UNIT 17 CORRELATION AND REGRESSION

Structure

17.1 INTRODUCTION

So far we have been dealing with the distributions of the data involving only one variable. Such a distribution is called a univariate distribution. Very often, we have to deal with the situations where more than one variables are involved. For example, we may like to study the relationship between the heights and weights of adult males, quantum of rainfall and the yield of wheat in India over a number of years, doses of drug and a response viz.a dose of insulin and blood sugar levels in a person, the age of individuals and their blood pressure, etc.

In such situations, our main purpose is to determine whether or not a relationship exists between the two v iriables. If such a relationship can be expressed by a mathematical formula, then we shall be able to use it for an analysis and hence make certain predictions.

Correlation and regression are methods that deal with the analysis of such relationships between various variables and possible predictions. In this unit, we shall confine ourselves to analysing the linear relationship between two variables. However, we can extend the methods for two variables to the situations where more than two variables are studied simultaneously.

Objectives

After reading this, you should be able to

- describe the correlation between two variables
- compute and interpret correlation coefficient
- describe simple linear regression line, and
- explain how to fit a linear regression line using least squares method.

17.2 CORRELATION AND SCATTER DIAGRAM

In studying the linear relationship between two variables, we try to examine the question "Are the two variables mutually related to each other?" In other words, we may ask whether the changes in one variable are accompanied by some corresponding changes in the other variable. For example, to find the relationship between the heights and weights of 100 persons, we can arrange them in increasing order of their heights and see whether or not the weight increases as the height increases. In other words, we are asking, "Do taller people tend to weigh more than shorter people?" Note carefully that we are not saying that if an individual is taller than another, he has to necessarily weigh heavier. Very seldom, a taller person may weigh less than a shorter person, but quite often taller persons have higher weights than shorter persons. That is, in general, we may expect to see that as the heights of 100 individuals are arranged in increasing order and the corresponding weights written down, the weights will show a tendency to increase.

In such a situation, two variables, then, are said to be mutually related or correlated. This process of mutual relationship is called **correlation** between two variables. Note that correlation need not be only in one direction. As one variable shows an increase, the second variable may show an increase or a decrease. We know, for example, as the altitudes of places increase, the atmospheric pressure decreases.

Hence, whenever two variables are related to each other in such a way that change in the one creates a corresponding change in the other, then the variables are said to be correlated.

An easy way of studying the correlation of two quantitative variables is to plot them on a graph sheet taking one of the variables on the X-axis and the other on the Y-axis. The resulting diagram is called a **scatter diagram** because it shows how the pairs of observations are scattered on the graph sheet. Note that the points representing the values of x and y may lie very close to a straight line. This means that we can approximate the relationship between the values of x and y by a straight line or by some other geometrical curve. If this is a straight line, then we say that the relationship between x and y is linear. The relationship between x and y may be a curve other than a straight line. The study of such relationships is beyond the scope of the present syllabus. Hence, we shall confine our discussion to the linear relationship between x and y.

Consider the following five artificial sets of 10 pairs of observations each, listed in Table 1. The two variables in each set are designated as X and Y.

TABLE 1

		1	2	3	4	5	6	7	8	9	10
Set I	X :	5	10	20	15	6	18	17	12	15	25
	Y :	20	8	10	5	12	16	22	17	12	17
Set II	X :	5	10	7	15	17	12	20	25	22	15
	Y :	4	6	9	12	27	14	17	24	21	17
Set III	X :	5	10	8	15	13	18	20	23	25	27
	Y :	25	22	17	20	15	13	17	12	19	10

Now plot the values of X and Y as (x,y) coordinates in the X-Y plane, as shown on next page in Fig. 1.

The Figure 1 presents the scatter diagrams of these 3 sets of data, set by set. Study these scatter diagrams carefully.

