of the estimate of the portfolio variance thus depends on the accuracy of this assumption. For example, if the covariance between the residuals for different securities is positive, not zero as assumed, the true residual variance of the portfolio will be underestimated and the model will be in accurate. In such cases models with more than one index are available and can be used. We shall not be discussing such models here.

Before we end this unit you may try the following exercises.

E15) How is the covariance between any two securities calculated with the single-index model?

E16) How would you compare the Markowitz model with the Sharpe model?

We now end this unit by giving a summary of what we have covered in it.

2.4 SUMMARY

In this unit, we have covered the following points:

1. Once the essential characteristics of the real world problem are identified its conversion into a mathematical description in terms of equations can be done in different ways according to the objectives of the study as illustrated in the case of motion of a simple pendulum.
2. Portfolio selection problem: How can an investor choose an optimal portfolio, from a feasible set of risky securities i.e., choose a portfolio that provides him/her the maximum level of satisfaction in terms of return and risk.
3. Markowitz portfolio theory provides the way to select an optimal portfolio based on using the full information about securities.
4. An efficient portfolio has the highest expected return for a given level of risk or the lowest level of risk for a given level of expected return.
5. The Markowitz analysis determines the efficient set of portfolios, all of which are equally desirable. The efficient set in an expected return-standard deviation space is a curve which is upward concave.
6. The efficient frontier gives the investment possibilities that exist from a given set of securities. Indifference curves express investor preferences.

7. The optimal portfolio for a risk-averse investor occurs at the point of tangency between the investor's highest (northwest) indifference curve and the efficient set of portfolios.
8. The Sharpe model relates returns on each security to the returns on a common index.
9. The Sharpe model provides an alternative expression for portfolio variance, which is easier to calculate in comparison to the case of the Markowitz analysis.

2.5 SOLUTIONS/ANSWERS

E1) i) Dynamic, discrete, probabilistic
Essentials: risk, premium, number of customers, profit, etc.
Non-essentials: religion and gender of the people working in the company.

ii) Discrete, dynamic, probabilistic
Essentials: price, advertising, competition, population changes, seasonal variations, etc.
Non-essentials: name, religion and gender of the people working in the company.

E2) State the problem giving reasons for why there is no other treatment of the problem possible.

E3) i) The meteorological quantities like rainfall, atmospheric temperature, pollution are all continuous and stochastic. Similarly give examples for ii) and iii).

E4) i) 20% ii) −7.5% iii) 1.67%

E5) $\mathbf{P}_1 = \left(0, \frac{1}{3}, \frac{2}{3}\right)$, $\mathbf{P}_2 = \left(0, \frac{1}{4}, \frac{3}{4}\right)$, $\mathbf{P}_3 = \left(0, \frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$. Similarly find others.

E6) $E(R) = 0.0832$.

E7) $E(R) = 0.1075$, $\sigma = 7.75\%$.

E8) $\sigma_1 = 0.0337$, $\sigma_2 = 0.0362$. Security 2 is more risky than security 1.

E9) $E(1) = 0.1264$, $E(2) = 0.1443$, $\sigma_1 = 0.0337$, $\sigma_2 = 0.0362$

$$\sigma_{12} = 0.33 \times (0.0636) \times (-0.0543) + (0.25) \times (0.0436) \times (0.0157) + 0.22 \times (-0.0164)\ (0.0357) + 0.2 \times (-0.0264) \times (-0.0357)$$

E10) a) 0.1975 b) 0.207 c) 0.235

E11) $\sigma_P = 0.0371$.

E12) There are large number of portfolios on an efficient frontier. Markowitz efficient set is the set of portfolios having highest expected return for a

given level of risk or the lowest level of risk for a given level of expected returns.

E13) Yes, all indifference curves are upward sloping.

E14) Let three be a point X of intersection of two indifference curves I_1 and I_2. Show that all portfolios on I_1 must be as desirable as those on I_2 and then reach a contradiction since I_1 and I_2 are two curves that are supposed to represent different levels of desirability.

E15) Explain Eqn. (15).

E16) Comparison between the two models can be made in terms of risk and return evaluation, in terms of number of estimates required, etc.

—x—

UNIT 3 DATA ANALYSIS AND FITTING MODELS TO DATA

Structure **Page No**

3.1 INTRODUCTION

Most scientific disciplines are concerned with measuring items and collecting data. One reason for this is the increasingly quantitative approach employed in all the sciences, business and many other activities which directly affect our lives. Volumes of data on various items like population, taxes, wealth, exports, imports, crop production, etc. are collected, processed and disseminated to the public for one reason or the other. Data relevant for a study can be obtained either from published works or from the collections of the government or research organisations or through direct field work. In order to draw any meaningful conclusion from the data, it is important to represent and analyse the collected data in an effective way. With the help of statistics, data can be used to

- analyse and summarise a situation;
- model experimental outcomes;
- quantify uncertainty;
- make decisions;
- predict future outcomes.

Statisticians generally obtain data by taking a **sample** from some larger **population**. We take samples because the entire population may be too large to be examined in a timely and economic manner. Regression analysis is one of the techniques used in statistics for modelling and analysing numerical data. Regression models describe the variation in one dependent variable (also called response variable or measurement) when one or more independent variables (also known as explanatory variables or predictors) vary. Applications of regression are numerous and occur in almost every field including engineering, physical sciences, economics, management, life and biological sciences, and social sciences.

A general regression model consists of a function describing how the response is related to one or more predictor variables and a term which models the random variation in the response variable. This relationship may be linear or nonlinear combination of one or more model parameters and accordingly the

model is classified as linear or nonlinear. In this unit, we shall discuss some of the **linear regression models**. For example, models relating to time and distance, price and sale, performance of a car related to engine displacement and the horsepower of the engine, etc.

We shall start the unit by giving you in Sec. 3.2, the general idea of data visualization i.e., visual representation of data. We shall discuss linear regression models with one predictor in Sec. 3.3 and with two or more predictors in Sec. 3.4. Second order polynomial models in one variable are also introduced in this section. You are also required to do computer practical exercises on this unit which are given at the end of the unit.

Objectives

After studying this unit, you should be able to

- use quantitative and graphical techniques for data visualization;
- distinguish between simple and multiple linear regression models;
- write down simple or multiple linear regression models appropriate for a given set of data;
- use the method of least squares for estimating the parameters of linear regression model;
- fit a line/curve/plane/surface, appropriate, for a given set of data.

3.2 DATA VISUALIZATION

Data visualization is the study of the visual representation of data. It is an information which has been abstracted in some schematic form including attributes or variables for the units of information. Data visualization is closely related to information graphics, information visualization, scientific visualization and statistical graphics. It refers to the more technologically advanced techniques which allow visual interpretation of data through the representation, modelling and display of solids, surfaces, properties and animations, involving the use of graphics, image processing, computer vision and user interfaces. It encompasses a much broader range of techniques than specific techniques as solid modelling.

The origin of this field are in the early days of computer graphics in the 1950s, when the first graphs and figures were generated by computers. With the rapid increase of computing power, larger and more complex numerical models were developed, resulting in the generation of huge numerical data sets. Also, large data sets were generated by data acquisition devices such as medical scanners and microscopes, and data was collected in large databases containing text, numerical information and multimedia information. Advanced computer graphics techniques were needed to process and visualize these massive data sets. Once the data are converted to a visual form, the trend and patterns are often immediately apparent. Fig. 1 shows an example of large data set that has been converted to colour-coded display. It shows the Indian map with the population classified and presented using different colours.

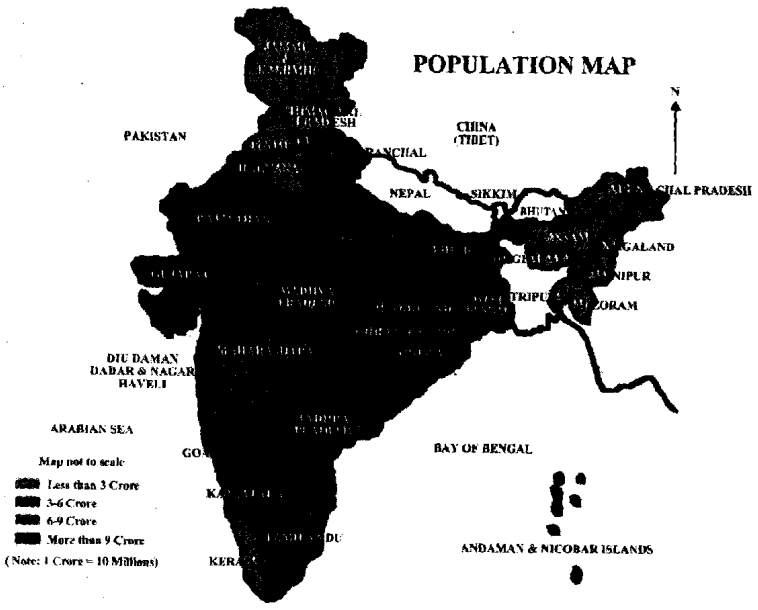


Fig.1: Population map of India-2001.

