14 SIMPLE LINEAR REGRESSION AND CORRELATION

14.1 RELATIONS BETWEEN VARIABLES

The concept of a relation between two variables such as family incomes and family expenditures for housing, is a familiar one. We now distinguish between a *functional relation* and a *statistical relation*, and consider each of them in turn.

14.1.1 Functional Relation between Two Variables. A *functional relation* between two variables, is a perfect relation, where the value of the dependent variable is uniquely determined from the value of the independent variable. A functional relation is expressed by a mathematical formula. If x is the *independent* variable and y is the *dependent* variable, a functional relation is of the form

$$y = f(x)$$

Given a particular value of x, the function $f(x)$ gives the corresponding value of y. The observations, when plotted on a graph, all fall directly on the line or curve of the functional relationship. This is the main characteristic of all functional relationships.

14.1.2 Statistical Relation between Two Variables. A *statistical relation* is a relation where the value of the dependent variable is not uniquely determined when the level of the independent variable is specified. A statistical relation, unlike a functional relation, is not exact. The value of y is not uniquely determined from knowledge of x. The observations, when plotted on a graph, do not fall directly on the line or curve of the relationship. This is the main characteristic of all statistical relationships.

In many fields such as business, economics and administration exact relations are not generally observed among the variables, but rather statistical relationships prevail. For example: (*i*) The grade point Y secured by a student in the college is undoubtedly related to his grade point x secured in the school. (*ii*) The consumption expenditure Y of a household is related to its income x. (*iii*) The maintenance cost Y per year for an automobile is related to his age x. (*iv*) The yield Y of wheat is related to the quantity x of a fertilizer. (*v*) The amount of sales Y of a newly produced item may be related to its advertising cost x. (*vi*) The weight Y of a baby is certainly related to his age x. (*vii*) The saving Y of a person or a firm is related to his/its income. (*viii*) The height Y of a son is undoubtedly related to the height x of his father, *etc.*

Causal Relation. Another factor to consider is whether a causal relationship exists between two variables. In the example of steel output and labour input, it is clear that a causal relationship does exist. The number of workers will influence the number of tons of steel produced. There is also a causal relationship between hours of sunshine and the rate of growth of tulips. Conversely, it is less clear that more steel output will cause a rise in the number of workers, nor will we make the sunshine more by forcing tulips to grow faster. It is important to note regression analysis and correlation analysis make no assertions about causality.

14.2 REGRESSION ANALYSIS

One of the most common and important tasks that statisticians must face is to determine the existence and nature of relationships between variables in a problem. We are interested in relationships between variables because we may often possess information about some variables and wish to use that information to draw conclusions about another variable. In many situations, we face the problems that involve two or more variables and we are to make inferences about how the changes in one variable are related to the changes in other variables, and how one set of variables is considered to predict or account for the other variable. These problems can be dealt with measuring statistical relationships between variables, representing the relationships in mathematical (functional) form and evaluating the significance of the relationships.

The *regression analysis* provides a method of estimating an average relationship (often linear) between two or more variables, which allows the investigator to explain and predict and this is, in a sense, the best possible approximation. The regression analysis provides an equation that can be used for estimating the average value of one variable from given values of other variables.

14.2.1 Simple Regression. The *simple regression* is a relationship that describes the dependence of the expected value of the dependent random variable for a given value of the independent non-random variable. In statistical relationships, if only two values are involved:

Regressor. The variable, that forms the basis of estimation or prediction, is called the regressor. It is also called as the predictor variable or independent variable or controlled variable or explanatory variable. It is usually denoted by x.

Regressand. The variable, whose resulting value depends upon the selected value of the independent variable, is called the regressand. It is also called as the response variable or the predictand variable or dependent variable or explained variable. It is usually denoted by Y.

The values of the independent variable x are determined by the experimenter and they are fixed in advance. They are arbitrarily selected constants and thus have no error attached with them. The independent variable is not random but a mathematical variable and we can choose the values we give to it. On the other hand, however, the problem is usually complicated by the fact that the dependent variable is subject to experimental variation or scatter. Besides depending upon the regressor variable, there is a random error in determining the response variable. Thus the response variable possesses a random character, it is left free to take on any value that may be possibly associated to a given value of the independent variable.

Let Y be the response variable and x be the regressor variable and $\mu_{Y|x} = E(Y|x)$ be the expected value of the distribution of the random variable Y for a given value of the non-random variable x, then the simple regression is given by

$$\mu_{Y|x} = f(x)$$

where $f(x)$ is a function that describes the relationship between the regressor x and the response Y and $f(x)$ may be of linear, quadratic, exponential, geometric, or any other form.

14.2.2 Regression Function. When we look for a relationship $\mu_{Y|x} = f(x)$, where the function $f(x)$ is to be determined, *i. e.*, given the points only we have to 'work backwards' or regress to the original function $f(x)$. Hence this function is called *regression function.*

14.2.3 Regression Curve. The *regression curve* is the locus (a continuous set of points) of the expected value of the response variable for given values of the regressor variable. If several measurements are made on the response variable Y at the same value of the regressor variable x, then the results will form a distribution. The curve which joins the expected values of these distributions for different values of x is called the simple regression curve of Y on x.

14.3 CURVE FITTING

Curve fitting is a process of estimating, from an observed sample, the parameters of the population regression function of a response variable on a regressor variable.

14.3.1 Least Squares Principle. The *principle of least squares* says that the sum of squares of the residuals of observed values from their corresponding estimated values should be the least possible. This principle was given by a French mathematician Adrien Legendre.

14.3.2 Least Squares Fit. Among all the curves approximating a given data, the curve is called a *least squares fit* for which the sum of squares of the residuals of the observed values from their corresponding estimated values is the least.

For a given set of observed data, different curves have different values of the sum of squares of the residuals. The best fitting curve is the one having the smallest possible value of the sum of squares of the residuals. To avoid the personal bias in fitting a curve to observed data, the method of least squares is used.

14.3.3 Scatter Diagram. The *scatter diagram* is a set of points in a rectangular co-ordinate system (with x measured horizontally and y measured vertically), where each point represents an observed pair of values. To aid in determining an equation connecting the two variables, a first step is the collection of the data showing the paired values of the variables under consideration.

Let us suppose that n pairs of observations (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n) are made on two variables. The next step in the investigation is to plot the data on a graph to get a scatter diagram. The choice of the regression curve to fit may be influenced by theory, by experience or simply by looking at the scatter diagram. For example, the experiment may have been designed to verify a particular relationship between the variables. Alternately, the function form may be selected after inspecting the scatter diagram, as it would be useless to try to fit a straight line to some data if the relationship was clearly curvilinear. In practice the experimenter may choose the one which gives the best fit.

It is often possible to see, by looking at the scatter diagram that a smooth curve can be fitted to the data. In particular, if a straight line can be fitted to the data, then we say that a linear relationship exists between two variables, otherwise the relationship is curvilinear. A visual examination of a scatter diagram gives some useful indications of the nature and strength of the relationship between two variables and aids in choosing the appropriate type of model for estimation.

For example, if the points on the scatter diagram tend to run from the lower left side to the upper right side (that is, if the Y variable tends to increase as x increases), there is said to be a *direct* relationship between the two variables. On the other hand if the points on the scatter diagram tend to run from the upper left side to the lower right side (that is, if the variable Y tends to decrease as x increases), there is said to be *inverse* relationship between the two variables. The scatter diagram gives an indication whether a straight line appears to be an adequate

description of the average relationship between two variables. If a straight line is used to describe the average relationship between two variables, a linear relationship is said to exist. If the points on the scatter diagram appear to lie along a curve, a curvilinear relationship is said to be present.

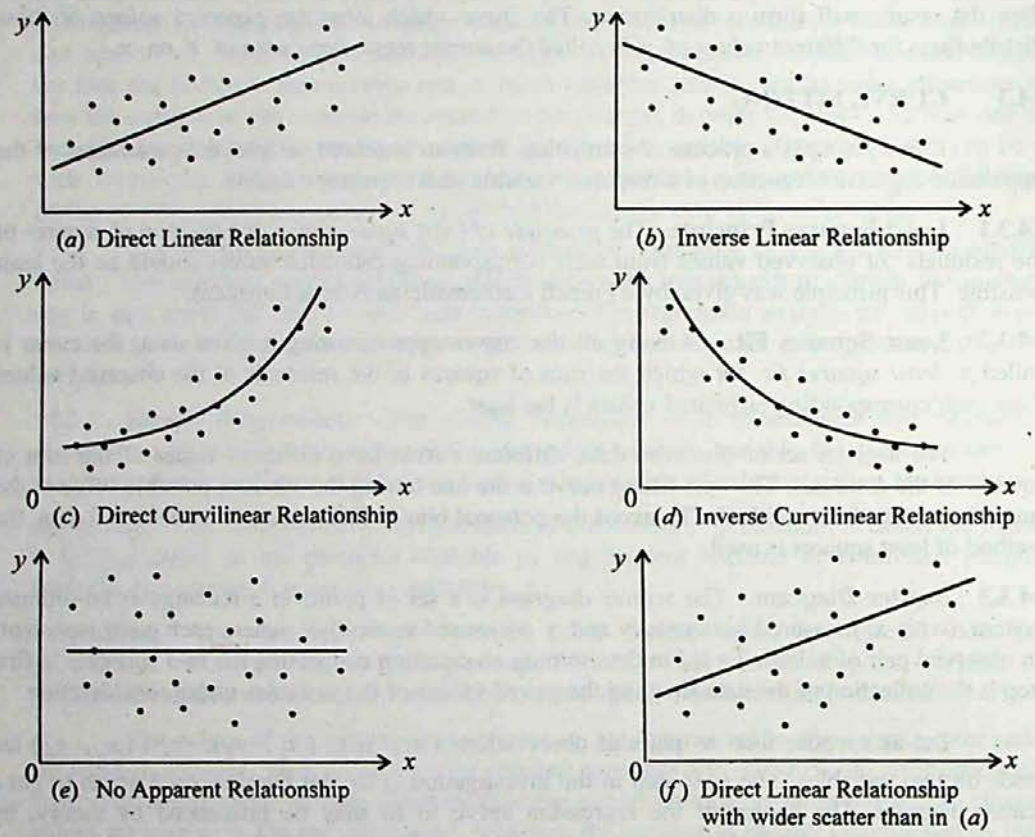

Fig. 14.1 Types of relationships found in scatter diagrams

Parts (*a*), (*b*), (*c*) and (*d*) of Fig 14.1 show direct linear, inverse linear, direct curvilinear and inverse curvilinear relationships. The points tend to follow a straight line with positive slope in (*a*), a straight line with negative slope in (*b*), a curve with positive slope in (*c*), and a curve with negative slope in (*d*). Of course the relationships are not always so obvious.