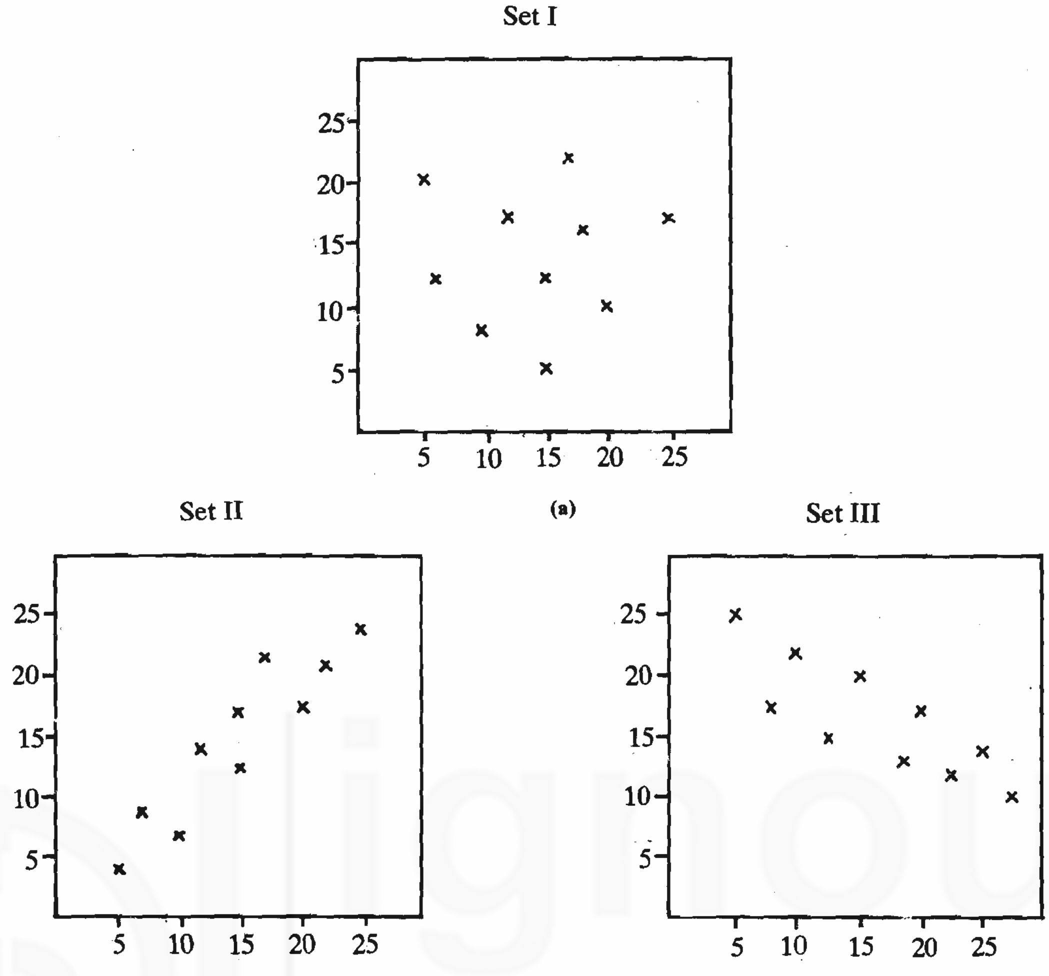
(i) Set I represents a situation in which we do not see any relationship between the values of X and Y. High values of X do not appear to be associated with either high or low values of Y. This indicates that the sample values of X and Y vary independently or that there is no evidence of a relationship between X and Y. When such a relationship is absent, as in the case of the scatter diagram of Set I (Fig. 1 (a)), then we say that the two variables are **uncorrelated (or not correlated).**

Note that in studying the relationship between X and Y, we are studying pairs of values of X and Y.

(ii) Set II indicates relationship of a particular type. As X increases, Y also tends to increase. In other words, high values of X tend to pair with high values of Y and low values of X pair with low values of Y. The scatter diagram of Set II, as in Fig.1 (b), gives us an evidence of what we call a "**positive correlation**" between X and Y. Thus, we have the following definition.