Visualization in scientific computing was used initially to refer to visualization as a part of a process of scientific computing: the use of computer modelling and simulation in scientific and engineering practice. Some of the most popular examples of scientific visualizations are computer-generated images that show real spacecraft in action, out in the void far beyond earth, or on other planets. TV also offers scientific visualization when it shows computer drawn and animated reconstructions of road or airplane accidents. More recently, visualization is also increasingly concerned with data from other sources including large and heterogeneous data collections found in business and finance, administration, digital media, etc.

A new research area called Information Visualization was launched in the early 1990s to support analysis of abstract and heterogeneous data sets in many application areas. Therefore, the phrase "Data Visualization" is gaining acceptance to include both the scientific and information visualization fields.

Two main parts of data visualization mainly presumed are statistical graphics, and thematic maps.

Statistical Graphics

Statistical graphics, also known as graphical techniques, are information graphics in the field of statistics used to visualize quantitative data. Statistics and data analysis procedures can broadly be split into two parts: **quantitative techniques** and **graphical techniques**. Quantitative techniques are the set of statistical procedures that yield numeric or tabular output. Examples of quantitative techniques include hypothesis testing, analysis of variance, point estimation, confidence intervals, and least squares regression. These and similar techniques are all valuable and are mainstream in terms of classical analysis.

On the other hand, there is a large collection of statistical tools that we generally refer to as graphical techniques. You must be familiar with most of

these techniques through your undergraduate statistics course. These include: scatter plots, bar diagrams, histograms, probability plots, residual plots, box plots, block plots, and biplots. Exploratory data analysis (EDA) relies heavily on these and similar graphical techniques. Fig. 2 gives the bar diagram for the data showing the marks obtained (out of 10) by 100 students in an examination.

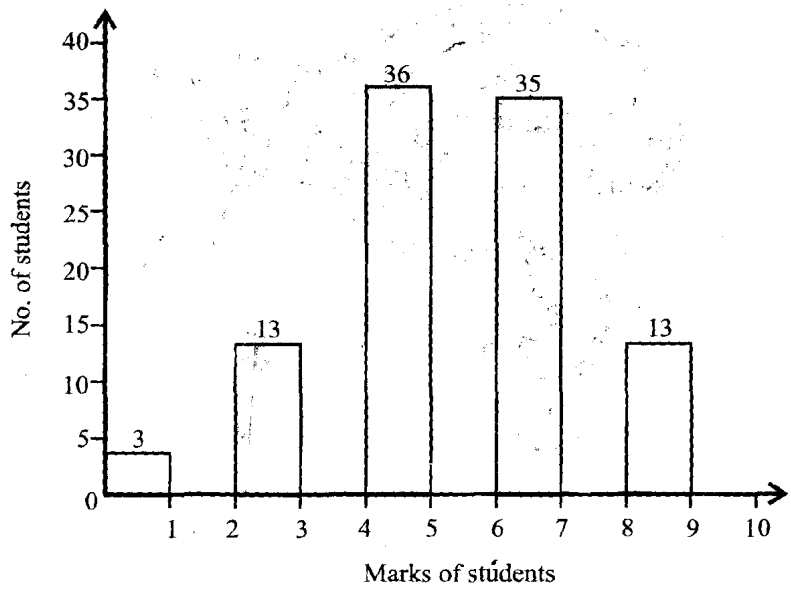


Fig. 2: Bar diagram of marks obtained by students.

Graphical procedures are not just tools used in an EDA context; such graphical tools are the shortest path to gaining insight into a data set in terms of testing assumptions, model selection and statistical model validation, estimator selection, relationship identification, factor effect determination, and outlier detection. In addition, good statistical graphics can provide a convincing means of communicating the underlying message, that is present in the data, to others. To sum up we can say that the graphical statistical methods have the following four objectives:

- The exploration of the content of a data set;
- The use to find structure in data;
- Checking assumptions in statistical models;
- Communicate the results of an analysis.

Thematic Maps

A thematic map displays spatial pattern of a theme or series of attributes. In contrast to reference maps which show many geographic features (forests, roads, political boundaries), thematic maps emphasise spatial variation of one or a small number of geographic distributions. These distributions may be physical phenomena such as climate or human characteristics such as population density and health issues. These types of maps are sometimes referred to as graphic essays that portray spatial variations and interrelationships of geographical distributions. Location, of course, is also important to provide a reference base of where selected phenomena are occurring. While general reference maps show where something is in space, thematic maps tell a story about that place. Fig. 3 gives a Pie chart showing the

proportion of people from different states of India living in a certain town of West Bengal.

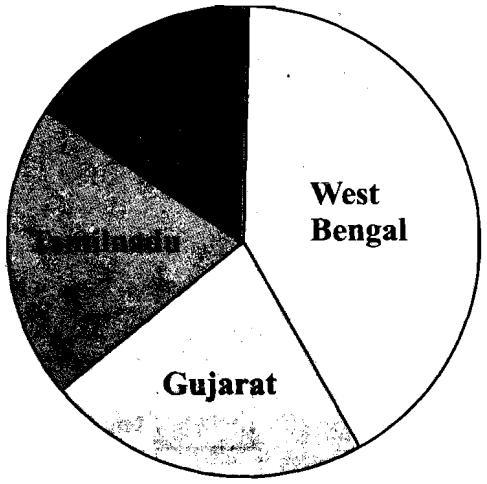


Fig. 3: Pie chart showing the proportion of people state-wise.

You may now try the following exercises.

E1) Explain at least four statistical graphical techniques you are familiar with. Illustrate the advantages of each of them through examples.

E2) Give at least two examples of the thematic maps with which you are familiar. Also explain the purpose of using each of them.

Once the data related to any study is collected, the next step is to make use of this data to draw meaningful conclusions, i.e., analyse the data about the subject under study. As we have already mentioned, regression models which are statistical models are useful in almost all the areas biological, physical, social sciences, business, engineering, etc. in both the planning stages of research and analysis of the resulting data. Care should be devoted to accurate data collection because the conclusions from the analysis depend on the data. A good data collection will result in better analysis and more applicable model. If the data used in a regression model are not representative of the system studied, then conclusions drawn from the model are likely to be in error.

Let us now discuss some linear regression models one by one

3.3 SIMPLE LINEAR REGRESSION MODELS

In simple linear regression, relationship between two variables, for example, income and number of years of education; height and weight of people; length and width of envelopes; temperature and output of an industrial process; altitude and boiling point of water; or dose of a drug and response are modelled. The linear relationship between any two such variables can be represented by a straight line. Let us now illustrate through an example the formulation of simple linear regression model.

Example 1 (Driving a car at a constant speed): Suppose you are driving a car at 50 km/h. The two variables of interest to you are time and distance. If

you are driving at constant speed, then the theoretical relationship between time and distance covered is given by the straight line as shown in Fig. 4(a).

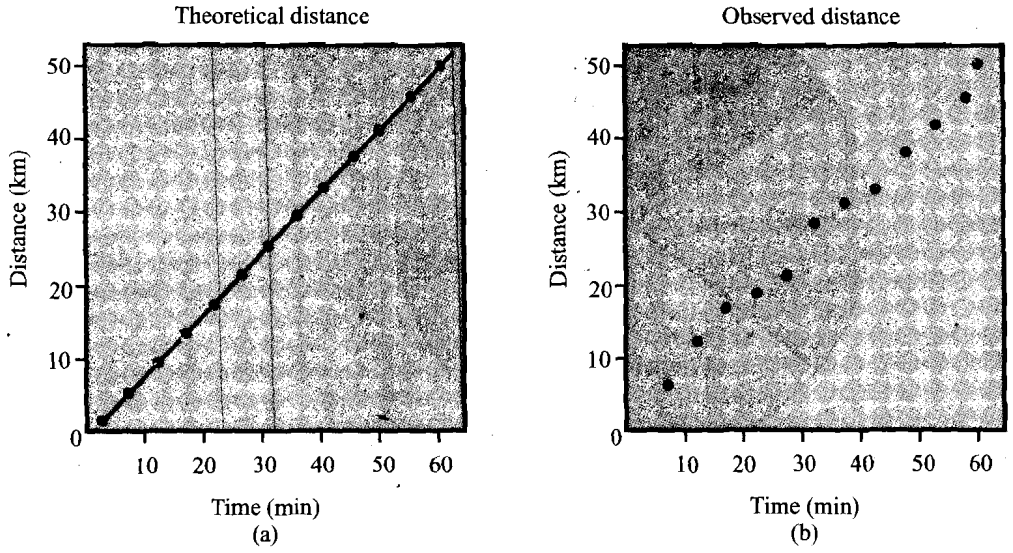


Fig. 4: Scatter plot of the time and distance data.