In (*e*) the points appear to follow a horizontal line. This type of scatter diagram depicts "no correlation" or no evident relationship between x and Y variables because the horizontal line implies no change, on the average, as x increases. In (*f*) the points follow a straight line with positive slope as in (*a*) but there is a much wider scatter of points around the line than in (*a*).

Note that a scatter diagram is primarily used to determine the appropriateness of a particular type of equation for describing the data. The approximate "goodness of fit" of the equation is also apparent from a scatter diagram, for example the fit in (*a*) is quite good as compared to the fit in (*f*). However, "goodness of fit" can and should be defined and determined precisely.

14.4 SIMPLE LINEAR REGRESSION

If the simple regression describes the dependence of the expected value of the dependent random variable Y as a linear function of the independent non-random variable x, then the regression is called *simple linear regression.* It is given by

$$\mu_{Y|x} = \alpha + \beta x$$

which implies that $\alpha = \mu_{Y|x}$ when $x = 0$. Thus α is the intercept of the line along *y-axis.* The β indicates the change in the mean of the probability distribution of Y per unit increase in x.

14.4.1 Simple Linear Regression Coefficient. The *simple linear regression coefficient* is the relative change in the expected value of the dependent random variable with respect to a unit increase in the independent non-random variable. It is denoted by β. The slope of the line β remains constant at each value of x.

It is measured by $\tan\theta$ where θ is the angle made by the line with the positive side of the *x-axis.* The slope of the line depends upon the value of the β. If the value of β is positive, the line will slope upward like the solid line in the Fig. 14.2. On the other hand, if the value of β is negative, the line will slope downward like the broken line in the Fig. 14.2.

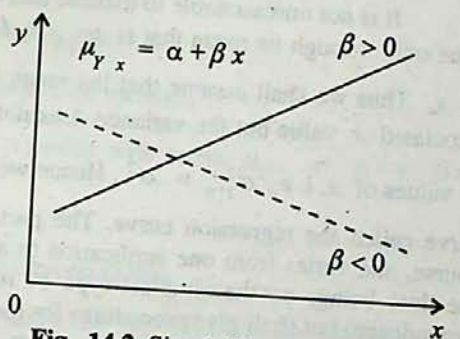

Fig. 14.2 Simple linear regression

14.5 THE SIMPLE LINEAR REGRESSION MODEL

We used the scatter diagram to illustrate the problem of selecting the regression line that provides the "best" estimate of the relationship between the independent and dependent variables in a regression problem. The next step is to specify the mathematical formulation of the linear regression model to provide a basis for statistical analysis.

14.5.1 An Example of Regression Model. In this section we will give an example to illustrate how the regression model between two variables can be used to simulate real world problems.

Suppose that we are to investigate the relationship between the consumption expenditure and the disposable income of households in a certain city for some given period of time. We know that as one's income increases, there is a tendency to spend more. What kind of relation is there between income and expenditure? Is it proportional, or is there any other form of a relationship, how close this relationship between income and expenditure is? Certainly there is no functional relationship between disposable income and consumption expenditure. Now let Y denote the consumption expenditure and x denote the disposable income.

Let us suppose that we have already divided our households into various groups on the basis of income levels. We do not expect that all the households within the group which have some given (fixed, predetermined) income x will display an identical expenditure. Some will spend more than the others, some will spend less, but we do expect a clustering of the expenditure figures around a central value with some variance. For each possible value of x chosen non-randomly there are several values of Y that could occur. Thus Y becomes a random

variable that possesses a distribution or population of associated y values for any given value of x. This distribution of associated y values, for any given x, is described either by a probability density function $f_Y(y|x)$ or by a probability mass function $p_Y(y|x)$ if the population has a discrete set of possible values. This distribution represents the relative likelihood of different values of Y occurring.

The mean of each probability distribution of Y values varies in some constant and systematic manner with the independent variable x. The mean of any distribution of Y for given x will be denoted by $\mu_{Y|x} = E(Y|x)$ and the variance of this distribution by $\sigma^2_{Y|x} = Var(Y|x)$. These are unknown parameters. They are constant for any fixed value of x but may vary between the distribution of Y_i for different x_i. The mean of Y for all values will be dented by μ_Y and the variance by σ_Y^2 or σ^2.

It is not unreasonable to assume that the random variable Y depends on the associated x value only through its mean that is $\mu_{Y|x} = f(x)$, but any higher moments of Y do not depend on x. Thus we shall assume that the mean value of the random variable Y depends upon the associated x value but the variance does not. We shall further assume that $\sigma^2_{Y|x}$ is constant for all values of x, *i. e.*, $\sigma^2_{Y|x} = \sigma^2$. Hence we assume that all the means $\mu_{Y|x}$ lie on a continuous curve called the regression curve. The particular form of the regression curve is arbitrary, of course, and varies from one application to another. We shall only concentrate our attention, for the time being, on the simplest type of regression curve, namely, the straight line (a linear dependence) but shall give procedures for more general models.

14.5.2 Mathematical Formulation of Regression Model. Our observed paired values of x and Y are only sample values from a large population. However, for a moment we are concerned with constructing a model for the population of all possible paired values. If, for example, a linear relationship is considered to be appropriate, that is, the average relationship between the dependent random variable Y and the independently varying non-random variable x is assumed to be linear. Since we are interested in the conditional expectation $\mu_{Y|x} = E(Y|x)$. By assuming that Y and x are linearly related, we are saying that all possible conditional means $\mu_{Y|x}$ which might be calculated one for each possible value of x must lie on a single straight line. This line is called the population regression line. To specify this line we need to know its slope and intercept. Let α be the y-intercept and β be the slope of the line. The population regression line is written as follows:

$$\mu_{Y|x} = \alpha + \beta x$$

This line is unknown. When some exact value of x is specified from its domain, it is customary to denote this value as x_i. Associated with this value x_i of the independent non-random variable x, there exists a random variable Y_i with a distribution or population with mean $\mu_{Y|x_i}$ and variance $\sigma^2_{Y|x_i}$. Assume that

$$\mu_{Y|x_i} = \alpha + \beta x_i, \qquad \sigma^2_{Y|x_i} = E(Y_i - \mu_{Y|x_i})^2 = \sigma^2$$

Now we define a deviation of the random variable Y_i from its unknown mean $\mu_{Y|x_i}$ and call this deviation as population regression error. For this reason, this difference is usually called the random "error" and denoted by ε_i. Therefore, we can define the random variable ε_i as

$$\varepsilon_i = Y_i - \mu_{Y|x_i} = Y_i - (\alpha + \beta x_i) = Y_i - \alpha - \beta x_i$$

There are three generally recognized sources of errors in such regression problems: (1) specification (or equation) error, arising from the omission of one or more relevant independent variables; (2) sampling error, arising from random variation of observations around their expected value and (3) measurement error, arising from the lack of precision in measuring variables. These errors are assumed to have zero mean and the constant variance identical to the variance of Y for a given value of x. We can now define what is called the population regression model as

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

where, x_i = a predetermined value of a non-random variable.

Y_i = associated with x_i a random variable with mean $\mu_{Y|x_i} = \alpha + \beta x_i$ and variance $\sigma^2_{Y|x_i} = \sigma^2$.

α = the population y-intercept of the regression line.

β = the population slope of the regression line also known as the population regression coefficient.

ε_i = the deviation $(Y_i - \mu_{Y|x_i})$ in the population.

This model is said to be simple, linear in parameters and linear in independent variable. It is *simple*, in that there is only one independent variable, *linear* in the parameters because no parameter appears as an exponent or is multiplied or divided by another parameter, and *linear* in the independent variable, because the variable appears only in the first power.

14.5.3 The Sample Simple Linear Regression Model. In our population regression model α, β, $\mu_{Y|x}$ and σ^2 are unknown parameters we wish to estimate these parameters statistically on the basis of our sample observations on x and Y, and we may wish to test hypothesis and construct confidence intervals about these parameters. In this regard sampling is accomplished as follows:

(*i*) A set of n values of x in its domain is observed and denoted by $x_1, x_2, \cdots, x_n$. The x's are not random variables, but they may be selected either by some random procedure or by purposeful selection.

(*ii*) Each x_i determines a distribution or population whose mean is $\alpha + \beta x_i$ and whose variance is σ^2. From this distribution a value (a sample of size one) is selected at random and denoted by Y_i.

Thus we have a set of n pairs of observations denoted by $(Y_1, x_1), (Y_2, x_2), \cdots, (Y_n, x_n)$ which we have written to stress the fact that each sample of Y that we take has an associated x value. The values of x may or may not be all distinct, but as we shall see, we must have at least two different values of x represented if we are to estimate both α and β. We can write the n actually observed sample pairs as $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ without incorporating any further assumptions into our model, we can obtain estimates of α, β and $\mu_{Y|x}$. It is customary to let a be the best estimate of α, b be the best estimate of β and to let $\hat{y}$ be the resulting estimate of $\mu_{Y|x}$ and the line resulting a, b and $\hat{y}$ is called the best fitted regression line. This line has the same form as the population line. Thus the sample simple linear regression is

$$\hat{y} = a + bx$$

where $\hat{y}$ = the ordinate of the estimated line for any given value of x which is the best point estimate of $\mu_{Y|x}$

a = the y-intercept of the estimated line which is the best point estimate of α

b = the slope of the estimated line which is the best point estimate of β

Thus, if x_i is a specific value of x, then

$$\hat{y}_i = a + bx_i$$

is the equation for finding $\hat{y}_i$, which is the best estimate of $\mu_{Y|x_i}$ for this value x_i.

We can specify a sample regression model just as we did in the population regression model. Again we need to define an error term, which in this case is the deviation of actual value y_i from predicted value $\hat{y}_i$. This error term is denoted by e_i, which means that the sample regression error e_i is an estimate of the population error ε_i. The errors e_i ($i = 1, 2, \cdots, n$) are often called *residuals* or *deviations* or *prediction errors*.