Definition : If two variables deviate in the same direction simultaneously, then they are said to be positively correlated.

In other words, if the two variables increase or decrease simultaneously (i.e., when one increases, the other also increases or when one decreases, the other also decreases), then the correlation between the two variables is said to be a **positive correlation.** In this case as shown in Fig. 1 (c), the points follow a line of positive



Set I

(a)

Set II

(b)

Set III

(c)

Figure I

slope. For example, the correlation between heights and weights of a group of persons is a positive correlation.

(iii) In the scatter diagram of Set III (Fig. (1 (c)), it is seen that high values of X pair with low values of Y and low values of X pair with high values of Y. In such a case, we say that X and Y are **negatively correlated.** In other words, the scatter diagram of Set III gives the negative correlation. Thus, we have the following definition:

Definition : If two variables deviate in opposite direction, then they are said to be negatively correlated or inversely correlated.

In other words, if the increase in one variable creates a decrease in the other, or the decrease in one creates an increase in the other, then the correlation between the two variables is said to be **negative correlation.** In this case, the points of the scatter diagram follow a line of negative slope.

For example, the correlation between the price and demand of a commodity is a negative correlation.

Thus, you see how useful the scatter diagram is to study the correlation between two variables. Now, you should try the following exercise:

E 1) Plot the following two sets of paired variables in a scatter diagram and comment on their relationship.

Set I :	X :	7	7	1	6	5	3	8	9	3	1
	Y :	5	9	6	1	3	1	9	4	6	8
Set II :	X :	9	7	6	1	3	1	9	4	6	8
	Y :	1	3	5	6	9	8	6	4	1	9

Thus, we have seen that correlation between two variables may or may not exist. In case, there exists a correlation between two variables then it is either a positive correlation or a negative correlation. How to measure such a correlation, if it exists? The most popular method used for this purpose is the **Pearson Product-Moment Correlation Coefficient** method due to Karl Pearson, a noted statistician as already mentioned in Unit 14.

17.3 CORRELATION COEFFICIENT

In the previous section, we saw how a scatter diagram helps us visually to judge whether pairs of values of two variables are correlated with each other. If related, whether positively or negatively. It would, however, be essential to measure the strength of this correlation in order that we may, compare two sets of variables both indicating positive or negative correlation.

As a first step in this direction, we define a measure called **'Co-Variance'**. You have already studied the term **'variance'** in Unit 12. You proved that the variance of a variable x is defined as the average sum of the squares of deviations of the values of the variable from its mean value. Algebraically, it is given by the formula,

$$\text{Variance (x)} = \frac{\Sigma\,(x - \bar{x})^2}{n}$$

where $\bar{x}$ is the mean of the sample and 'n' the number of observations in the sample.

We now define co-variance (analogously) as

$$\text{Cov (x,y)} = \text{Co-variance (x,y)} = \frac{\Sigma(x - \bar{x})\ (y - \bar{y})}{n}.$$

Let us understand this quantity carefully by means of the following illustrations:

Take 10 pairs of values of the two variables as

x : $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$
y : $y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8, y_9, y_{10}$

where (x_1, y_1) forms the first pair (x_2, y_2) forms the second pair and so on, till the 10th pair (x_{10}, y_{10}).

We can measure the deviation of each x value from its mean $\bar{x}$ and the deviation of each y value from its mean $\bar{y}$. If x and y vary together positively, then whenever $(x-\bar{x})$ is positive, $(y-\bar{y})$ will also be positive, and whenever $(x-\bar{x})$ is negative, $(y-\bar{y})$ will also be negative. However, if x and y vary together negatively, then whenever $(x-\bar{x})$ is positive, $(y-\bar{y})$ will be negative and vice-versa. Then you take the sum of the products of $(x-\bar{x})$ and $(y-\bar{y})$. Since there are 10 pairs of x and y values, we divide this by 10 to get an average measure of covariance as

$$\frac{\Sigma(x-\bar{x})\quad(y-\bar{y})}{10}.$$

In general, when n pairs of values of the two variables x and y are given, then the covariance is given as

$$\text{Cov. (x, y)} = \frac{\Sigma\,(x - \bar{x})\ (y - \bar{y})}{n}.$$

Covariance is a direct measure of correlation between two variables, but it cannot be used for meaningfully measuring the strength of the relationship between the two variables. This is because covariance can take all values from an infinitely large negative value to an infinitely large positive value. In the language of mathematics, it can vary from $-\infty$ (−infinity) to $+\infty$ (+ infinity).