In a perfect world, where speed and distance could be measured without error, all observations would lie exactly on this straight line. However, in reality it is impossible to keep the speed exactly constant and to measure the precise distance. Therefore, in a scatter plot of 'real' data, the points would deviate from the theoretical straight line. A realistic scatter plot might look like Fig. 4(b). For this example, we need a model that will describe the linear relationship between the two variables, and, at the same time, take the variation away from the line into account.

Formulation

If we let y represents the distance covered and x represents time then the equation of the straight line in Fig. 4(a), relating these two variables is

$$y = b_0 + b_1 x . \qquad (1)$$

where b_0 is the intercept and b_1 is the slope. Now the data points in Fig. 4(b) do not fall exactly on a straight line so Eqn. (1) should be modified to account for this. Let the difference between the observed value of y and the straight line $(b_0 + b_1 x)$ be an error e. That is, it is a device that accounts for the failure of the model to fit the data exactly.

Thus, a more appropriate model for the data in Fig. 4(b) is

$$y = b_0 + b_1 x + e . \qquad (2)$$

Eqn. (2) is called a **linear regression model.** Variable x is called the **predictor** or **regressor** variable and y the **response** variable. b_0 and b_1 are the parameters of the model and e is **a random error component.** The errors are assumed to have mean zero and unknown variance σ^2. We also assume that the errors are uncorrelated. This means that the value of one error does not depend on the value of any other error. Because Eqn. (2) involves only one

regressor variable it is called a **simple linear regression** model. In gerenal, the response may be related to k regressors, $x_1, x_2, \ldots, x_k$, so that

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k + e. \qquad (3)$$

Eqn. (3) represents a **multiple linear regression** model, which we shall discuss in the next section. A model is called **linear** because it is linear in the parameters $b_0, b_1, \ldots, b_k$, and not because y is a linear function of x's. You will see that many models in which y is related to the x's in a nonlinear fashion are treated as linear regression models as long as the equation is linear in the b's. Once model (2) is formulated, we would like to use it to obtain information on the y's for specific x values. For that we have to estimate the unknown parameters in the regression model by making the error minimum and this process is called **fitting the model to the data**. There are several parameter estimation techniques available. Here we shall be using the method of least squares for parameter estimation.

Least squares estimation of the parameters

The parameters b_0 and b_1 are unknown and must be estimated using sample data. Suppose we have n pairs of data, say $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$. Then the estimation of b_0 and b_1 is done as follows:

Estimation of b_0 and b_1

Let us use the method of least squares to estimate b_0 and b_1. That is, we will estimate b_0 and b_1 so that the sum of the squares of the difference between the observations y_i and the straight line is a minimum. From Eqn. (2), we may write

$$y_i = b_0 + b_1x_i + e_i, \qquad i = 1, 2, \ldots, n. \qquad (4)$$

Eqn. (2) may be viewed as a **population** regression model while Eqn. (4) is a **sample** regression model, written in terms of the n pairs of data (y_i, x_i), $i = 1, 2, \ldots, n$. Thus, the least squares criterion is that the error

$$S(b_0, b_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1x_i)^2 \qquad (5)$$

should be minimum.

The least squares estimators or the 'best' estimates of b_0 and b_1, say $\hat{b}_0$ and $\hat{b}_1$, are the values that minimize e and, therefore, must satisfy

$$\left.\frac{\partial S}{\partial b_0}\right|_{\hat{b}_0, \hat{b}_1} = -2\sum_{i=1}^{n} (y_i - \hat{b}_0 - \hat{b}_1x_i) = 0 \qquad (6)$$

and

$$\left.\frac{\partial S}{\partial b_1}\right|_{\hat{b}_0, \hat{b}_1} = -2\sum_{i=1}^{n} (y_i - \hat{b}_0 - \hat{b}_1x_i)\, x_i = 0. \qquad (7)$$

Simplifying Eqns. (6) and (7), we get

$$n\,\hat{b}_0 + \hat{b}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \,. \tag{8}$$

$$\hat{b}_0 \sum_{i=1}^{n} x_i + \hat{b}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \,. \tag{9}$$

Eqns. (8) and (9) are called the **least squares normal equations**. The solution to the normal equations is obtained as

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x} \tag{10}$$

and

$$\hat{b}_1 = \frac{\sum_{i=1}^{n} y_i x_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2} \tag{11}$$

where, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (12)

are the averages of y_i and x_i, respectively. Therefore, $\hat{b}_0$ and $\hat{b}_1$ in Eqns. (10) and (11) are the least squares estimators of the **intercept** and **slope**. The fitted simple linear regression model can then be written as

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x \,. \tag{13}$$

This line is called the **least squares line** or the **regression line**. For a given set of data, the least squares lines is the best-fitting line, in the sense that it minimize e. The values $\hat{y}$ are referred to as **predicted values** or **fitted values**. The difference between the observed value y_i and the corresponding fitted value $\hat{y}_i$ is a **residual**.

Mathematically, the i^{th} residual is

$$e_i = y_i - \hat{y}_i = y_i - (\hat{b}_0 + \hat{b}_1 x_i), \quad i = 1, 2, \ldots, n \,. \tag{14}$$

Residuals play an important role in investigating the adequacy of the fitted regression model and in detecting departures from the underlying assumptions.

Using simple algebra you can write the denominator and numerator of Eqn. (11) in a more compact notation as

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{15}$$

and

$$S_{xy} = \sum_{i=1}^{n} y_i x_i - n\,\bar{x}\,\bar{y} = \sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y}) \,. \tag{16}$$

Thus, Eqn. (11) can be conveniently written as

$$\hat{b}_1 = \frac{S_{xy}}{S_{xx}} \,. \tag{17}$$

Before proceeding further, you may try the following exercise.

E3) Verify Eqns. (15) and (16).

In addition to estimating b_0 and b_1, we also need an estimate of σ^2, the **variance of error**. It gives the variance of the data points around the least squares line.

Estimation of σ^2

The estimate of σ^2 is obtained from the **residual** or **error sum of squares**

$$SS_e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 . \quad (18)$$

Substituting $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$ in Eqn. (18) and simplifying, we get

$$SS_e = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \hat{b}_1 S_{xy} . \quad (19)$$

But, $\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 = S_{yy}$.

Thus, $SS_e = S_{yy} - \hat{b}_i S_{xy}$. (20)

The residual sum of squares has $n-2$ degrees of freedom, because two degrees of freedom are associated with the estimates $\hat{b}_0$ and $\hat{b}_1$ involved in obtaining $\hat{y}_i$. Thus, an estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{SS_e}{n-2} = MS_e . \quad (21)$$

The quantity MS_e is called the **error mean square** or the **residual mean square**. The square root of $\hat{\sigma}^2$ is called the **standard error of regression**, and has the same units as the response variable y. Because $\hat{\sigma}^2$ depends on the residual sum of squares, any violation of the assumptions on the model errors or any misspecification of the model form may damage the usefulness of $\hat{\sigma}^2$ as an estimate of σ^2. We now illustrate the estimation of the model parameters through examples.

Example 2 (Demand for homes): Find a linear demand equation that best fits the following data, and use it to predict annual sales of homes priced at Rs.14,00,000.

x = Price (lakhs of Rs.)	16	18	20	22	24	26	28
y = sales of new homes this year	126	103	82	75	82	40	20

Solution: Calculations are shown in Table 1.

Table 1

	x	y	xy	x^2
	16	126	2016	256
	18	103	1854	324
	20	82	1640	400
	22	75	1650	484
	24	82	1968	576
	26	40	1040	676
	28	20	560	784
sums	$\Sigma x = 154$	$\Sigma y = 528$	$\Sigma xy = 10{,}728$	$\Sigma x^2 = 3500$

Substituting these values in the formula given by Eqns. (10) and (11), we ge

$$\text{slope} = \hat{b}_1 = \frac{n(\Sigma xy) - (\Sigma x)\,(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{7(10,728) - (154)\,(528)}{7(3500) - 154^2} \approx -7.929$$

$$\text{intercept} = \hat{b}_0 = \frac{\Sigma y - m(\Sigma x)}{n} = \frac{528 - (-7.929)\,(154)}{7} \approx 249.9.$$

Thus, our least squares line is $\hat{y} = 249.9 - 7.929x$.

We can now use this equation to predict the annual sales of homes priced at Rs.14,00,000. Remembering that x is the price in lakhs of rupees, we set $x = 14$, and solve for y, getting $y \approx 139$. Thus, our model predicts that approximately 139 homes priced at Rs. 14,00,000 will be sold.

***

Example 3: Consider the data shown in Table 2

Table 2

x	2	9	3	5	1
y	1	17	3	9	0

Use a best fit line to estimate the value of y for $x = 6$ and 8. Also obtain the estimate of the error variance of the best fit.

Solution: Calculations of the least squares line are given in Table 3.

Table 3

	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$
	2	1	−2	−5	10	4
	9	17	5	11	55	25
	3	3	−1	−3	3	1
	5	9	1	3	3	1
	1	0	−3	−6	18	9
Sum	20	30	0	0	89	40

Therefore, $\hat{b}_1 = \frac{89}{40} = 2.225$

and $\hat{b}_0 = 6 - 2.225(4) = 6 - 8.9 = -2.9$.