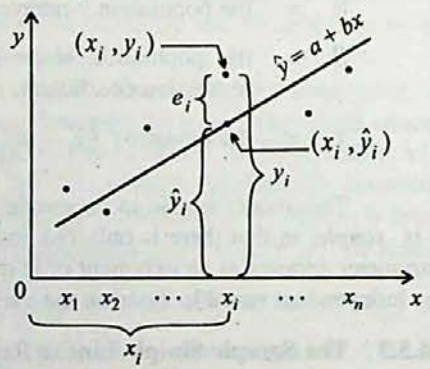

Fig. 14.3 Residual of y_i

Residual: $e_i = y_i - \hat{y}_i = y_i - (a + bx_i) = y_i - a - bx_i$

These residuals from the estimated line will be positive or negative as the actual value lies above or below the line. Thus the sample regression model is

$$y_i = a + bx_i + e_i$$

14.5.4 Covariance of Two Variables. If $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ are n pairs of observations on two variables X and Y, then the *covariance*, denoted by s_{xy}, is defined as

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n} \qquad i = 1, 2, \cdots, n$$

It is a measure of the *linear mutual variability* of the two variables. Its sign reflects the direction of the mutual variability: if the variables tend to move in the same direction, the covariance is positive; if the variables tend to move in opposite directions, the covariance is negative. It can be easily expressed as

$$s_{xy} = \frac{\sum x_i y_i}{n} - \bar{x}\,\bar{y} \qquad i = 1, 2, \cdots, n$$

If $x_1, x_2, \cdots, x_n$ and $y_1, y_2, \cdots, y_n$ are two series of n observations each, and if $z_i = x_i \pm y_i$, then

(*i*) $s_z^2 = s_x^2 + s_y^2 \pm 2 s_{xy}$

(*ii*) $s_z^2 = s_x^2 + s_y^2$ when X, Y are independent variables

14.5.5 Least Squares Point Estimation of α, β and $\mu_{Y|x_i}$ (Fitting of Straight Line). We now face the problem of estimating the linear regression between a dependent random variable Y and an independently varying non-random variable x given a sample of y values with their associated values of x. A general method of estimating the parameters of a regression line is the method of least squares which is explained in the following theorem.

***Theorem* 14.1** *Let $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ be n observed values of a random variable Y, with their associated x values, where the regression line is $\mu_{Y|x} = \alpha + \beta x$.*

(*i*) *The least squares line is given by $\hat{y} = a + b\,x$, where*

$$b = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2} = \frac{\sum x_i y_i - n \bar{x}\,\bar{y}}{\sum x_i^2 - n \bar{x}^2}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}$$

$$a = \frac{\sum y_i - b \sum x_i}{n} = \bar{y} - b\bar{x}$$

(*ii*) *The least squares line always passes through the point of means $(\bar{x}, \bar{y})$.*

(*iii*) *The least squares estimate of $\mu_{Y|x_i}$ is*

$$\hat{y}_i = \bar{y} + b(x_i - \bar{x})$$

***Example* 14.1** *The following sample of 8 grade point averages and marks in matriculation was observed for students from a college.*

Score	480	490	510	510	530	550	610	640
GPA	2.7	2.9	3.3	2.9	3.1	3.0	3.2	3.7

Find the least squares line. Estimate the mean GPA *of students scoring* 600 *marks.*

Solution. The estimated regression line is $\hat{y} = a + bx$

x_i	y_i	x_i^2	$x_i\, y_i$
480	2.7	230400	1296
490	2.9	240100	1421
510	3.3	260100	1683
510	2.9	260100	1479
530	3.1	280900	1643
550	3.0	302500	1650
610	3.2	372100	1952
640	3.7	409600	2368
4320	24.8	2355800	13492

The least squares estimates a and b are

$$b = \frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{n\Sigma x_i^2 - (\Sigma x_i)^2} = \frac{8(13492) - (4320)(24.8)}{8(2355800) - (4320)^2} = 0.00435$$

$$a = \frac{\Sigma y_i - b\Sigma x_i}{n} = \frac{24.8 - 0.00435(4320)}{8} = 0.751$$

The best fitted line is $\hat{y} = 0.751 + 0.00435\, x$

For $x = 600$, we have $\hat{y} = 0.751 + 0.00435\,(600) = 3.361$

14.5.6 Properties of the Least Squares Line. The line fitted by the method of least squares has a number of properties worth noting.

(1) The sum of the residuals is zero, that is

$$\Sigma e_i = 0$$

However, rounding errors may, of course, be present in any particular case. Hence, in minimizing Σe_i^2 the least squares method automatically sets $\Sigma e_i = 0$.

(2) The sum of the observed values y_i equals the sum of the fitted values $\hat{y}_i$

$$\Sigma y_i = \Sigma \hat{y}_i$$

Therefore, it follows that

(*i*) the mean of the fitted values $\hat{y}_i$ is the same as the mean of the observed values y_i

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{\Sigma \hat{y}_i}{n} = \bar{\hat{y}}_i$$

(*ii*) $\Sigma(\hat{y}_i - \bar{y})$ is also equal to zero.

(3) The sum of the squares of the residuals Σe_i^2 is minimum.

$$\Sigma e_i^2 = \Sigma y_i^2 - a\Sigma y_i - b\Sigma x_i\, y_i$$

(4) The regression line always passes through the point of means $(\bar{x}, \bar{y})$, the centre of gravity of the observed data. That is, whenever $x_i = \bar{x}$, we have $\hat{y}_i = \bar{y}$.

***Example* 14.2** *Given the following data*

x_i	0	1	2	3	4
y_i	1.0	1.8	3.3	4.5	6.3

(*a*) *Determine the least squares line taking x as independent variable.*

(*b*) *Find the estimated values for given values of x and show that*

(*i*) $\sum y_i = \sum \hat{y}_i$

(*ii*) $\sum e_i = 0$

(*c*) *Calculate the sum of the squares of the residuals.*

(*d*) *Verify that* $\sum e_i^2 = \sum y_i^2 - a\sum y_i - b\sum x_i\, y_i$.

Solution. (*a*) The estimated regression line is $\hat{y} = a + bx$

x_i	0	1	2	3	4	$\sum x_i = 10$
y_i	1.0	1.8	3.3	4.5	6.3	$\sum y_i = 16.9$
$x_i\, y_i$	0	1.8	6.6	13.5	25.2	$\sum x_i\, y_i = 47.1$
x_i^2	0	1	4	9	16	$\sum x_i^2 = 30$

The least squares estimates a and b are

$$b = \frac{n\sum x_i\, y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2} = \frac{5(47.1) - (10)(16.9)}{5(30) - (10)^2} = 1.33$$

$$a = \frac{\sum y_i - b\sum x_i}{n} = \frac{16.9 - 1.33(10)}{5} = 0.72$$

The best fitted line is $\hat{y} = 0.72 + 1.33x$

(*b*) The estimated values $\hat{y}_i$ for the given values of x and the residuals $e_i = y_i - \hat{y}_i$ are obtained as shown in the following table.

x_i	y_i	$\hat{y}_i = 0.72 + 1.33x_i$	$e_i = y_i - \hat{y}_i$	e_i^2	y_i^2
0	1.0	0.72 + 1.33 (0) = 0.72	0.28	0.0784	1.00
1	1.8	0.72 + 1.33 (1) = 2.05	− 0.25	0.0625	3.24
2	3.3	0.72 + 1.33 (2) = 3.38	− 0.08	0.0064	10.89
3	4.5	0.72 + 1.33 (3) = 4.71	− 0.21	0.0441	20.25
4	6.3	0.72 + 1.33 (4) = 6.04	0.26	0.0676	39.69
Sum	16.9	16.90	0	0.2590	75.07

It is verified that

(*i*) $16.9 = \sum y_i = \sum \hat{y}_i = 16.9$

(*ii*) $\sum e_i = \sum (y_i - \hat{y}_i) = 0$

(*c*) The sum of the squares of the residuals is $\sum e_i^2 = 0.259$

(*d*) We are to verify that $\Sigma e_i^2 = \Sigma y_i^2 - a\Sigma y_i - b\Sigma x_i y_i$

$$0.259 = 75.07 - 0.72(16.9) - 1.33(47.1)$$

$$0.259 = 0.259$$

14.5.7 Coding and Scaling. In many cases the process of coding and scaling by a linear transformation can simplify the job of estimating the regression line or curve.

***Theorem* 14.2** *The sample linear regression coefficient b is independent of change of origin but it is not independent of change of scale.*

Let $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ be n observed values of a random variable Y, with their associated x values, then the sample regression coefficient is

$$b_{yx} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$$

Let $u_i = \dfrac{x_i - p}{h} \Rightarrow x_i = p + hu_i \Rightarrow \bar{x} = p + h\bar{u}$

and $v_i = \dfrac{y_i - q}{k} \Rightarrow y_i = q + kv_i \Rightarrow \bar{y} = q + k\bar{v}$

In this transformation we choose the constants p, q and h, k so that the transformed values u_i and v_i become as simple as possible. Then

$$b_{yx} = \frac{k}{h} b_{vu}$$

A Special Coding and Scaling. If the values $x_1, x_2, \cdots, x_n$ of the independent variable x are equally spaced at an interval h, then calculations involved in solving the normal equations can be made much simple by taking the origin at $\bar{x}$ and choosing a suitable unit of measurement. The choice of origin and unit is explained below in the two cases.

(*i*) If the sample size is odd, say $n = 2m - 1$, then we take the origin at the middle value x_m which is equal to $\bar{x}$, *i. e.*, $\bar{x} = x_m$. If h is the common interval, we take h as a new unit of measurement, then changing each x_i into u_i by a linear transformation $u_i = (x_i - \bar{x})/h$, the variable u takes the values $-(m-1), -(m-2), \cdots, -2, -1, 0, 1, 2, \cdots, (m-2), (m-1)$. Thus, we get

$$\Sigma u_i = 0 = \Sigma u_i^3 = \Sigma u_i^5 = \cdots$$

(*ii*) If the sample size is even, say $n = 2m$, then we take the origin at the average of the two middle values x_m and x_{m+1} which is equal to $\bar{x}$, *i.e.*, $\bar{x} = (x_m + x_{m+1})/2$. If h is the common interval, we take $h/2$ as a new unit of measurement, then changing each x_i into u_i by a linear transformation $u_i = (x_i - \bar{x})/(h/2)$, the variable u takes the values $-(2m-1), -(2m-3), \cdots, -3, -1, 1, 3, \cdots, (2m-3), (2m-1)$. Thus, we get

$$\Sigma u_i = 0 = \Sigma u_i^3 = \Sigma u_i^5 = \cdots$$

The values of the controlled variable are coded into integers symmetrically about zero. When the values of the coded variable u sum to zero, the least squares line of Y upon u becomes

$$\hat{y} = a + b\,u$$

where $a = \dfrac{\Sigma y_i}{n} = \bar{y}$ and $b = \dfrac{\Sigma u_i\, y_i}{\Sigma u_i^2}$

In the end, we must change the least squares line of Y on u into the least squares line of Y on x by transforming back the coded variable u into the original variable x. Sometimes, for the sake of further convenience each observed value y_i of the dependent variable can also be transformed into v_i by a linear transformation $v_i = (y_i - q)/k$ where q and k have arbitrary values.