We would, however, like to measure the strength of the relationship between two variables by means of a single number. The correlation coefficient is such a number with the property that its value will vary from −1 to +1, taking the value of −1 when the relationship between the two variables is perfectly negative, the value of +1 when the relationship is perfectly positive (refer back to Fig. 1 of this unit), and the value of 0 when there is no relationship.

In other words, a value r = − 1 will occur when all the points lie exactly on a straight line having a negative slope. The value r = + 1 will indicate that all points lie on a straight line have positive slope. If r is close to + 1 or − 1, then we say that the linear relationship (related by a linear equation) between the two variables is strong and we have high correlation. However, if r is close to zero, then we say that the relationship between the two variables is weak or perhaps even does not exist.

The correlation coefficient is denoted by r and is computed by the formula

$$r = \frac{\text{Covariance between x and y}}{\sqrt{(\text{Variance of X})\,(\text{Variance of Y})}}$$

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

$$= \frac{\Sigma\,(x-\bar{x})\,(y-\bar{y})}{\sqrt{(x-x)^2\;\Sigma\,(y-y)^2}}\,.$$

Note the symbol 'r' is used to denote the correlation coefficient calculated from a sample of a population (see Unit 15). The symbol for the correlation coefficient in the population is ρ (pronounced as 'Rho'). In other words, r is usually used for the **Sample Correlation Coefficient** which means a value computed from sample of n pairs of measurements whereas ρ is generally referred to as the **Population Correlation Coefficient** which means a linear correlation coefficient for the entire population. Sometimes, the correlation coefficient r is also mentioned as the product-moment correlation coefficient to distinguish it from the notation ρ in the following way:

The correlation coefficient 'r' between n pairs of observations whose values are $(x_1, y_1)\,(x_2, y_2)\,(x_3, y_3) \ldots (x_n, y_n)$ is

$$r = \frac{\sum_{i=1}^{n} (x_i-\bar{x})\;(y_i-\bar{y})}{\sqrt{[\,\overset{n}{\Sigma}\,(x_i-\bar{x})^2\,]\;\;[\,\Sigma\,(y_i-\bar{y})^2\,]}}$$

The correlation coefficient between the two random variables X and Y is

$$\rho = \frac{\text{Cov.}\,(x, y)}{\sqrt{\text{Var}\,(x)\;\text{Var}\,(y)}}\,.$$

In any case the two are equivalent to each other. For computational purposes, 'r' can be writtten as

$$r = \frac{\Sigma\,xy - \frac{\Sigma x\,.\,\Sigma y}{n}}{\sqrt{[\,\Sigma x^2 - \frac{(\Sigma x)^2}{n}\,]\,.\,[\Sigma y^2 - \frac{(\Sigma y)^2}{n}]}}$$

where

i) Σxy is the sum of products of the 'n' pairs of x and y,
ii) Σx^2 and Σy^2 are the sums of squares of x's and y's respectively,
iii) Σx and Σy are the sums of x and y respectively,
iv) 'n' is the number of pairs of observation.

Often this correlation coefficient is referred to as Pearson's correlation coefficient.

There is one important point to be noted in using the Pearson's correlation coefficient. **It is applicable** only **when** the **relationship** between the two variables under consideration is linear or in other words the two variables have a straight line relationship. If the relationship between two variables is non-linear or curvilinear, Pearson's correlation coefficient should not be used to measure the strength of the relationship. The scatter diagram, which we learnt in Section 17.2, is very helpful to decide whether the relationship is linear or not.