The least squares line is $\hat{y} = -2.9 + 2.225x$. Thus, if $x = 6$, we would predict that

$$\hat{y} = -2.9 + 2.225(6) = 10.45.$$

Similarly, if $x = 8$ we would predict

$$\hat{y} = -2.9 + 2.225(8) = 14.9.$$

Calculations for SS_e are done in Table 4.

Table 4

	x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
	2	1	1.550	−0.550	0.3025
	9	17	17.125	−0.125	0.015625
	3	3	3.775	−0.775	0.600625
	5	9	8.225	0.775	0.600625
	1	0	−0.675	0.675	0.455625
Sum				0	1.975

Thus, $SS_e = 1.975$, and it follows from Eqn. (21) that

$\hat{\sigma}^2 = 1.975/(5-2) = 0.6583$.

***

Once the parameters of the model have been estimated, it is important to evaluate the model adequacy. For this purpose, let us determine the coefficient of determination.

Coefficient of Determination

The quantity R^2 defined by

$$R^2 = \frac{SS_R}{S_{yy}} = 1 - \frac{SS_e}{S_{yy}} \qquad (22)$$

is called the coefficient of determination where $SS_R = \sum_i (\hat{y}_i - \bar{y})^2$ is the **regression sum of squares** and $S_{yy} = \sum_i (y_i - \bar{y})^2$ is the **total sum of squares**. The total sum of squares can be partitioned into

$$S_{yy} = SS_R + SS_e$$

i.e., $$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 . \qquad (23)$$

You may **notice** that S_{yy} determines the variability in y without considering the effect of the regressor variable x, and SS_e is a measure of the variability in y remaining after x has been considered. Thus, R^2 in Eqn. (22) gives the **proportion of variation in y explained by the regressor x**. Because $0 \le SS_e \le S_{yy}$, it follows that $0 \le R^2 \le 1$. Values of R^2 that are close to 1 imply that most of the variability in y is explained by the regression model. For the data in Example 3, we have

$$R^2 = 1 - \frac{SS_e}{S_{yy}} = 1 - \frac{1.975}{300} = 0.9934$$

that is, 99.34 percent of the variability in the data is accounted for by the regression model.

The statistic R^2 should be used with caution, since it is always possible to make R^2 large by adding enough terms to the model. For example, if there are no repeat points (more than one y value at the same x value) a polynomial of degree $n-1$ will give a perfect fit ($R^2 = 1$) to n data points. When there are repeat points, R^2 can never be exactly equal to 1, because the model cannot explain the variability due to 'pure' error. Although R^2 increases if we add a regressor variable to the model, this does not necessarily mean the new model is superior to the old one. Unless the error sum of squares in the new model is reduced by an amount equal to the original error mean square, the new model will have a larger error mean square than the old one because of the loss of one degree of freedom for error. Thus, the new model will actually be worse than the old one.

We now list some of the limitations of the regression models. Limitations which we are going to list are in fact the limitations of regression analysis in general and apply to all the regression models.

Limitations

1. Regression models are **intended as interpolation equations** over the range of the regressor variable(s) used to fit the model. They may not be valid for extrapolation outside of this range. For example, see Fig. 5.

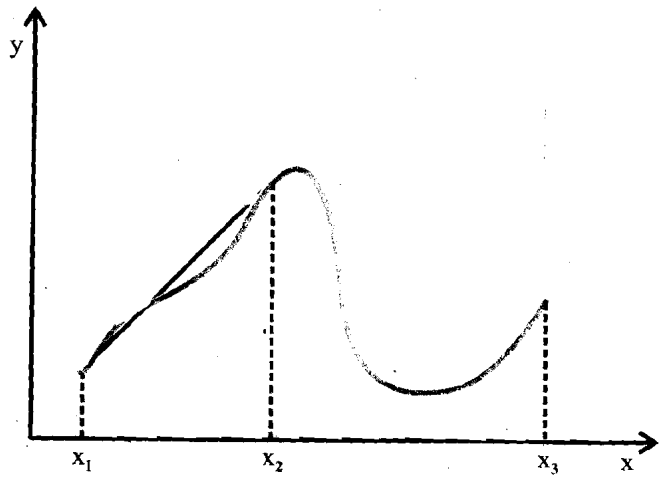


Fig. 5: The risk of extrapolation in regression.

Suppose that data on y and x were called in the interval $x_1 \leq x \leq x_2$. Over this interval, the linear regression equation shown in Fig. 5 is a good approximation of the true relationship. However, if this equation is used to predict values of y for values of the regressor variable in the region $x_2 \leq x \leq x_3$, then model is useless over this range of x because of equation error.

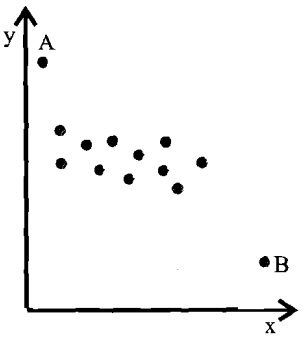


Fig. 6: Two extreme observations.

2. The position of the **x-values play an important role** in the least squares fit. While determining the height of the line, all points have equal weight. Whereas, the slope of the line is strongly influenced by the remote values of x. For example, consider the data in Fig. 6. The slope in the least squares fit depends heavily on either or both of the points A and B. The remaining data would give a very different estimate of the slop if A and B were deleted.

3. Regression techniques indicate a strong relationship between two variables, this **does not imply** that the variables are related in any **causal sense**. Our expectations of discovering cause and effect relationship from regression should be modest. For example, if you look at the data given in E6), you will see that the linear trend between x and y does not establish cause and effect between homework and test results.

4. In some applications of regression the value of the regressor variable x required to predict y is unknown. For example, consider predicting maximum daily load on an electric power generation system from a regression model relating the load to the maximum daily temperature. To predict tomorrow's maximum load, we must first predict tomorrow's maximum temperature. Thus, the prediction of maximum load is conditional on the temperature forecast. The accuracy of the maximum load forecast depends on the accuracy of the temperature forecast. All these considerations must be taken into account when evaluating model performance.

You may now try some exercises.

E4) In 1978, a company conducted a study of the amount of additional oil that can be extracted from existing oil wells by "enhanced recovery techniques" (such as injecting solvents or steam into an oil well to lower the density of the oil). The following table gives the study's estimates of recoverable oil based on the price per litre

Price per Litre (Rs.)	92	102	97	75
Recovery	21.2	29.4	41.6	49.2

Use a best fit line to estimate the additional amount of oil that can be economically recovered.

E5) The percentage of new plant and equipment expenditure by companies on pollution control is as shown.

1975	1980	1981	1984	1987
8.4	8.1	7.3	6.8	5.1

Use a linear regression model to estimate the figure for 1985.

E6) Students in a statistics class claimed that doing the homework had not helped prepare them for the midterm exam. The exam score y and homework score x for the 18 students in the class were as follows:

y	x	y	x	y	x
95	96	72	89	35	0
80	77	66	47	60	30
0	0	98	90	72	59
0	0	90	93	55	77
79	78	0	18	75	74
77	64	95	86	66	67

Fit a simple linear regression model to the data and interpret the result. Calculate R^2 for the data.

Fitting Exponential Curves

So far, we have seen how to fit a straight line to a set of bivariate data i.e., data giving relationship between two variables. A straight line is the simplest model for a set of bivariate data, and is not appropriate if the scatter plot of the data shows curvature. One class of models that accounts for curvature is the class of exponential curves. Some examples of exponential curves are

$$y = b_0 x^{b_1} \tag{24}$$

$$y = b_0 e^{b_1 x} \tag{25}$$

$$e^y = b_0 x^{b_1}. \tag{26}$$

Models (24)-(26) have two important features. They are monotone (either increasing or decreasing), and they are easy to fit. Monotonicity is important because often, in practice, a monotone relationship exists between two variables. For example, if x is the level of traffic in a communications network and y is the number of packets lost, it only makes sense for y to increase as x increases. Models (24)-(26) are easy to fit using the technique for fitting lines. This is because, when we transform them to a different scale, they are actually lines. Which, if any, of these models is appropriate for a particular set of data is best determined by drawing scatter plots of the transformed data and see which transformation makes the data look the most linear. Let us see how these transformations are done.

The Model $y = b_0 x^{b_1}$.

Taking logarithms on both sides of Eqn. (24) and then adding an error term, we obtain

$$\ln y = \ln b_0 + b_1 \ln x + e$$

or, $$y^* = b_0^* + b_1 x^* + e \qquad (27)$$

where, $y^* = \ln y$, $x^* = \ln x$ and $b_0^* = \ln b_0$.

You may **notice** that the model (27) is a line in the variables y^* and x^*. In order to estimate the parameters, we can simply calculate ln (y) and ln (x) for all the data points, and then use least squares to fit a straight line. We now illustrate the method through an example.