***Example* 14.3** *The following table shows the tons of steel produced versus the number of workers in a small steel mill.*

Observation number i	*Number of workers* x_i	*Tons of steel produced* y_i
1	1	4
2	2	6
3	3	10
4	4	10
5	5	15
6	6	15
7	7	16
8	8	20

Estimate the line of regression using $u = \dfrac{x - 4.5}{1/2}$.

Solution. We have $\bar{x} = 4.5$ and $h = 1$. Let $u_i = \dfrac{x_i - \bar{x}}{h/2} = \dfrac{x_i - 4.5}{1/2}$

x_i	y_i	$u_i = \dfrac{x_i - 4.5}{1/2}$	$u_i\, y_i$	u_i^2
1	4	−7	− 28	49
2	6	−5	− 30	25
3	10	−3	− 30	9
4	10	−1	− 10	1
5	15	1	15	1
6	15	3	45	9
7	16	5	80	25
8	20	7	140	49
Sum	96	0	182	168

The estimated regression line of Y on u is $\hat{y} = a + b\,u$

The least squares estimates of a and b are

$$a = \frac{\Sigma y_i}{n} = \frac{96}{8} = 12$$

$$b = \frac{\Sigma u_i\, y_i}{\Sigma u_i^2} = \frac{182}{168} = 1.0833$$

The best fitted line of Y on u is $\hat{y} = 12 + 1.0833\,u$

Substituting $\dfrac{x - 4.5}{1/2}$ for u, we get the best fitted line of Y on x as

$$\hat{y} = 12 + 1.0833\left(\frac{x - 4.5}{0.5}\right) = 12 + \frac{1.0833}{0.5}(x - 4.5)$$

$$= 12 + 2.167\,(x - 4.5) = 12 + 2.167\,x - 9.75$$

$$= 2.25 + 2.167\,x$$

***Example* 14.4** *The following data show, in convenient units, the yield Y of a chemical reaction run at various different temperature x.*

$$n = 7, \quad \Sigma x_i = 980, \quad \Sigma y_i = 27.4, \quad \Sigma x_i\, y_i = 3958,$$

$$\Sigma x_i^2 = 140000, \quad \Sigma y_i^2 = 115.54$$

Assuming that a linear regression model $Y_i = \alpha + \beta x_i + \varepsilon_i$ is appropriate estimate the regression line of yield on temperature. Find the residuals sum of squares.

Solution. The estimated regression line is $\hat{y} = a + b\,x$

The least squares estimates a and b are

$$b = \frac{n\,\Sigma x_i\, y_i - (\Sigma x_i)(\Sigma y_i)}{n\,\Sigma x_i^2 - (\Sigma x_i)^2}$$

$$= \frac{7(3958) - (980)(27.4)}{7(140000) - (980)^2} = 0.04357$$

$$a = \frac{\Sigma y_i - b\Sigma x_i}{n} = \frac{27.4 - 0.04357(980)}{7} = -2.1855$$

The best fitted line is $\hat{y} = -2.1855 + 0.04357\,x$

The sum of squares of the residuals is

$$\Sigma e_i^2 = \Sigma y_i^2 - a\Sigma y_i - b\Sigma x_i\, y_i$$

$$= 115.54 - (-2.1855)(27.4) - 0.04357\,(3958) = 2.97$$

14.5.8 Limitations of Linear Regression. There are a number of limitations and cautions that must be kept in mind when using linear regression. They are.

Firstly, the linear regression is applicable only to relationships that can be described by a straight line. Non-linear regression methods exist to deal with some non-linear relationships. If you are in doubt about whether the data are approximately linear, a scatter diagram will help you to decide.

Secondly, the procedure used to find the regression coefficients a and b will give us a linear equation which is the best fit (*i. e.*, has the lowest value of $\sum e_i^2$) for the data, even when a linear relationship is non-existent. Therefore, the test of significance must be made to determine whether the regression coefficient b is "real".

Thirdly, the regression equation predicts values of the dependent variable based on values of the independent variable. It is therefore an *asymmetrical* measure. The regression equation predicting Y based on x (called "regression Y on x") cannot be used to derive the equation that will predict x based on Y.

Finally, the regression equation holds only for the range of values actually observed. The regression equation will not necessarily hold beyond this range.

Exercise 14.1

1. (*a*) What is a scatter diagram? Describe its role in the theory of regression.

(*b*) Explain what is meant by

(*i*) regression, (*ii*) regressand, (*iii*) regressor

(*c*) Explain what is meant by

(*i*) simple linear regression, (*ii*) simple linear regression coefficient.

2. (*a*) The following measurements of the specific heat of a certain chemical were made in order to investigate the variation in specific heat with temperature.

Temperature ($^{\circ}C$)	x_i	0	10	20	30	40
Specific heat	y_i	0.51	0.55	0.57	0.59	0.63

Plot the points on a scatter diagram and verify that the relationship is approximately linear. Estimate the regression line of specific heat on temperature, and hence estimate the value of the specific heat when the temperature is $25\ ^{\circ}C$.
($\hat{y} = 0.514 + 0.0028\,x$; $\hat{y} = 0.584$)

(*b*) Determine the estimated regression equation $\hat{y} = a + b\,x$ in each of the following cases

(*i*) $n = 10,\ \sum x_i = 20,\ \sum y_i = 260,\ \sum x_i\,y_i = 3490,\ \sum x_i^2 = 3144$

(*ii*) $n = 100,\ \bar{x} = 125,\ \bar{y} = 80,\ \sum x_i\,y_i = 1007425,\ \sum x_i^2 = 1585000$

(*iii*) $\bar{x} = 52,\ \bar{y} = 237,\ \sum (x_i - \bar{x})^2 = 2800,\ \sum (x_i - \bar{x})(y_i - \bar{y}) = 9871$

(*iv*) $n = 8,\ \bar{x} = 7,\ \bar{y} = 5,\ \sum x_i\,y_i = 364,\ \sum (x_i - \bar{x})^2 = 132$

($\hat{y} = 24.0864 + 0.9568\,x$; $\hat{y} = 38.75 + 0.33\,x$; $\hat{y} = 53.70 + 3.525\,x$; $\hat{y} = 0.5459 + 0.6363\,x$)

3. (*a*) Estimate the regression line of Y on x for the following data.

x_i	25	30	35	40	45	50
y_i	78	70	65	58	48	42

Is it possible from the equation you have just found

(*i*) an estimate for the value of x when $y = 54$?

(*ii*) an estimate for the value of y, when $x = 37$? In each case, if the answer is "Yes", calculate the estimate. If the answer is "No", say why not.

{ $\hat{y} = 114.4 - 1.45x$; (*i*) No, x is controlled; (*ii*) 61 }

(*b*) From n pairs of values (x_i, y_i), $i = 1, 2, \cdots, n$ the following quantities are calculated

$$n = 20, \quad \Sigma x_i = 400, \quad \Sigma y_i = 220,$$
$$\Sigma x_i^2 = 8800, \quad \Sigma y_i^2 = 2620, \quad \Sigma x_i y_i = 4300$$

Find the linear regression equation of y on x and x on y. Which would be the more useful if.

(*i*) x is the age (in years) and y is the reaction time (in milliseconds) of 20 people;

(*ii*) x is the cost (in ,000 Rs.) and y the floor-space (in 100 ft ^{2}) of 20 buildings

{ $\hat{y} = 13.5 - 0.125x$; $\hat{x} = 25.5 - 0.5y$; (*i*) y on x; (*ii*) x on y }

4. (*a*) The following table shows the ages x and systolic blood pressures Y of 12 women.

Age (years) x_i	56	42	72	36	63	47	55	49	38	42	68	60
Blood pressure y_i	147	125	160	118	149	128	150	145	115	140	152	155

Assuming that a linear regression model $Y_i = \alpha + \beta x_i + \varepsilon_i$ is appropriate, estimate the linear regression of blood pressure on age. Estimate the expected blood pressure of a woman whose age is 45 years. What is the change in blood pressure for a unit change in age?

($\hat{y} = 80.78 + 1.138x$; 132; 1.138)

(*b*) Suppose that four randomly chosen plots where treated with various levels of fertilizer, resulting in the following yields of corn.

Fertilizer (*kg*/Acre) x_i	100	200	400	500
Production (Bushels/ Acre) y_i	70	70	80	100

(*i*) Estimate the linear regression $\mu_{Y|x} = \alpha + \beta x$ of production Y on fertilizer x.

(*ii*) Estimate the yield when no fertilizer is applied.

(*iii*) Estimate the yield when the average amount of fertilizer is applied.

(*iv*) Estimate how much yield is increased for every kilogram of fertilizer applied.

{ (*i*) $\hat{y} = 59 + 0.070x$; (*ii*) when $x = 0$, $\hat{y} = 59$; (*iii*) when $x = \bar{x} = 300$, $\hat{y} = 80$; (*iv*) 0.070 bushels per kg of fertilizer }

5. (*a*) Describe the properties of the least squares regression line.

(*b*) Determine the regression line and estimate the weight of a student whose height is 68 inches.

Height (inches) x_i	72	66	67	69	74	61	66	62	70	63
Weight (pounds) y_i	178	141	158	165	180	133	159	140	160	136

Find also the estimated values for given values of height. Show that the sum of the estimated values is equal to the sum of the observed values of weight. Find the deviations $e_i = y_i - \hat{y}_i$. Show that these deviations add to zero.

($\hat{y} = -94.4 + 3.72x$; 158.76)

6. (*a*) Four identical money boxes contain different numbers of a particular type of coin and no coin of other types. The information on the combined weights, is given below.

Number of coins in box x_i	10	20	30	40
Combined weight of coins and box y_i	312	509	682	865

Estimate the regression line of Y on x. Estimate from your regression line,

(*i*) the weight of an empty box,

(*ii*) the mean weight of a single coin. State the co-ordinates of one point through which the line of regression of Y upon x must pass.

{ $\hat{y} = 134 + 18.32x$; (*i*) 134, (*ii*) 18.32; (25, 592) }

(*b*) Fifteen boys took two examination papers in the same subject and their marks as percentages were as follows, where each boy's marks are in the same column.

Paper I x_i	65	73	42	52	84	60	70	79	60	83	57	77	54	66	89
Paper II y_i	78	88	60	73	92	77	84	89	70	89	73	88	70	85	89

Calculate the equation of the line of regression of Y on x. Two boys were each absent from one paper. One scored 63 on paper I, the other scored 81 on paper II. In which case can you use your regression line to estimate the mark that the boy should be allocated for the paper he did not take, and what is that mark ?