Example 1 : Compute the correlation coefficient for the following data:

x	12	10	14	11	12	9
y	18	17	23	19	20	15

Solution: From the data, we have

$\sum_{i=1}^{6} x_i = 68$, $\sum_{i=1}^{6} y_i = 112$, $\sum_{i=1}^{6} x_i y_i = 1292$ $\sum_{i=1}^{6} x_i^2 = 786$, $\sum_{i=1}^{6} y_i^2 = 2128$

Therefore, $r = 0.947$ (calculate yourself)

Try the following exercises:

E 2) Compute Pearson's correlation coefficient for the data given in Set II of Exercise E 1) using the computational formula for r.

E 3) Draw the scatter diagram of the following data and comment if it is valid to represent the relationship seen by a Pearson's correlation coefficient.

x :	2	3	4	5	6	7	8	9
y :	9	12	15	18	14	11	9	7

17.4 REGRESSION

In the previous section you have seen that the data giving the corresponding values of two variables can be graphically represented by a scatter diagram. Also, you were introduced to a method of finding the relationship between these two variables in terms of the correlation coefficient. Very often, in the study of relationship of two variables, we come across instances where one of the two variables depends on the other. In other words, what is the possible value of the dependent variable when the value of independent variable is known. For example, the bodyweight of a growing child depends on the nutrient intake of the child or the weight of an individual may be dependent on his height or the response ʋ a drug can be dependent on the dose of the drug or the agricultural yield may depena on the quantum of rainfall. In such situations, where one of the variables is dependent and the other independent, you may ask "can we find a method of estimating the numerical relationship between two variables so that given a value of the independent variable, we can predict the average value of the dependent variable?".

Note that we are trying to predict or estimate the average value of the dependent variable for a given value of the independent variable. We cannot determine the exact value of the dependent variable when the value of the independent variable is known. What perhaps we can do is just to make an estimation of the value of the dependent variable, knowing fully well that there could be an error in our estimation. This is because of the reason that there is no certainty that the estimated value of the variable would be exactly the same as the value actually observed. This is also because for a given value of the independent variable, the dependent variable will usually show some variations in its values. For example, not all persons of a given height, say of 5′ 6″ have the same weight. Some will be heavier than others. This is why we talk of predicting the average value of the dependent variable for a given value of the independent variable.

Let us consider another situation, where it is not logically meaningful to consider one variable as dependent on the other. The heights of brothers and sisters, we expect, will be related. However, it would not be logical to consider the height of one member of a pair to be dependent on the height of the other member. The height of a sister is not dependent on the height of the brother while both of them may be related through the heights of their parents.

When two variables are so related that it is logical to think of estimating the value of one variable as dependent on the other, then the relationship can be described by what is called a **Regression Function** or **Regression Equation,** a term introduced by Sir Francis Galton (1822-1911).

The words 'dependent' and 'independent' are used in the study of probability with different meaning now-a-days. Therefore, in the context of regression, the dependent variable is called the "outcome" variable and independent variable is called the "predictor" variable. We shall stick to this terminology in studying regression, since in regression, our interest is in prediction. Hence Y will be called the outcome variable, and X the predictor variable.

The simplest form of the regression function relating the predictor and outcome variables is that of a straight line as shown in Fig. 2 below. That is, the nature of relationship between Y and X is linear.