Example 4: Consider the set of data given in Table 5.

Table 5

x	9	2	5	8	4	1	3	7	6
y	253.5	7.1	74.1	165.2	32.0	1.9	18.1	136.1	100.0

Use a best fit line to estimate the value of y when $x = 3.36$. Also obtain the residual for the fitted line.

Solution: A graph of these data in Fig. 7 shows that a line is not the appropriate model and suggests that model (24) might be reasonable.

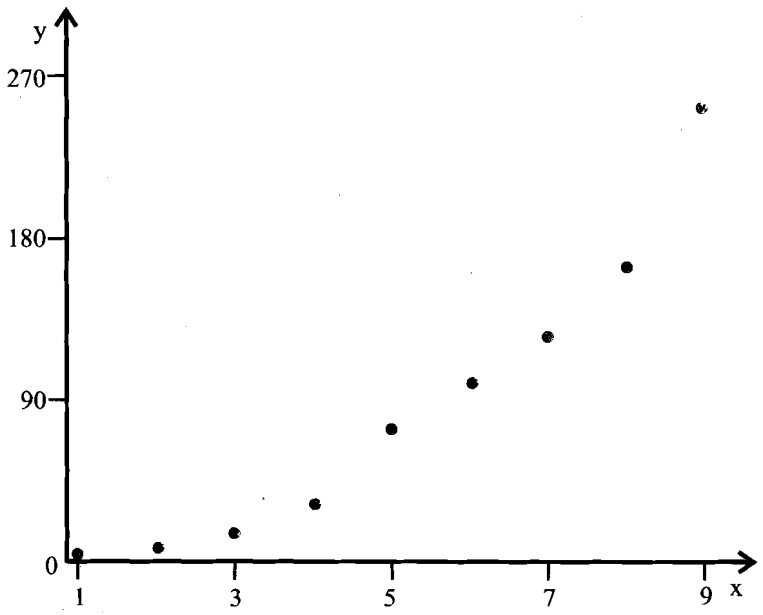


Fig. 7: A scatter plot of the data in Example 4.

Taking logarithms of all the data points we obtain the values as given in Table 6.

Table 6

$x^* = \ln(x)$	2.197	0.693	1.609	2.079	1.386	0	1.099	1.946	1.792
$y^* = \ln(y)$	5.54	1.96	4.30	5.11	3.47	0.65	2.90	4.91	4.61

A scatter plot of the log-transformed data in Fig. 8 looks quite linear.

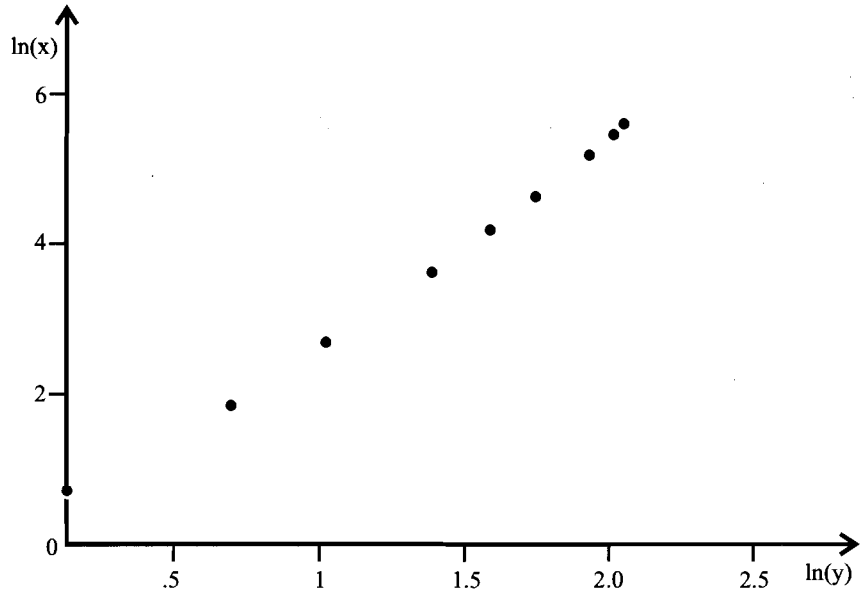


Fig. 8: A scatter plot of the log transformed data in Example 4.

To obtain the estimates for the parameters in the model we calculate

$$\overline{x^*} = 1.422$$

$$\overline{y^*} = 3.715$$

$$\sum (x^* - \overline{x^*})^2 = 4.138$$

$$\sum (x^* - \overline{x^*})\,(y^* - \overline{y^*}) = 9.315.$$

This gives,

$$\hat{b}_1 = 9.315/4.138 = 2.251$$

$$\hat{b}_0^* = 3.715 - 2.251(1.422) = 0.51$$

and our fitted model is

$$\hat{y}^* = 0.51 + 2.251x^*$$

$$\ln(\hat{y}) = 0.51 + 2.251 \ln(x)$$

or,

$$\hat{y} = 1.67\ x^{2.251}.$$

Thus, when $x = 3.36$, we would predict

$$\hat{y} = 1.67\ (3.36)^{2.251} = 25.6.$$

Residuals for this fitted line could be computed using Eqn. (14). For example,

$$\hat{e}_1 = 5.54 - [0.51 + 2.251 \ln(2.197)] = 0.08.$$

***

Similarly, the transformations for the models (25) and (26) can also be obtained.

For the Model $y = b_0 e^{b_1 x}$, taking logarithms on both sides of Eqn. (25) and adding an error terms, we obtain,

$$\ln(y) = \ln b_0 + b_1 x + e$$

or,

$$y^* = b_0^* + b_1 x + e. \tag{28}$$

Thus, we can estimate the parameters of model (25) by calculating $\ln(y)$ for each data point and then using least squares with $\ln(y)$ as the dependent variable and x as the independent variable.

For the model $e^y = b_0 x^{b_1}$, once again, taking logarithms, on both sides of Eqn. (26) and adding an error term, we get

$$y = \ln b_0 + b_1 \ln(x) + e$$

$$= b_0 + b_1 x^* + e. \tag{29}$$

To have a better understanding of models (25) and (26), you can solve the following exercises.

E7) Find a linear regression equation that best fit the data given in Table 7.

Table 7

x	4.99	1.96	2.98	9.00	4.04
y	4.35×10^{9}	1.97×10^{4}	5.99×10^{4}	3.73×10^{17}	1.67×10^{8}

x	6.06	0.88	8.02	6.97
y	7.20×10^{10}	180	5.97×10^{14}	3.67×10^{12}

E8) Find a linear regression equation that best fit the data given in Table 8.

Table 8

x	17,000	47	150	1000	14	4300	200	18
y	45	15	18	25	9	32	20	10

Just as we can fit a line or a curve to two-dimensional data, we can also fit a plane or curved surface to three-dimensional data and a hyper-plane or hyper-surface to four- or higher-dimensional data.

We shall now discuss models that are suitable to fit three and higher dimensional data.

3.4 MULTIPLE LINEAR REGRESSION MODELS

The process of fitting models for surfaces and hypersurfaces is the same as for lines. Let us start with an example of the situation involving two predictor variables.

Formulation

Suppose we want to develop a model for estimating the effective life of a cutting tool based on two variables namely, the cutting speed and the depth of cut. A regression model that might describe this relationship is

$$y = b_0 + b_1 x_1 + b_2 x_2 + e \tag{30}$$

which is the equation for a plane in three dimensions where y denotes the effective tool life, x_1 denotes the cutting speed, and x_2 denotes the depth of cut. b_0, b_1, b_2 are the unknown parameters of the model and e is the random error. Eqn. (30) is a **multiple linear regression** model **with two predictor** variables. The term linear is used because Eqn. (30) is a linear function of the unknown parameters b_0, b_1 and b_2.

In general, the response y may be related to k predictor variables. The model

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + e \tag{31}$$

is called a **multiple linear regression** model **with k predictors**. The parameters b_j, $j = 0, 1, \ldots, k$ are called the regression coefficients and it represents the expected change in the response y per unit change in x_j when all the remaining predictor variables x_i $(i \neq j)$ are held constant. For this reason the parameters b_j, $j = 1, 2, \ldots, k$ are often called **partial regression coefficients**.

Least Squares Estimation of the Parameters

Suppose that $n > k$ observations are available, and let y_i denotes the i^{th} observed response and x_{ij} denote the i^{th} observation or level of regressor x_j. We assume that the error term e has the mean zero and variance σ^2, and that the errors are uncorrelated.

Estimation of Regression Coefficients

Let us use the method of least squares for estimating the regression coefficients in model (31).