($\hat{y} = 35.53 + 0.665x$, 63 on I $\Rightarrow$ 78 on II)

7. (*a*) The following data shows the son's height and father's height.

Father's height (inches) x_i	59	61	63	65	67	69	71	73	75
Son's height (inches) y_i	64	66	67	67	68	69	70	72	72

Estimate the regression line $\mu_{Y|x} = \alpha + \beta x$ of son's height on father's height using $u_i = (x_i - 67)/2$ and $v_i = y_i - 68$. Predict the mean height of sons whose fathers are 70 inches in height.

($\hat{y} = 35.95 + 0.4833x$; 69.78)

(*b*) For 9 observations on supply X and price Y the following data was obtained

$\Sigma(x_i - 90) = -25$, $\Sigma(x_i - 90)^2 = 301$, $\Sigma(y_i - 127) = 12$,

$\Sigma(y_i - 127)^2 = 1006$, $\Sigma(x_i - 90)(y_i - 127) = -469$

Obtain the estimated line of regression of X on Y and estimate the supply when the

price is Rs. 125.
($\hat{x}$ = 143.69 − 0.44 y; 88.69)

(c) Number of revolutions x (per minute) and power y (hp) of a diesel engine are

x_i	400	500	600	700	800
y_i	580	1030	1420	1880	2310

Determine the regression line of the y-values on the x – values of the sample using $x_i = 100\,u_i + 600$ and $y_i = 10\,v_i + 1400$ estimate y when $x = 750$.
($\hat{y}$ = − 1142 + 4.31 x; 2090.5)

8. (a) Fit a straight line taking x as independent variable

$3x_i + 2$	5	8	11	14	17	20	23	26
$3y_i - 2$	7	10	16	16	25	28	28	34

Also estimate y for $x = 5/3$.
($\hat{y}$ = 1.714 + 1.286 x; 3.86)

(b) Fit a least squares line to following data taking (*i*) Y as dependent variable (*ii*) X as dependent variable.

x_i	1	3	4	6	8	9	11	14
y_i	1	2	4	4	5	7	8	9

Show that the two least squares lines obtained intersect at the point $(\bar{x}, \bar{y})$. estimate the mean value of y when $x = 7$. Estimate the mean value of x when $y = 6$.
($\hat{y}$ = 0.5455 + 0.6364 x, $\hat{y}$ = 5 for x = 7;
$\hat{x}$ = −0.5 + 1.5 y, $\hat{x}$ = 8.5 for y = 6)

9. (a) A random sample of 5 pairs of observations (x_i, y_i) is given below

x_i	3	2	5	1	4
y_i	13	9	27	8	18

Determine the least squares linear regression $\hat{y} = a_{yx} + b_{yx}\,x$ and estimate y for $x = 6$. Also find the least squares linear regression $\hat{x} = a_{xy} + b_{xy}\,y$ and use this to find that value of y for which $\hat{x} = 6$. Account for the difference.
($\hat{y}$ = 0.9 + 4.7 x, $\hat{y}$ = 29.1 for x = 6; $\hat{x}$ = 0.09 + 0.194 y, y = 30.46 for $\hat{x}$ = 6 which is a useless estimate, because the regression analysis does not permit the inverse use of the least squares line.)

(b) Compute the regression coefficients in each of the following cases:

(*i*) $n = 10$, $\Sigma(x_i - \bar{x})^2 = 170$, $\Sigma(y_i - \bar{y})^2 = 140$, $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 92$

(*ii*) $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 148$, $s_x = 7.933$, $s_y = 16.627$, $n = 15$

($b_{yx} = 0.54$, $b_{xy} = 0.66$; $b_{yx} = 0.16$, $b_{xy} = 0.04$)

◆

14.6 SIMPLE LINEAR CORRELATION

The *simple linear correlation* measures the strength or closeness of linear relationships between two variables. The purpose of simple linear correlation is to determine whether or not two variables are related, that is, whether one variable tends to increase (or decrease) as the other variable increases. The correlation analysis is performed keeping in view the following two aspects.

(*i*) It measures the closeness of the linear regression to the distribution of observations of a dependent variable with associated values of an independent variable.

(*ii*) It measures the degree (extent or strength) of covariability between two variables.

We have discussed this first aspect in the preceding chapter. We shall now discuss the second aspect of correlation. This approach to the problem of understanding the relationship between two variables is to leave the type or form of the relationship unspecified and concentrate on measuring the strength of the relationship itself.

14.6.1 Positive Correlation. The correlation is said to be *positive* (or *direct*) if the two random variables tend to move in the same direction, *i. e.*, increase (or decrease) simultaneously. That is, the correlation is positive if the least squares regression lines have positive slopes.

Perfect Positive Correlation. The correlation is said to be *perfect positive* if the relationship between the two random variables is perfectly linear with positive slope.

14.6.2 Negative Correlation. The correlation is said to be *negative* (or *inverse*) if the two random variables tend to move in opposite directions, *i. e.*, one random variable decreases as the other random variable increases. That is, the correlation is negative if the least squares regression lines have negative slopes.

Perfect Negative Correlation. The correlation is said to be *perfect negative* if the relationship between the two random variables is perfectly linear with negative slope.

14.6.3 No Correlation. If one least squares regression line is horizontal and the other least squares regression line is vertical then there is no correlation between the two random variables. That is, if X and Y are independent, then $Cov(X, Y) = 0$ which implies that $\rho = 0$ and we say that there is no correlation.

14.7 CORRELATION ANALYSIS

One of the most widely used statistical techniques applied by statistician is *correlation analysis.* In purely correlation problems both the variables X and Y are random and the relationship between them is considered simultaneously and symmetrically. Examples of correlation problems are: (*i*) heights and weights of persons, (*ii*) ages of husbands and ages of wives at the time of their marriages, (*iii*) I. Q. of brothers and I. Q. of sisters, (*iv*) marks of students in economics and in statistics, (*v*) income and I. Q. of persons, (*vi*) demand and supply of a commodity, (*vii*) daily wages and overtime wages, (*viii*) gold prices and silver prices, (*ix*) the height and the circumference of head of babies at the time of their birth, (*x*) the greatest and the smallest diameters of hen eggs, *etc.*

In correlation problems, we sample from a population, observing two measurements on each individual in the sample. For example, if a person is selected at random, and both his height and weight are left free to take any possible values. Thus we have a joint distribution of two random variables or we may say that we have bivariate distribution. The data are assumed to be obtained by taking a random sample of values of X and Y.

14.7.1 Sample Correlation Coefficient. If $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ is a random sample of n pairs of observations from a bivariate population, then the sample correlation coefficient, denoted by r or more appropriately r_{xy}, is defined as

$$r = \frac{s_{xy}}{s_x s_y}$$

It can be expressed as

$$r = \frac{s_{xy}}{s_x s_y} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})/n}{\sqrt{\{\Sigma(x_i - \bar{x})^2/n\}\{\Sigma(y_i - \bar{y})^2/n\}}}$$

$$= \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}} = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\Sigma x_i^2 - n\bar{x}^2)(\Sigma y_i^2 - n\bar{y}^2)}}$$

$$= \frac{\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n}{\sqrt{\{\Sigma x_i^2 - (\Sigma x_i)^2/n\}\{\Sigma y_i^2 - (\Sigma y_i)^2/n\}}},$$

$$= \frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{\sqrt{\{n\Sigma x_i^2 - (\Sigma x_i)^2\}\{n\Sigma y_i^2 - (\Sigma y_i)^2\}}}$$

This r is the maximum likelihood estimate of ρ. The process of subtracting $\bar{x}$ and $\bar{y}$ indicates that the origin has been shifted to $(\bar{x}, \bar{y})$.

14.7.2 Properties of Sample Correlation Coefficient r. The sample correlation coefficient r has the following properties.

(1) r is symmetrical with respect to the variables X and Y, that is

$$r_{xy} = r_{yx}$$

(2) r is the covariance of values of the two variables X and Y measured in standard units, that is

$$r = Cov(z_x, z_y)$$

(3) ***Change of Origin and Scale.*** The value of r remains unchanged if constants are added to or subtracted from the values of the variables or if the values of the variables are multiplied or divided by constants having the same sign, but the value of r changes in sign only if the values of the variables are multiplied or divided by constants having opposite signs. That is, the magnitude of the sample correlation coefficient $|r|$ is independent of change of origin and scale.

(4) r always lies between -1 and $+1$, *i.e.*,

$$-1 \leq r \leq 1$$

(5) $|r|$ is the geometric mean of the two regression coefficients b_{yx} and b_{xy}, that is

$$r = (+/-)\sqrt{b_{yx} \times b_{xy}}.$$

$$\text{Thus} \quad r = \begin{cases} +\sqrt{b_{yx} \times b_{xy}} & \text{if } b_{yx} \text{ and } b_{xy} \text{ are positive} \\ -\sqrt{b_{yx} \times b_{xy}} & \text{if } b_{yx} \text{ and } b_{xy} \text{ are negative} \end{cases}$$

(6) r is zero when one of the variables X or Y is constant.

***Theorem* 14.3** *The correlation coefficient is independent of the origin and the scale of measurement of the variables.*

Let $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ *be a random sample of* n *pairs of observations from a bivariate population.*

If we let $u_i = (x_i - p)/h$ *and* $v_i = (y_i - q)/k$ *then* $r_{xy} = \dfrac{hk}{|h||k|} r_{uv}$

$$\text{That is,} \quad r_{xy} = \begin{cases} r_{uv} & \text{if } h \text{ and } k \text{ have same signs} \\ -r_{uv} & \text{if } h \text{ and } k \text{ have different signs} \end{cases}$$

Similarly, if we let $u_i = p + h x_i$ *and* $v_i = q + k y_i$, *then* $r_{uv} = \dfrac{hk}{|h||k|} r_{xy}$

$$\text{That is,} \quad r_{uv} = \begin{cases} r_{xy} & \text{if } h \text{ and } k \text{ have same signs} \\ -r_{xy} & \text{if } h \text{ and } k \text{ have different signs} \end{cases}$$

***Example* 14.5** *The following are the measurements of height and weight of* 8 *men.*

Height (inches) x_i	78	89	97	69	59	79	68	61
Weight (pound) y_i	125	137	156	112	107	136	123	104

(*i*) *Calculate the correlation coefficient between the height and weight of eight men by using the deviations from their means.*

(*ii*) *Again compute the correlation coefficient by taking the deviations of variable* X *from* 70 *and of variable* Y *from* 120.