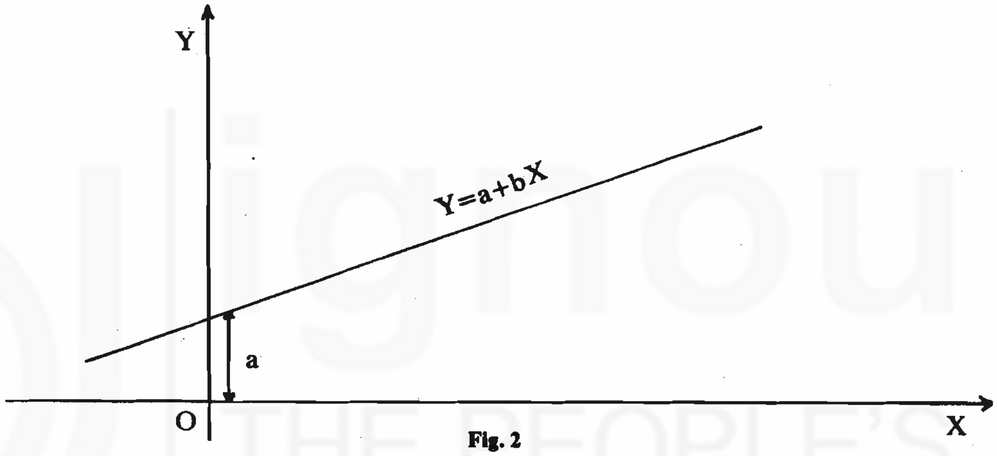


Fig. 2

In this situation, the regression function takes the form, Y = a + bX where Y is the outcome variable, X the predictor variable, and a and b are constants to be estimated from a given set of data. Figure 3 below presents a straight line relationship.

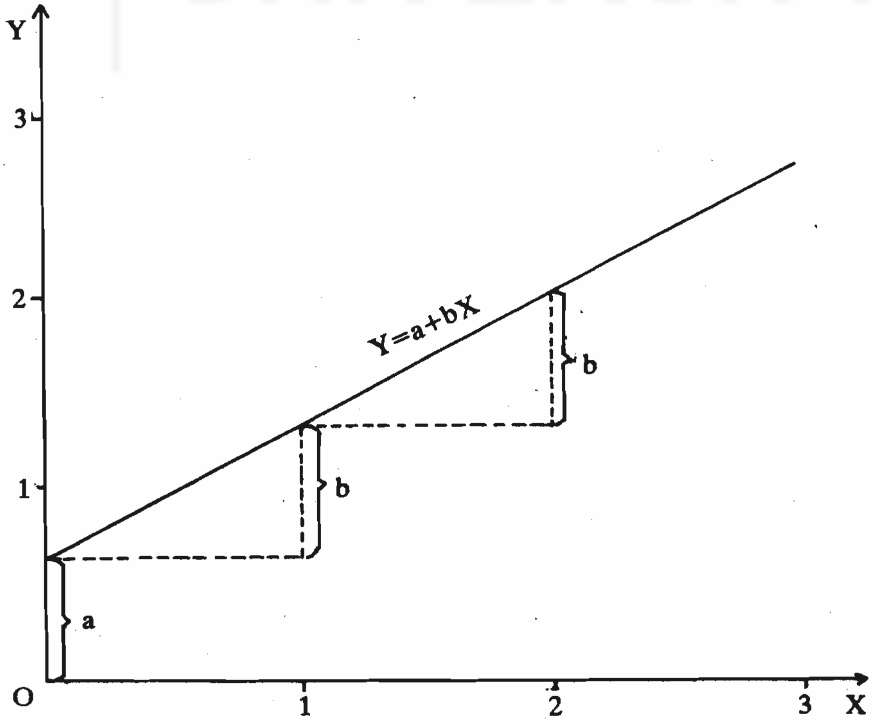


Fig. 3

In the Fig. 3

i) 'a' is the intercept that the straight line Y = a + bX makes on the Y axis, that is, Y = a, when X = 0.

ii) 'b' is the slope of the line, that is, the increase in the value of Y for every unit increase in the value of X.

Note that b is marked on the graph as the height that the straight line gains corresponding to an increase of one unit in X value. Further, b is positive when Y increases when X increases and is negative when Y decreases as X increases or Y decreases as X increases. If Y takes the value 0 when X is 0, then 'a' becomes equal to 0 so that the regression line passes through the origin—that is where both X and Y are 0. The regression equation then is written as

Y = bX

Very often, the results of an investigation (either by an experiment or by a survey) may be summarized and approximated by a relation or an equation which represents a line. In the foregoing example, a line of the form y = a + bX has been used to approximate the given data. Thus there is a problem of finding equations to straight lines which fit the given data. We shall, therefore, try to find the solution to this problem in the next section.