We write the **sample model** corresponding to Eqn. (31) as

$$\begin{aligned} y_i &= b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik} + e_i \\ &= b_0 + \sum_{j=1}^{k} b_j x_{ij} + e_i, \qquad i = 1, 2, \ldots, n. \end{aligned} \tag{32}$$

The least squares function is

$$S(b_0, b_1, \ldots, b_k) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{n} b_j x_{ij} \right)^2 \tag{33}$$

The function S is to be minimized with respect to $b_0, b_1, \ldots, b_k$. The least squares estimators $\hat{b}_0, \hat{b}_1, \hat{b}_2, \ldots, \hat{b}_k$ of $b_0, b_1, b_2, \ldots, b_k$ must satisfy

$$\left.\frac{\partial S}{\partial b_0}\right|_{\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_k} = -2\sum_{i=1}^{n}\left(y_i - \hat{b}_0 - \sum_{j=1}^{k}\hat{b}_j\, x_{ij}\right) = 0 \qquad (34)$$

and

$$\left.\frac{\partial S}{\partial b_j}\right|_{\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_k} = -2\sum_{i=1}^{n}\left(y_i - \hat{b}_0 - \sum_{j=1}^{k}\hat{b}_j\, x_{ij}\right) x_{ij} = 0, \quad j = 1, 2, \ldots, k. \qquad (35)$$

Simplifying Eqns. (34) and (35), we obtain $(k+1)$ least squares normal equations, one for each of the unknown regression coefficients. The solution to the normal equations will be the least squares estimators, $\hat{b}_0, \hat{b}_1, \ldots \hat{b}_k$.

It is more convenient to deal with multiple regression models in matrix notation. Using matrix notation Eqn. (32) can be written as

$$\mathbf{y} = \mathbf{x}\mathbf{b} + \mathbf{e} \qquad (36)$$

where $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, $\mathbf{x} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$

$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix}$ and $\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$.

Now we want to find the vector $\hat{\mathbf{b}}$, which minimizes

$$S(\mathbf{b}) = \sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{x}\mathbf{b})'\,(\mathbf{y} - \mathbf{x}\mathbf{b})$$

$$= \mathbf{y}'\mathbf{y} - \mathbf{b}\,\mathbf{x}'\mathbf{y} - \mathbf{y}'\,\mathbf{x}\mathbf{b} + \mathbf{b}'\mathbf{x}'\,\mathbf{x}\mathbf{b}$$

$$= \mathbf{y}'\mathbf{y} - 2\mathbf{b}\,\mathbf{x}'\mathbf{y} + \mathbf{b}'\mathbf{x}'\,\mathbf{x}\mathbf{b} \qquad (37)$$

since $\mathbf{b}'\mathbf{x}'\mathbf{y}$ is a (1×1) matrix, or a scalar, and its transpose $(\mathbf{b}'\mathbf{x}'\mathbf{y})' = \mathbf{y}'\mathbf{x}\mathbf{b}$ is the same scalar.

The least squares estimators then must satisfy

$$\left.\frac{\partial S}{\partial \mathbf{b}}\right|_{\hat{\mathbf{b}}} = -2\mathbf{x}'\mathbf{y} + 2\mathbf{x}'\mathbf{x}\hat{\mathbf{b}}' = \mathbf{0}$$

which gives $\mathbf{x}'\mathbf{x}\hat{\mathbf{b}} = \mathbf{x}'\mathbf{y}$ (38)

Eqns. (38) are the least squares normal equations. To solve the normal equations multiply both sides of Eqns. (38) by the inverse of $\mathbf{x}'\mathbf{x}$. Thus, the least squares estimator of $\mathbf{b}$ is

$$\hat{\mathbf{b}} = (\mathbf{x}'\mathbf{x})^{-1}\,\mathbf{x}'\mathbf{y} \qquad (39)$$

provided $(\mathbf{x}'\mathbf{x})^{-1}$ exists. The $(\mathbf{x}'\mathbf{x})^{-1}$ matrix will always exist if the regressors are linearly independent. Using the value of $\hat{\mathbf{b}}$ given in Eqn. (39), the fitted value of $\mathbf{y}$ can be written as

$$\hat{\mathbf{y}} = \mathbf{x}\,\hat{\mathbf{b}}$$

$$= \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\,\mathbf{x}'\mathbf{y}$$

$$= \mathbf{H}\,\mathbf{y}. \tag{40}$$

The $n \times n$ matrix $\mathbf{H} = \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\,\mathbf{x}'$ is called the '**hat**' matrix. The **residual** error vector can be written as

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} \tag{41}$$

$$= \mathbf{y} - \mathbf{x}\hat{\mathbf{b}}$$

$$= \mathbf{y} - \mathbf{H}\,\mathbf{y} = (1 - \mathbf{H})\,\mathbf{y}. \tag{42}$$

We shall now illustrate this concept through an example.

Example 5: Find a linear regression equation that best fit the data given in Table 9.

Table 9

x_1	0	2	2	2	4	4	4	6	6	6	8	8
x_2	2	6	7	5	9	8	7	10	11	9	15	13
y	2	3	2	7	6	8	10	7	8	12	11	14

Solution: Computations for $\hat{\mathbf{b}}$ are as follows:

$$\mathbf{y} = \begin{pmatrix} 2 \\ 3 \\ 2 \\ 7 \\ 6 \\ 8 \\ 10 \\ 7 \\ 8 \\ 12 \\ 11 \\ 14 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} 1 & 0 & 2 \\ 1 & 2 & 6 \\ 1 & 2 & 7 \\ 1 & 2 & 5 \\ 1 & 4 & 9 \\ 1 & 4 & 8 \\ 1 & 4 & 7 \\ 1 & 6 & 10 \\ 1 & 6 & 11 \\ 1 & 6 & 9 \\ 1 & 8 & 15 \\ 1 & 8 & 13 \end{pmatrix}, \quad \mathbf{x}'\mathbf{x} = \begin{pmatrix} 12 & 52 & 102 \\ 52 & 395 & 536 \\ 102 & 536 & 1004 \end{pmatrix}$$

$$\mathbf{x}'\mathbf{y} = \begin{pmatrix} 90 \\ 482 \\ 872 \end{pmatrix}, \quad (\mathbf{x}'\mathbf{x})^{-1} = \begin{pmatrix} 0.975 & 0.243 & -0.229 \\ 0.243 & 0.162 & -0.111 \\ -0.229 & -0.111 & 0.084 \end{pmatrix}$$

$$\hat{\mathbf{b}} = (\mathbf{x}'\mathbf{x})^{-1}\,\mathbf{x}'\mathbf{y} = \begin{pmatrix} 5.375 \\ 3.012 \\ -1.286 \end{pmatrix}.$$

Thus the least squares fit is

$$\hat{y} = 5.375 + 3.012x_1 - 1.286x_2 .$$

***

In addition to estimating the regression coefficients we now obtain an estimate of the variance of error for the multiple regression model.

Estimation of σ^2

As in simple linear regression, we obtain an estimator of σ^2 from the **residual sum of squares**

$$SS_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

$$= \mathbf{e}'\mathbf{e}. \qquad (43)$$

Substituting $\mathbf{e} = \mathbf{y} - \mathbf{x}\hat{\mathbf{b}}$ in Eqn. (43), we get

$$SS_e = (\mathbf{y} - \mathbf{x}\hat{\mathbf{b}})'\,(\mathbf{y} - \mathbf{x}\,\hat{\mathbf{b}})$$

$$= \mathbf{y}'\mathbf{y} - \hat{\mathbf{b}}'\mathbf{x}'\mathbf{y} - \mathbf{y}'\mathbf{x}\hat{\mathbf{b}} + \hat{\mathbf{b}}'\mathbf{x}'\mathbf{x}\hat{\mathbf{b}}$$

$$= \mathbf{y}'\mathbf{y} - 2\hat{\mathbf{b}}'\mathbf{x}'\mathbf{y} + \hat{\mathbf{b}}'\mathbf{x}'\mathbf{x}\hat{\mathbf{b}}. \qquad (44)$$

Since $\mathbf{x}'\mathbf{x}\hat{\mathbf{b}} = \mathbf{x}'\mathbf{y}$, Eqn. (44) reduces to

$$SS_e = \mathbf{y}'\mathbf{y} - \hat{\mathbf{b}}'\mathbf{x}'\mathbf{y}. \qquad (45)$$

The residual sum of squares has $n - (k+1)$ degrees of freedom associated with it since $(k+1)$ parameters are already estimated in the regression model.

Thus, an estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{SS_E}{n-(k+1)} = MS_e, \text{ the } \textbf{residual mean square}. \qquad (46)$$

Example 6: Estimate the error variance σ^2 for the multiple regression model fit to the data in Example 5.

Solution: For the data in Example 5, we have

$$SS_e = \mathbf{y}'\mathbf{y} - \hat{\mathbf{b}}'\mathbf{x}\mathbf{y}$$

$$= 840 - (5.375,\ 3.012, -1.286)\begin{pmatrix} 90 \\ 482 \\ 872 \end{pmatrix}$$

$$= 840 - 814.541 = 25.459$$

$$\hat{\sigma}^2 = \frac{SS_e}{n-k-1} = \frac{25.459}{12-3} = 2.829.$$

***

Just as in the case of simple linear regressions model adequacy can be evaluated in the case of multiple regression models also.