(*iii*) *Do the results in* (*i*) *and* (*ii*) *agree?*

Solution. (*i*) The coefficient of correlation between X and Y is

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
78	125	3	0	9	0	0
89	137	14	12	196	144	168
97	156	22	31	484	961	682
69	112	−6	−13	36	169	78
59	107	−16	−18	256	324	288
79	136	4	11	16	121	44
68	123	−7	−2	49	4	14
61	104	−14	−21	196	441	294
600	1000	0	0	1242	2164	1568

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{600}{8} = 75$$

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{1000}{8} = 125$$

$$r_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \, \Sigma(y_i - \bar{y})^2}} = \frac{1568}{\sqrt{(1242)(2164)}} = 0.956$$

(*ii*) Let the assumed mean for x be $p = 70$, and $u_i = x_i - p = x_i - 70$. Let the assumed mean for y be $q = 120$, and $v_i = y_i - q = y_i - 120$.

x_i	y_i	$u_i = x_i - 70$	$v_i = y_i - 120$	u_i^2	v_i^2	$u_i v_i$
78	125	8	5	64	25	40
89	137	19	17	361	289	323
97	156	27	36	729	1296	972
69	112	−1	−8	1	64	8
59	107	−11	−13	121	169	143
79	136	9	16	81	256	144
68	123	−2	3	4	9	−6
61	104	−9	−16	81	256	144
		40	40	1442	2364	1768

The coefficient of correlation between U and V is

$$r_{uv} = \frac{n \Sigma u_i v_i - (\Sigma u_i)(\Sigma v_i)}{\sqrt{\{n \Sigma u_i^2 - (\Sigma u_i)^2\}\{n \Sigma v_i^2 - (\Sigma v_i)^2\}}}$$

$$= \frac{8(1768) - (40)(40)}{\sqrt{\{8(1442) - (40)^2\}\{8(2364) - (40)^2\}}} = 0.956$$

(*iii*) The results in (*i*) and (*ii*) are same, since $0.956 = r_{xy} = r_{uv} = 0.956$.

***Example* 14.6** The following data were obtained for a sample of 10 persons from a height and weight distribution.

$$\Sigma x_i = 700, \quad \Sigma y_i = 1550, \quad \Sigma x_i^2 = 49120, \quad \Sigma y_i^2 = 240550, \quad \Sigma x_i y_i = 108650$$

Compute the coefficient of correlation.

Solution. The coefficient of correlation between X and Y is

$$r = \frac{n \Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{\sqrt{\{n \Sigma x_i^2 - (\Sigma x_i)^2\}\{n \Sigma y_i^2 - (\Sigma y_i)^2\}}}$$

$$= \frac{10(108650) - (700)(1550)}{\sqrt{\{10(49120) - (700)^2\}\{10(240550) - (1550)^2\}}} = 0.79$$

14.7.3 Goodness of Fit of a Linear Regression Equation. One approach to correlation analysis emphasized the covariability of the two random variables. The other approach to correlation analysis is related to regression analysis and provides a measure of the strength of closeness of the linear relationship between two variables; thus correlation coefficient is a measure of the goodness of fit of the linear regression equation. Consider a sample from a bivariate distribution of X and Y. There are two regression functions, each obtained by considering that variable as dependent whose mean value is to be estimated and treating the other variable as independent. The two linear regression functions of Y on X and of X on Y are

$$\mu_{Y|x} = \alpha_{YX} + \beta_{YX} X, \qquad \mu_{X|y} = \alpha_{XY} + \beta_{XY} Y$$

where β_{YX} and β_{XY} are the population regression coefficients of Y on X and of X on Y respectively. Their corresponding least squares sample regression lines are

$$\hat{y} = a_{yx} + b_{yx}x \qquad \hat{x} = a_{xy} + b_{xy}y$$

where b_{yx} and b_{xy} are the sample regression coefficients of Y on X and of X on Y respectively. The least squares estimates are

$$b_{yx} = \frac{s_{xy}}{s_x^2}, \qquad a_{yx} = \bar{y} - b_{yx}\bar{x}$$

$$b_{xy} = \frac{s_{xy}}{s_y^2}, \qquad a_{xy} = \bar{x} - b_{xy}\bar{y}$$

The regression equation of Y on X becomes

$$\hat{y} = \bar{y} + b_{yx}(x - \bar{x})$$

and the regression equation of X on Y becomes

$$\hat{x} = \bar{x} + b_{xy}(y - \bar{y})$$

The regression coefficients are related to the correlation coefficient as

$$b_{yx} = \frac{s_{xy}}{s_x^2} = \frac{r\,s_y}{s_x}, \qquad b_{xy} = \frac{s_{xy}}{s_y^2} = \frac{r\,s_x}{s_y}$$

Thus the regression equation of Y on X becomes

$$\hat{y} = \bar{y} + \frac{r\,s_y}{s_x}(x - \bar{x})$$

and the regression equation of X on Y becomes

$$\hat{x} = \bar{x} + \frac{r\,s_x}{s_y}(y - \bar{y})$$

Since s_x and s_y are positive, the sign of s_{xy}, b_{yx}, b_{xy} and r_{xy} will always be same. Note also that if any one of these four quantities is zero then all others must equal to zero. A positive sign of r_{xy} indicates that X and Y are directly related. A direct relationship between X and Y is associated with an upward sloping regression line; that is as one variable increases other variable also increases. A negative sign of r_{xy} indicates that X and Y are inversely related. An inverse relationship between X and Y is associated with a downward sloping regression line, that is as one variable increases, the other variable decreases.

***Theorem* 14.4** *In the correlation analysis the two regression lines intersect at the* point $(\bar{x}, \bar{y})$.

***Theorem* 14.5** *The correlation coefficient r is the slope of the regression lines for standard scores.*

***Theorem* 14.6** *The graphs of the regression lines of Y on X and X on Y are identical if all the points of the given sample lie on a straight line.*

***Example* 14.7** *The following data were obtained for a sample of* 10 *men from a height and weight distribution.*

$$\bar{x} = 70, \qquad \bar{y} = 155, \qquad \sum (x_i - \bar{x})^2 = 120,$$

$$\sum y_i^2 = 240550, \qquad \sum (x_i - \bar{x})(y_i - \bar{y}) = 150$$

Calculate covariance, correlation coefficient, the two regression lines.

Solution. The variance of X, variance of Y, covariance and correlation coefficient are

$$s_x^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{120}{10} = 12$$

$$s_y^2 = \frac{\sum y_i^2}{n} - \bar{y}^2 = \frac{240550}{10} - (155)^2 = 30$$

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{150}{10} = 15$$

$$r = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{15}{\sqrt{(12)(30)}} = 0.79$$

The estimated regression line of Y on X is $\hat{y} = a_{yx} + b_{yx} x$

The least squares estimates of a_{yx} and b_{yx} are

$$b_{yx} = \frac{s_{xy}}{s_x^2} = \frac{15}{12} = 1.25$$

$$a_{yx} = \bar{y} - b_{yx} \bar{x} = 155 - 1.25(70) = 67.5$$

The best fitted line of Y on X is $\hat{y} = 67.5 + 1.25 x$

The estimated regression line of X on Y is $\hat{x} = a_{xy} + b_{xy} y$

The least squares estimates of a_{xy} and b_{xy} are

$$b_{xy} = \frac{s_{xy}}{s_y^2} = \frac{15}{30} = 0.50$$

$$a_{xy} = \bar{x} - b_{xy} \bar{y} = 70 - (0.50)(155) = -7.5$$

The best fitted line of X on Y is $\hat{x} = -7.5 + 0.5 y$

***Example* 14.8** *Find the coefficient of correlation if the two regression coefficients have the following values*

(*i*) 0.45 and 0.8, (*ii*) -0.1 and -0.4.

Solution.

(*i*) $r = (+/-) \sqrt{b_{yx} \times b_{xy}} = + \sqrt{(0.45)(0.8)} = 0.6$

(*ii*) $r = (+/-) \sqrt{b_{yx} \times b_{xy}} = - \sqrt{(-0.1)(-0.4)} = -0.2$

***Example* 14.9** *The coefficient of correlation, for a sample of* 20 *pairs of observations is* 0.6. *If* $\bar{x} = 12$, $\bar{y} = 20$, $s_x = 1.5$ and $s_y = 2$, *estimate the lines of the regression. Estimate the mean of* Y *for* $x = 10$. *Estimate the mean of* X *for* $y = 22$.

Solution. The estimated regression line of Y on X is $\hat{y} = a_{yx} + b_{yx}\, x$

The least squares estimates of a_{yx} and b_{yx} are

$$b_{yx} = \frac{r\, s_y}{s_x} = \frac{0.6(2)}{1.5} = 0.8$$

$$a_{yx} = \bar{y} - b_{yx}\, \bar{x} = 20 - 0.8\,(12) = 10.4$$

The best fitted line of Y on X is

$$\hat{y} = 10.4 + 0.8\, x$$

For $x = 10$, we have

$$\hat{y} = 10.4 + 0.8\,(10) = 18.4$$

The estimated regression line of X on Y is $\hat{x} = a_{xy} + b_{xy}\, y$

The least squares estimates of a_{xy} and b_{xy} are

$$b_{xy} = \frac{r\, s_x}{s_y} = \frac{0.6(1.5)}{2} = 0.45$$

$$a_{xy} = \bar{x} - b_{xy}\, \bar{y} = 12 - 0.45\,(20) = 3$$

The best fitted line of X on Y is

$$\hat{x} = 3 + 0.45\, y$$

For $y = 22$, we have

$$\hat{x} = 3 + 0.45\,(22) = 12.9$$

***Example* 14.10** *The following results are given from paired data of two variables* X *and* Y.

Estimate of variance of $X = 9$

Estimated regression line of X *on* Y: $40\,\hat{x} - 18\, y = 214$

Estimated regression line of Y *on* X: $8\, x - 10\,\hat{y} = -66$

Find (i) *The coefficient of correlation between* X *and* Y, (ii) *Standard deviation of* Y, (iii) *Mean values of* X *and* Y.