17.5 FITTING OF LINEAR REGRESSION

The regression line Y = a + bX has to be estimated from a given set of n pairs of observations of X and Y. We have seen earlier that the first essential step in trying to understand the nature of the relationship between two variables is to plot them on a scatter diagram. Consider the following 16 pairs of observations of heights and weights of 16 adult males.

Table 2

Heights (cm) and Weights (kg) of 16 Males

Serial No. of individual	Height (cm)	Weight (kg.)
1	178	80
2	176	75
3	170	72
4	174	74
5	165	68
6	162	64
7	178	76
8	165	66
9	174	72
10	172	70
11	160	63
12	162	61
13	160	60
14	170	69
15	176	72
16	172	67

These 16 pairs of observations are presented in a scatter diagram in Figure 4 along with the best fitting line described in Figure 4. You can easily see that the relationship between the two variables is linear.

We would now like to fit a straight line to this set of points on the scatter diagram. One way to do it is to fit a straight line by visual judgement. However, since this is subjective, we may not at all fit the same straight line to the given set of data. In order to avoid subjectivity and hence differences in deciding which line "best" fits the data we would like to adopt a mathematical approach which uniquely determines the "best" fitting straight line. The approach we are going to adopt is called the **Method of Least Squares**. The procedure is as follows:

Method of Least Squares: Let (x_1, y_1), (x_2, y_2) ... (x_n, y_n) denote the data points in the scatter diagram as given in Figure 4. For a given value of x say x_1, there will be a difference between the value y_1 and the corresponding value as determined from the straight line which fit to data. We denote this difference by d_1, which is generally called as the deviation. Of all the straight lines that can be fitted to this data, we define that line as the "best" which minimises the sum of squares of deviations of the individual points from the line, the deviations being measured along the Y-axis, i.e., if d_i denote the deviations then $\Sigma\, d_i^2$ is minimum. In Figure 4, the deviations of the points from the best fitting line are marked in dotted lines. Since we are attempting to fit a straight line of the form