Coefficient of Multiple Determinations

The coefficient of multiple determination R^2 is defined exactly the same way as in Eqn. (22) i.e.,

$$R^2 = \frac{SS_R}{S_{yy}} = 1 - \frac{SS_e}{S_{yy}}.$$

R^2 in this case is a measure of the reduction in the variability of y obtained by using the regressor variables $x_1, x_2, \ldots, x_k$. As in the simple linear regression case, we must have $0 \le R^2 \le 1$. However, a large value of R^2 does not necessarily imply that the regression model is a good one. Adding a regressor to the model will always increase R^2, regardless of whether or not the additional regressor contributes to the model. Thus, it is possible for models that have large values of R^2 to perform poorly in prediction or estimation.

Limitations

All the limitations mentioned in Sec. 3.3 apply to multiple linear regression models also.

The linear regression model $\mathbf{y} = \mathbf{xb} + \mathbf{e}$ is a general model for fitting any relationship that is linear in the unknown parameters $\mathbf{b}$. This includes the important class of polynomial regression models. We now introduce you to polynomial regression models in **one variable**.

Polynomial Regression Models in One Variable

Polynomial models in one variable are models of the form

$$y = b_0 + b_1 x + b_2 x^2 + e \tag{47}$$

or, $$y = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + e \tag{48}$$

and so on. Model (47) is a quadratic model and model (48) is a cubic curve. In general the k^{th} order polynomial model in one variable is

$$y = b_0 + b_1 x + b_2 x^2 + \cdots + b_k x^k + e. \tag{49}$$

Unlike exponential models, polynomial models are non-monotonic. A quadratic curve has one inflection point and a cubic curve has two inflection points. Thus, a quadratic model is useful for cases where increasing the explanatory variable has one effect up to a point, and the opposite effect thereafter. For example, in a chemical experiment, increasing the reaction temperature might increase the yield up to some optimal temperature, after which a further increase in temperature would cause the yield to drop off. In an agricultural experiment, increasing fertilizer will probably increase plant growth up to a point, after which more fertilizer will cause a chemical imbalance in the soil and may hinder plant growth. There are examples where

quadratic and cubic models can be useful. However, it is usually difficult to justify the use of a high-order polynomial because, generally, no physical explanation justifies the large number of inflection points. Thus, such models are seldom used. Moreover, there are several problems that arise when fitting a high-order polynomial in one variable. We shall not be going into those details here. Here we shall only concentrate on quadratic polynomial and illustrate its best fit for a given data through an example.

If we let $x_1 = x$ and $x_2 = x^2$ in the model given by Eqn. (47) then it reduces to

$$y = b_0 + b_1 x_1 + b_2 x_2 + e \qquad (50)$$

which is a multiple linear regression model with two regressor variables and can be treated by the method discussed in previous section. Let us consider an example to understand the procedure.

Example 7: The data given in Table 13 is obtained by a process engineer while studying the relationship between a vacuum setting and particle size distribution for a granular product.

Table 13

Vacuum setting = x	18	18	20	20	22	22	24	24	26	26
Particle size = y	4.0	4.2	5.6	6.1	6.5	6.8	5.4	5.6	3.3	3.6

Fit a linear regression model to the data.

Solution: A scatter plot of the given data is shown in Fig. 9. It appears that the data has a somewhat parabolic curvature

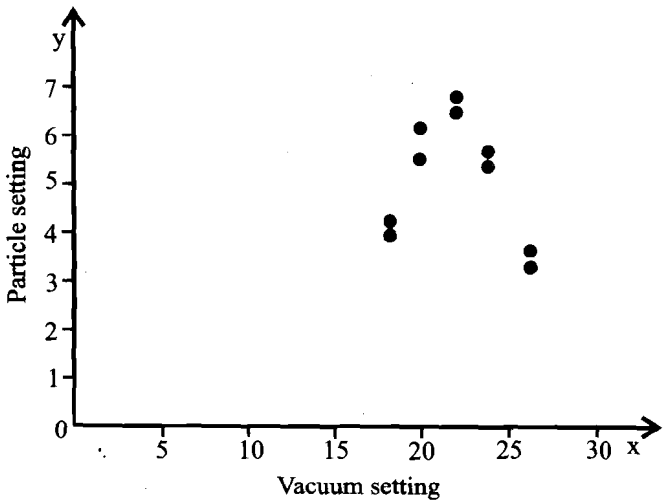


Fig. 9: A scatter plot of the data in Example 7.

We fit the model

$$y = b_0 + b_1 x + b_2 x^2 + e$$

to the given data. Using $x_1 = x$ and $x_2 = x^2$, we can transform the above model to

$$y = b_0 + b_1 x_1 + b_2 x_2 + e.$$

We now use the method used in Example 5 to estimate the regression coefficients.

$$\mathbf{y} = \begin{pmatrix} 4.0 \\ 4.2 \\ 5.6 \\ 6.1 \\ 6.5 \\ 6.8 \\ 5.4 \\ 5.6 \\ 3.3 \\ 3.6 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} 1 & 18 & 324 \\ 1 & 18 & 324 \\ 1 & 20 & 400 \\ 1 & 20 & 400 \\ 1 & 22 & 484 \\ 1 & 22 & 484 \\ 1 & 24 & 576 \\ 1 & 24 & 576 \\ 1 & 26 & 676 \\ 1 & 26 & 676 \end{pmatrix}, \quad \mathbf{x'x} = \begin{pmatrix} 10 & 220 & 4920 \\ 220 & 4920 & 111760 \\ 4920 & 111760 & 2575968 \end{pmatrix}$$

$$\mathbf{x'y} = \begin{pmatrix} 51.10 \\ 1117.60 \\ 24774.40 \end{pmatrix}$$

$$(\mathbf{x'x})^{-1} = \begin{pmatrix} \frac{5119}{10} & \frac{-1181}{40} & \frac{17}{16} \\ \frac{-1881}{40} & \frac{2427}{560} & \frac{-11}{12} \\ \frac{17}{16} & \frac{-11}{112} & \frac{1}{448} \end{pmatrix}$$

$$\hat{\mathbf{b}} = (\mathbf{x'x})^{-1}\ \mathbf{x'y} = \begin{pmatrix} -74.25 \\ 7.42 \\ -0.17 \end{pmatrix}.$$

Thus, the least squares fit is

$$\hat{y} = -74.25 + 7.42x - 0.17x^2.$$

The above quadratic curve has a maximum, which can be obtained as

$$\frac{d\hat{y}}{dx} = 7.42 - 0.34x = 0$$

$$\Rightarrow \quad x = 21.82.$$

Corresponding to $x = 21.82$ the predicted value of y is

$$\hat{y} = -74.25 + 7.42(21.82) - 0.17(21.82)^2$$

$$= 6.5.$$

Thus, to maximize the particle size in the product the best setting for the vacuum is at 21.82.

***

You may now try the following exercises.

E9) The yearly fluctuations in the groundwater table is believed to be dependent on the annual rainfall and the volume of water pumped out from the basin. The data collected on these variables for a period of 10 years is given in Table 10.

Table 10

Year	Water table (in cm.)	Annual rainfall (in cm.)	Groundwater volume pumped out (in cm^3.)
1971	26.9	70.7	4.1
1972	−35.8	42.7	5.2
1973	1.6	51.3	4.4
1974	−15.4	58.0	5.6
1975	31.0	45.3	3.9
1976	58.8	74.7	4.3
1977	−0.4	62.7	5.9
1978	−27.1	46.7	5.2
1979	0.7	79.3	4.4
1980	−4.9	34.0	6.4

Establish the multiple linear regression equation among the variables.

E10) The mileage performance of a car is considered to be dependent on the engine displacement and the horsepower of the engine. The data collected on these variables for 17 cars is given in Table 11. Find a linear regression equation that best fit the data. Also obtain the error variance of the best fit.

Table 11

Car	Miles per Gallon	Horsepower	Engine Displacement
1	27	130	112
2	28	81	90
3	30	93	135
4	28	113	97
5	31	90	114
6	33	63	81
7	40	55	61
8	30	102	97
9	33	92	91
10	32	81	90
11	29	103	113
12	31	90	97
13	33	74	98
14	35	73	73
15	29	102	97
16	32	78	89.
17	28	100	109

E11) Table 12 presents data concerning the strength of kraft paper and the percentage of hardwood in the batch of pulp from which the paper was produced.

Table 12

Hardwood concentration % = (x)	Tensile strength = y
1	6.3
1.5	11.1
2	20.0
3	24.0
4	26.1
4.5	30.0
5	33.8
5.5	34.0
6	38.1
6.5	39.9
7	42.0
8	46.1
9	53.1
10	52.0
11	52.5
12	48.0
13	42.8
14	27.8
15	21.9

Find a linear regression model that best fit the data.

E12) The sale price of a holiday cottage depends on the age and livable area of the cottage. Find a linear regression model that best fit the data given in Table 13. Also find the residual and the residual mean square for the data.