Solution. (i) The estimated regression line of X on Y is

$$40\hat{x} - 18\, y = 214 \quad \Rightarrow \quad \hat{x} = \frac{214}{40} + \frac{18}{40} y \quad \Rightarrow \quad b_{xy} = \frac{18}{40} = 0.45$$

The estimated regression line of Y on X is

$$8x - 10\hat{y} = -66 \quad \Rightarrow \quad \hat{y} = \frac{66}{10} + \frac{8}{10}x \quad \Rightarrow \quad b_{yx} = \frac{8}{10} = 0.8$$

The estimate of the correlation coefficient between X on Y is

$$r = (+/-)\sqrt{b_{yx} \times b_{xy}} = +\sqrt{(0.8)(0.45)} = 0.6$$

(ii) $s_x^2 = 9 \quad \Rightarrow \quad s_x = +\sqrt{9} = 3$

$$b_{yx} = \frac{r_{xy}\, s_y}{s_x}$$

$$0.8 = \frac{0.6\, s_y}{3} \quad \Rightarrow \quad 0.6\, s_y = 0.8(3) \quad \Rightarrow \quad s_y = 4$$

(iii) Since both the estimated regression lines pass through the point $(\bar{x}, \bar{y})$. Thus

$$40\bar{x} - 18\bar{y} = 214 \ldots\ldots (i)$$

$$8\bar{x} - 10\bar{y} = -66 \ldots\ldots (ii)$$

Multiplying *(ii)* by 5 and subtracting it from *(i)*

$$\begin{array}{rrr} 40\bar{x} - 18\bar{y} = & 214 \\ 40\bar{x} - 50\bar{y} = & -330 \\ - \quad + & + \\ \hline 32\bar{y} = & 544 \end{array} \quad \Rightarrow \quad \bar{y} = 17$$

Putting this value of $\bar{y}$ in *(ii)*, we have

$$8\bar{x} - 10(17) = -66 \quad \Rightarrow \quad \bar{x} = 13$$

14.7.4 Correlation and Causation. It is necessary to consider the sampling distribution of the sample statistic R to decide whether or not we should accept the hypothesis that the variables in the population are related. But aside from this technical aspect of a relation between two variables, it is necessary for a statistician to consider whether or not correlation indicates a cause and effect relationship. It is possible to correlate the temperature of Lahore city with the birth rate and it is possible that a high positive correlation may be found showing that when the temperature is high, the birth rate is high, and when the temperature is low the birth rate is low.

There is no meaning to such a correlation. There is no causal relationship between the two phenomena. This example illustrate that you can correlate anything, and there are chances you may obtain a high correlation which may have no significant meaning at all. A high correlation simply tells us that the data we have collected is *consistent* with the hypothesis we set up. That is, it supports our hypothesis. We may say the following situations that brought about a high correlation.

(*i*) X is the cause of Y.

(*ii*) Y is the cause of X.

(*iii*) There is a third factor Z that affects X and Y such that they show a close relation.

(*iv*) The correlation between X and Y may be due to chance.

Only by more thorough investigation we can come to some conclusion as to whether or not X is the cause of Y.

Exercise 14.2

1. **(a)** Differentiate between regression and correlation problems, giving examples.

(b) Define the terms correlation and product moment co-efficient of correlation.

(c) For a set of 50 pairs of observations on variables X and Y, we have $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 450$. Find the covariance.
$(s_{xy} = 9)$

2. **(a)** The simple correlation coefficient $r = s_{xy}/(s_x s_y)$ is given as

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \, \Sigma(y_i - \bar{y})^2}}$$

The following table gives the ages of husbands and the ages of wives at the time of their marriage.

Couple	i	1	2	3	4	5	6	7	8	9	10
Husband's age	x_i	25	29	30	30	31	32	33	35	37	38
Wife's age	y_i	20	22	24	29	23	31	29	31	30	31

Calculate the coefficient of correlation by using the above formula.
$(r = 0.82)$

(b) The simple correlation coefficient $r = s_{xy}/(s_x s_y)$ is given as

$$r = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\sqrt{\{\Sigma x_i^2 - n\bar{x}^2\}\{\Sigma y_i^2 - n\bar{y}^2\}}}$$

The following table gives the demand and supply of a commodity.

Supply	x_i	400	200	700	100	500	300	600
Demand	y_i	50	60	20	70	40	30	10

Calculate the coefficient of correlation by using the above formula.
$(r = -0.857)$

(c) The simple correlation coefficient $r = s_{xy}/(s_x s_y)$ is given as

$$r = \frac{\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n}{\sqrt{\{\Sigma x_i^2 - (\Sigma x_i)^2/n\}\{\Sigma y_i^2 - (\Sigma y_i)^2/n\}}}$$

The following table gives the traffic density and accident rate.

Traffic density	x_i	30	35	40	45	50	60	70	80	90
Accident rate	y_i	2	4	5	5	8	15	24	30	32

Calculate the coefficient of correlation by using the above formula.
($r = 0.983$)

(d) The simple correlation coefficient $r = s_{xy}/(s_x s_y)$ is given as

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{\{n\sum x_i^2 - (\sum x_i)^2\}\{n\sum y_i^2 - (\sum y_i)^2\}}}$$

The following table gives the number of persons employed and cloth manufactured in a textile mill..

Persons employed	x_i	137	209	113	189	176	200	219
Cloth manufactured	y_i	23	47	22	40	39	51	49

Calculate the coefficient of correlation by using the above formula.
($r = 0.963$)

3. (a) For a set of 22 pairs of observations, we have

$$\sum x_i = 983, \quad \sum y_i = 409, \quad \sum x_i^2 = 61339, \quad \sum y_i^2 = 8475, \quad \sum x_i y_i = 15811$$

Find the product moment correlation coefficient for the data.
($r = -0.6325$)

(b) For a sample of 20 pairs of observations, we have

$$\bar{x} = 2, \quad \bar{y} = 8, \quad \sum x_i^2 = 180, \quad \sum y_i^2 = 3424, \quad \sum x_i y_i = 604$$

Calculate the coefficient of correlation.
($r = 0.6133$)

(c) For a set of 8 pairs of observations, we have

$$\sum x_i = 448, \quad \sum y_i = 472, \quad \sum y_i^2 = 29958, \quad \sum x_i y_i = 26762, \quad s_x = 16.6$$

Compute the product moment correlation coefficient.
($r = 0.15$)

4. (a) For a set of 50 pairs of observations, the standard deviations of x and y are 4.5 and 3.5 respectively. If the sum of products of deviations of x and y values from their respective means be 420, find the Karl Pearson's coefficient of correlation.
($r = 0.53$)

(b) For a given set of data, we have $s_x^2 = 9.102, \quad s_y^2 = 2.204, \quad s_{xy} = 1.694$

Find the product moment correlation coefficient for the data.
($r = 0.378$)

5. (a) For a given set of data, we have $r = 0.48, \quad s_{xy} = 36, \quad s_x^2 = 16$. Find s_y.
($s_y = 18.75$)

(b) For a given set of data, we have

$$r = 0.5, \quad \sum(x_i - \bar{x})(y_i - \bar{y}) = 120, \quad s_y = 8, \quad \sum(x_i - \bar{x})^2 = 90$$

Find the number of pairs of values.
($n = 10$)

6. (a) A computer while calculating the correlation coefficient between two variables X and Y from 25 pairs of observations obtained the following results.

$$\sum x_i = 125, \quad \sum x_i^2 = 650, \quad \sum y_i = 100, \quad \sum y_i^2 = 460, \quad \sum x_i y_i = 508$$

It was, however, later discovered at the time of checking that he had copied down two pairs as:

x_i	6	8
y_i	14	6

while the correct values were:

x_i	8	6
y_i	12	8

Obtain the correct value of coefficient of correlation.
($r = 0.67$)

(*b*) The following data show the marks in economics and marks in statistics obtained by ten students.

Student	i	1	2	3	4	5	6	7	8	9	10
Economics	x_i	78	36	96	25	75	82	90	62	65	39
Statistics	y_i	84	51	91	60	68	62	86	58	53	47

(*i*) Compute the coefficient of correlation.

(*ii*) Again compute the coefficient of correlation by taking the deviations of variable X from 50 and of variable Y from 60.

(*iii*) Do the results in (*i*) and (*ii*) agree?

{ (*i*) 0.775, (*ii*) 0.775, (*iii*) Yes }

(*c*) Compute the correlation coefficient between the variables X and Y represented in the following table:

x_i	2	4	5	6	8	11
y_i	18	12	10	8	7	5

Multiply each x_i value by 2 and add 6. Multiply each y_i value by 3 and subtract 15. Find the correlation co-efficient between the two new sets of values, explaining why you do or do not obtain the same result as above.
(– 0.92)

7. (*a*) Interpret the meaning when

$r = -1$, $r = 0$, $r = 1$

(*b*) Sketch scatter diagrams which illustrate:

(*i*) positive linear correlation, (*ii*) perfect positive linear correlation,

(*iii*) negative linear correlation, (*iv*) perfect negative linear correlation,

(*v*) no correlation, between two variables X and Y.

8. (*a*) From the data given below, calculate the coefficient of correlation between the ages of husbands and ages of wives at the time of their marriage.

Couple	i	1	2	3	4	5	6	7	8	9	10
Husband's age	x_i	28	27	28	23	29	30	36	35	33	31
Wife's age	y_i	27	20	22	18	21	29	29	28	29	27

Find the regression coefficients. Verify that r is the geometric mean of the two regression coefficients
($r = 0.82$, $b_{yx} = 0.89$, $b_{xy} = 0.75$)

(*b*) The two regression coefficients have following values, find r.

(*i*) $b_{yx} = 0.86,\ b_{xy} = 0.95$

(*ii*) $b_{yx} = -0.52,\ b_{xy} = -1.02$

{ (*i*) $r = 0.90$, (*ii*) $r = -0.73$ }

(*c*) Find the two regression coefficients in each of the following cases.

(*i*) $\Sigma x_i = 17.6,\quad \Sigma y_i = 32.8,\quad \Sigma x_i y_i = 94.7,$
$\Sigma x_i^2 = 49.64,\quad \Sigma y_i^2 = 182,\quad n = 8$

(*ii*) $n = 10,\quad \Sigma(x_i - \bar{x})^2 = 170,\quad \Sigma(y_i - \bar{y})^2 = 140,$
$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 92$

(*iii*) $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 148,\quad s_x = 7.933,\quad s_y = 16.627,\quad n = 15$

(*iv*) $n = 8,\quad \bar{x} = 7,\quad \bar{y} = 5,\quad \Sigma x_i y_i = 364,$
$\Sigma(x_i - \bar{x})^2 = 132,\quad \Sigma(y_i - \bar{y})^2 = 56.$

(*v*) $r = 0.97,\quad s_x = 17.08,\quad s_y = 14.34$

{ (*i*) $b_{yx} = 2.06,\ b_{xy} = 0.47$; (*ii*) $b_{yx} = 0.54,\ b_{xy} = 0.66$; (*iii*) $b_{yx} = 0.16,$ $b_{xy} = 0.04$; (*iv*) $b_{yx} = 0.64,\ b_{xy} = 1.5$; (*v*) $b_{yx} = 0.81,\ b_{xy} = 1.16$ }

9. (*a*) Explain why the regression line of Y on X is not necessarily the same as the regression line of X on Y. How would you decide which is the appropriate regression in any particular situation. Answer the following.?

(*i*) When do the two lines coincide?

(*ii*) When are they at right angles?

{ (*i*) Exact linear relation. (*ii*) Uncorrelated X, Y (*i. e.*, $\rho = 0$) }

(*b*) Calculate the coefficient of correlation and obtain the lines of regression from the following data

Price x_i	3	4	5	6	7	8	9	10	11	12
Demand y_i	25	24	20	20	19	17	16	13	10	6

($r = -0.98,\ \hat{y} = 31.45 - 1.93\,x,\ \hat{x} = 16 - 0.5\,y$)

(*c*) Given the following data:

$n = 100,\quad \Sigma x_i = 5000,\quad \Sigma y_i = 6000,$

$\Sigma x_i y_i = 300300,\quad \Sigma x_i^2 = 250400,\quad \Sigma y_i^2 = 360900,$

Calculate

(*i*) s_x, s_y and r,

(*ii*) regression lines,

(*iii*) estimate the value of y for $x = 55$.

{ (*i*) $s_x = 2,\ s_y = 3,\ r = 0.5$, (*ii*) $\hat{y} = 22.5 + 0.75\,x,\ \hat{x} = 30.2 + 0.33\,y,$ (*iii*) 63.75 }

10. (*a*) Given the following data:

$$n = 10, \quad \Sigma x_i = 120, \quad \Sigma y_i = 250, \quad \Sigma x_i\, y_i = 3070.7, \quad s_x = 3.5, \quad s_y = 7.2$$

Calculate regression lines.

($\hat{y} = 18.04 + 0.58\,x$; $\hat{x} = 8.50 + 0.14\,y$)

(*b*) Given that means and variances of two series X and Y are

	X-series	*Y*-series
Mean:	25	38
Variance:	25	36

The correlation coefficient between X and Y is 0.75. Estimate the most plausible value of Y for $x = 40$ and most plausible value of X for $y = 58$.

($\hat{y} = 15.5 + 0.9\,x$, 51.5; $\hat{x} = 1.25 + 0.625\,y$, 37.5)

(*c*) If the mean height of 500 fathers is 68.65 inches with standard deviation of 2.8 inches and the mean height of their youngest sons is 69.65 inches with standard deviation of 2.85 inches and the coefficient of correlation between them is 0.52 obtain the two equations of the lines of regression in the simplest form.

($\hat{y} = 33.27 + 0.53\,x$; $\hat{x} = 33.13 + 0.51\,y$)

11. (*a*) Given the following data:

$$\bar{x} = 54, \quad \bar{y} = 28, \quad b_{yx} = -1.5, \quad b_{xy} = -0.2$$

Show that the two estimated lines of regression intersect at the point $(\bar{x}, \bar{y})$. Estimate the value of X when $Y = 30$ and the value of Y when $X = 55$.

Hint: *Show that the estimated value of X for $Y = \bar{y} = 28$ is 54 and the estimated value of Y for $X = \bar{x} = 54$ is 28.*

($\hat{y} = 26.5$; $\hat{x} = 53.6$)

(b) For a given set of data, the least squares regression lines are

Estimated regression line of Y on X: $\hat{y} = 20.8 - 0.219\,x$

Estimated regression line of X on Y: $\hat{x} = 16.2 - 0.785\,y$

Find the product moment correlation coefficient for the data.

($r = -0.415$)

12. (*a*) For the following set of data, use $u_i = x_i - 1000$, and $v_i = (y_i - 250)/5$ to find the product moment correlation coefficient and the least squares lines of regression of Y on X and of X on Y.

x_i	1000	1012	1009	1007	1010	1015	1010	1011
y_i	235	240	245	250	255	260	265	270

(0.583; $\hat{y} = -1386.1346 + 1.6235\,x$; $\hat{x} = 956.326 + 0.2096\,y$)

(*b*) On each of 30 items, two measurements are made on the variables X and Y. The following summations are given

$$\Sigma x_i = 15, \quad \Sigma y_i = -6, \quad \Sigma x_i^2 = 61, \quad \Sigma y_i^2 = 90, \quad \Sigma x_i\, y_i = 56$$

Calculate the product moment correlation coefficient and obtain the lines of regression of Y on X and X on Y. If the variable X is replaced by U where $u_i = (x_i - 1)/2$, find the correlation coefficient between U and Y and the regression lines of Y on U and U on Y.
(0.856; $\hat{y} = -0.751 + 1.10x$; $\hat{x} = 0.633 + 0.664y$; 0.856; $\hat{y} = 0.351 + 2.21u$; $\hat{u} = -0.184 + 0.332y$)

(c) The following table shows the marks in statistics and mathematics obtained by 10 students from a large group of students.

Marks in Statistics	x_i	75	80	93	65	87	71	98	68	84	77
Marks in Mathematics	y_i	82	78	86	72	91	80	95	72	89	74

Estimate the linear regression function considering
(*i*) X as independent variable,
(*ii*) Y as independent variable.
($\hat{y} = 29.13 + 0.661x$; $\hat{x} = -14.39 + 1.15y$)

13. (a) A random sample of 20 pairs of observations (x_i, y_i) gave the following

$$\bar{x} = 2, \quad \bar{y} = 8, \quad \Sigma x_i^2 = 180, \quad \Sigma y_i^2 = 1424, \quad \Sigma x_i y_i = 404$$

Estimate the linear regression function taking (*i*) X as independent variable, (*ii*) Y as independent variable.
{ $\hat{y} = 6.32 + 0.84x$; $\hat{x} = -2.67 + 0.583y$ }

(b) Given the following data:

$$n = 5, \quad \Sigma x_i = 15, \quad \Sigma y_i = 25, \quad \Sigma(x_i - \bar{x})^2 = 10, \quad \Sigma(y_i - \bar{y})^2 = 26,$$

$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 13$. Determine the two regression lines.
($\hat{y} = 1.1 + 1.3x$; $\hat{x} = 0.5 + 0.5y$)

(c) The correlation coefficient between the two variables X and Y is $r = 0.60$. If $s_x = 1.50$, $s_y = 2.00$, $\bar{x} = 10$ and $\bar{y} = 20$, find the equations of the two regression lines of Y on X and X on Y.
($\hat{y} = 12 + 0.8x$; $\hat{x} = 1 + 0.45y$)

◆

Exercise 14.2
Objective Questions

1. Fill in the blanks.

(*i*) The ———— is a relationship that describes the dependence of the expected value of the dependent random variable for a given value of the independent non-random variable. (*regression*)

(*ii*) The variable, that forms the basis of estimation, is called ————. (*regressor*)

(*iii*) The variable, whose resulting value depends upon the selected value of the independent variable, is called ————. (*regressand*)

(*iv*) The ———— diagram is a set of points in a rectangular co-ordinate system, where each point represents an observed pair of values. (*scatter*)

(*v*) The principle of least squares is used for finding the ———— a and b of the parameters α and β. (*estimates*)

(*vi*) The ———— regression line always passes through $(\bar{x}, \bar{y})$. (*estimated*)

2. Mark the statements as true or false.

(*i*) The simple linear regression model contains two parameters α and β. (*false*)

(*ii*) The simple linear regression model contains four parameters α, β, $\mu_{Y|x}$ and σ^2. (*true*)

(*iii*) The simple linear regression model is simple in that there is only one independent variable. (*true*)

(*iv*) The parameter α is called the slope and the parameter β is the intercept of the regression line. (*false*)

(*v*) The regression coefficient is denoted by α. (*false*)

(*vi*) The parameter α is called the y-intercept of the regression line. (*true*)

(*vii*) In a regression analysis the independent variable is always prefixed while the dependent variable is random. (*true*)

(viii) The principle of least squares says that the sum of squares of the residuals of observed values from their corresponding estimated values should be the least possible. (*true*)

(*ix*) The principle of least squares is used for finding the estimates a and b of the parameters α and β. (*true*)

(*x*) The constant b estimates the parameter β representing the slope of the regression line. (*true*)

(*xi*) The regression coefficient b is independent of change in origin and scale. (*false*)

(*xii*) The estimated regression equation of Y on x is used to estimate the mean value of Y for a given value of x. (*true*)

3. Fill in the blanks.

(*i*) The correlation analysis is possible when both the variables X and Y are ————. (*random*)

(*ii*) If the two variables move in the ———— direction, the correlation is positive. (*same*)

(*iii*) If the two variables move in ———— directions, the correlation is negative. (*opposite*)

(*iv*) The correlation coefficient is ———— of the change in origin and unit of measurement. (*independent*)

(*v*) The correlation coefficient r is the ———— mean of the two regression coefficients. (*geometric*)

(*vi*) $r = 0$ indicates that the two variables are linearly ———— (*independent*)

4. Mark off the following statements true or false.

(*i*) The strength of covariability between two random variables is called correlation. (*true*)

(*ii*) The sample correlation coefficient R is a point estimator of the population correlation coefficient ρ. (*true*)

(*iii*) The correlation coefficient r is not symmetrical with respect to X and Y. (*false*)

(*iv*) The correlation coefficient changes with a change in origin. (*false*)

(*v*) The correlation coefficient is not affected by change in origin. (*true*)

(*vi*) The correlation coefficient is not independent of the origin and the unit of measurement. (*false*)

(*vii*) The correlation coefficient is a pure number which is unitless. (*true*)

(*xiii*) The correlation coefficient r always lies between -1 and 1. (*true*)

(*xiv*) $r = 1$ indicates perfect positive correlation between the two variables and slope is positive. (*true*)

(*xv*) $r = -1$ indicates perfect negative correlation between the two variables and slope is negative. (*true*)

5. Mark off the following statements true or false.

(*i*) $r = 0$ indicates that one regression line is horizontal and the other regression line is vertical. (*true*)

(*ii*) The correlation coefficient r is the geometric mean of the two regression coefficients. (*true*)

(*iii*) Each of the two estimated regression lines passes through the point $(\bar{x}, \bar{y})$. (*true*)

(*iv*) We can always estimate the X and Y values from the regression equation of Y on X. (*false*)

(*v*) The regression coefficient of X on Y is -1.2 and of Y on X is 0.3. (*false*)

(*vi*) The regression coefficient of X on Y is -1.2 and of Y on X is -0.3. (*true*)

(*vii*) The regression coefficient of X on Y, regression coefficient of Y on X and correlation coefficient have same sign. (*true*)

(*viii*) If the regression coefficient of X on Y is -1.2 and of Y on X is -0.3, the correlation coefficient is 0.6. (*false*)

(*ix*) The regression coefficient of X on Y is always equal to the regression coefficient of Y on X. (*false*)