Table 13

Price in thousand rupees	Age (years)	Area (m^2)
745	36	66
895	37	68
442	47	64
440	32	53
1598	1	101

We now end this unit by giving a summary of what we have covered in it.

3.5 SUMMARY

In this unit, we have covered the following points.

1. **Data visualization** is the study of the visual representation of data. It provides information abstracted in some schematic form.

2. In statistics, regression analysis is a technique in which **regression models** are formulated to analyse numerical data consisting of values of a dependent variable (also called response variable) and of one or more independent variables (also known as explanatory variables or predictors).

3. Regression models involving only one predictor are called **simple regression models** and those involving two or more predictors are called **multiple regression models.**

4. Regression models are **linear** if they are linear in the parameters involved in the model.

5. Method of **least squares** can be used to fit a line or a curve to two-dimensional data and a plane or curved surface to three-dimensional data.

6. Second order **polynomial regression** models in **one-variable** fit a quadratic curve to a given set of data which can be converted to multiple regression model by using the transformation of the independent variables.

3.6 SOLUTIONS/ANSWERS

E1) You can discuss scatter plot, histogram, bar diagram, box plot, block plot, etc. You can illustrate them by considering a type of data for which each of them is most suited. For example, in case of continuous data histogram is preferred whereas, discrete data is commonly represented by a bar diagram.

E2) Cartographs in weather bulletins are used to compare data on a geographical basis. Think of other similar examples.

E3) a) $$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}(x_i^2+\bar{x}^2-2x_i\,\bar{x})$$

$$=\sum_{i=1}^{n}x_i^2+n\,\bar{x}^2-2n\,\bar{x}^2$$

$$=\sum_{i=1}^{n}x_i^2-n\bar{x}^2.$$

b) Do as in a) above.

E4) $\hat{b}_0=93.727$, $\hat{b}_1=-0.638$, $\hat{y}=93.727-0.638x$, where x is price per litre and y is recovery.

E5) $\hat{y}=538.155-0.268\,x$ when $x=1985$, $\hat{y}=6.2$

E6) $\hat{y}=1073+.8726x$

$R^2=0.8288$

Plotting the regression equation and 18 points, it is apparent that the slope $\hat{b}_1$ is the rate of change of $\hat{y}$ as x varies and the intercept $\hat{b}_0$ is the value of $\hat{y}$ at $x = 0$. But the linear trend does not establish cause and effect between homework and test results.

E7) Draw a scatter plot of the data and see that a line is not an appropriate model. Transform the data by taking log of y's and obtain

x	4.99	1.96	2.98	9.00	4.04	6.06	0.88	8.02	6.97
$y_1 = \ln y$	22.19	9.89	11.00	40.46	18.93	25.00	5.19	34.02	28.93

Check that the scatter plot of the above data looks linear.
The best fit linear regression line is

$$\hat{y}_1 = 0.73 + 4.209\ x$$

or, $\hat{y} = e^{0.73+4.209x} = 2.075\, e^{4.209x}$

E8) Taking log of x's the data transforms to

$x_1 = \ln x$	9.74	3.85	5.01	6.91	2.64	8.37	5.30	2.89
y	45	15	18	25	9	32	20	10

The scatter plot of the above data indicates that linear model is appropriate for the data. The best fit regression line is

$$\hat{y} = -4.1 + 4.62\ x_1$$
$$= -4.1 + 4.62\ \ln x.$$

E9) $\hat{y} = 47.5 + 0.56\ x_1 - 15.3\ x_2$, where x_1 is annual rainfall, x_2 is ground water volume and y is water table in cms.

E10) $\hat{y} = 46.429 - 0.109x_1 - 0.058x_2$. Error variance $= \dfrac{SS_e}{14}$, where x_1 is horsepower, x_2 is engine displacement and y is miles per gallon.

E11) Draw the scatter plot of the given data and check that the model to be fitted is

$$y = b_0 + b_1 x + b_2 x^2$$

Put $x_1 = x$ and $x_2 = x^2$ in the above model and proceed exactly as in Example 7 to obtain $\hat{y} = -6.6959 + 11.7703x - 0.635x^2$.

E12) $$(\mathbf{x}'\mathbf{x})^{-1} = \begin{pmatrix} 27.531 & -0.266 & -0.273 \\ -0.266 & 0.003 & 0.002 \\ -0.273 & 0.002 & 0.003 \end{pmatrix}$$

$$\mathbf{x}'\mathbf{y} = \begin{pmatrix} 4120 \\ 96387 \\ 323036 \end{pmatrix}, \quad \hat{\mathbf{b}} = \begin{pmatrix} -281.427 \\ -7.611 \\ 19.01 \end{pmatrix}$$

$$\hat{y} = -281.427 - 7.611x_1 + 19.01x_2$$

$$\text{residual} = \mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{x}\hat{\mathbf{b}}$$

$$\text{residual mean squares} = \frac{\mathbf{e}'\mathbf{e}}{2}.$$

—x—

PRACTICAL EXERCISES

Sessions 1 and 2

1. An electric utility is interested in developing a model relating peak hour demand (y) to total energy usage during the month (x). Data for 53 residential customers for the month of August, 2005 are shown in Table 1.

Table 1

Customer	x(KWH)	y(KW)	Customer	x(KWH)	y(KW)
1	679	.79	27	837	4.20
2	292	.44	28	1748	4.88
3	1012	.56	29	1381	3.48
4	493	.79	30	1428	7.58
5	582	2.70	31	1255	2.63
6	1156	3.64	32	1777	4.99
7	997	4.73	33	370	.59
8	2189	9.50	34	2316	8.19
9	1097	5.34	35	1130	4.79
10	2078	6.85	36	463	.51
11	1818	5.84	37	770	1.74
12	1700	5.21	38	724	4.10
13	747	3.25	39	808	3.94
14	2030	4.43	40	790	.96
15	1643	3.16	41	783	3.29
16	414	.50	42	406	.44
17	354	.17	43	1242	3.24
18	1276	1.88	44	658	2.14
19	745	.77	45	1746	5.71
20	435	1.39	46	468	.64
21	540	.56	47	1114	1.90
22	874	1.56	48	413	.51
23	1543	5.28	49	1787	8.33
24	1029	.64	50	3560	14.94
25	710	4.00	51	1495	5.11
26	1434	.31	52	2221	3.85
			53	1526	3.93

Write a program in C-language to

a) find a linear model that best fit the data using the least squares estimates and estimate the peak hour demand for a total of 2050 KWH energy usage.

b) find R^2 for the model.

2. A manufacturer of particle boards is interested in the strength of particle boards as a function of the baking temperature. The data from an experiment designed to study this relation are given in Table 2.

Table 2

Strength	Temp.	Strength	Temp.
66.30	40	75.78	55
64.84	40	72.57	55
64.36	40	76.64	55
69.70	45	78.87	60
66.26	45	77.37	60
72.06	45	75.94	60
73.23	50	78.82	65
71.40	50	77.13	65
68.85	50	77.09	65

Write a programme in C-language to

a) obtain a best fit line using the least squares estimates.

b) fit a quadratic model using the least squares estimates.

c) compare the models in a) and b) above by finding R^2 for each of them.

3. When gasoline is pumped into the tank of a car, vapors are vented into the atmosphere. An experiment was conducted to determine whether y, the amount of vapour, can be predicted using the following four variables based on initial conditions of the tank and the dispensed gasoline:

x_1 = tank temperature

x_2 = gasoline temperature

x_3 = vapour pressure in tank

x_4 = vapour pressure of gasoline

The data are given in Table 3.

Table 3

y	x_1	x_2	x_3	x_4	y	x_1	x_2	x_3	x_4
29	33	53	3.32	3.42	40	90	64	7.32	6.70
24	31	36	3.10	3.26	46	90	60	7.32	7.20
26	33	51	3.18	3.18	55	92	92	7.45	7.45
22	37	51	3.39	3.08	52	91	92	7.27	7.26
27	36	54	3.20	3.41	29	61	62	3.91	4.08
21	35	35	3.03	3.03	22	59	42	3.75	3.45
33	59	56	4.78	4.57	31	88	65	6.48	5.80
34	60	60	4.72	4.72	45	91	89	6.70	6.60
32	59	60	4.60	4.41	37	63	62	4.30	4.30
34	60	60	4.53	4.53	37	60	61	4.02	4.10
20	34	35	2.90	2.95	33	60	62	4.02	3.89
36	60	59	4.40	4.36	27	59	62	3.98	4.02
34	60	62	4.31	4.42	34	59	62	4.39	4.53
23	60	36	4.27	3.94	19	37	35	2.75	2.64
24	62	38	4.41	3.49	16	35	35	2.59	2.59
32	62	61	4.39	4.39	22	37	37	2.73	2.59

Write a program in C-language to

a) fit a linear regression model to the data using the least squares estimates.

b) find the 'hat' matrix.

c) find the residual error.

d) find R^2 for the model.