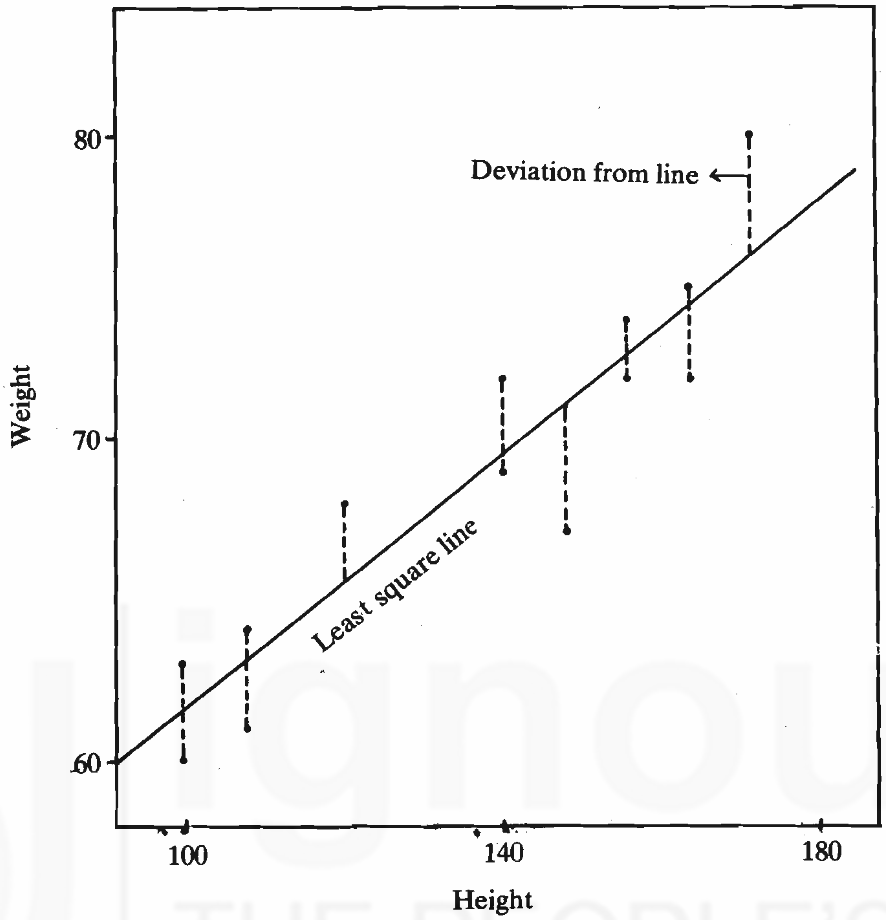


Fig. 4

$$y = a + bX$$

the deviations of the observed values of Y from the line have to be measured in the units in which the dependent variable Y is expressed. That is why in the diagram, you find the deviations of the observed Y values from the line marked (in dotted lines) in the direction of the Y axis. But then you may raise the following questions: "why should we sum the squares of these deviations and choose that line which minimises this sum? Why not just add up these deviations, some of which are positive and others negative and take that line which gives a minimum value to this sum?"

The reason why we do not take just the sum of these deviations is that more than one straight line may satisfy the condition of this sum having the same minimum value. For example, we may be able to fit two different straight lines to this data both of which may give the sum of the deviations as zero. We want that we should get a unique line by our method. Minimising the sum of squares of the deviations will give a unique line or only a single line. Let us express the above idea in simple algebraic terms.

We want $(Y - a - bX)^2$ for the 16 pairs of observations in our example to be minimum. In the equation $Y = a + bX$ we can substitute the values of X and Y but we have to estimate the two unknown constants a and b.

That is, we want to find the value of a and b such that $(Y - a - bX)^2$ is a minimum. This can be done by using calculus.

We are not actually doing this here but we only give the expressions for a and b which minimise $(Y - a - bX)^2$. The estimates of a and b are

$$a = \frac{\Sigma y}{n} - b.\frac{\Sigma x}{n} = \bar{y} - b\bar{x}$$

$$b = \frac{\Sigma (x - \bar{x})\ (y - \bar{y})}{\Sigma (x - x)^2}$$

$$= \frac{\Sigma\, xy - \dfrac{\Sigma x.\ \Sigma y}{n}}{\Sigma\, x^2 - \dfrac{(\Sigma x)^2}{n}}$$

Then calculate a by the formula

$$a = \frac{\Sigma y}{n} - b.\ \frac{\Sigma x}{n}$$

Thus a and b can be calculated from the given set of data. For the example given in Table 2 of 16 pairs of heights and weights, we have

$a = - 69.4935$; $b = 0.8183$

so that the regression equation is given by

$Y = - 69.4935 + 0.8183\ X.$

That is the line drawn on the scatter diagram in Figure 4. The Method of Least Squares as described here is the best method for fitting a straight line regression provided the following assumptions are satisfied:

1) For every value of X, the predictor variable, the value of Y have a mean that is given by the equation $\alpha + \beta x$ where α and β are parameters corresponding to a and b respectively that we estimated from the sample. In the example, we considered 16 pairs of individuals. If we had measured **all** the individuals in the population from which we got 16 pairs of observations, then corresponding to each height, the mean weight would be given by $\alpha + \beta x$.

2) The distribution of Y values around this mean $\alpha + \beta X$ corresponding to each X, will have the same standard deviation (i.e., the variability of Y remains constant at all values of X).

3) The distribution of Y will be Gaussian or normal as it is often called.

Let us try the following exercises:

E 4) The data below show the initial weight (grams) and gains in weight of 12 female rats on a high protein diet from the 20th to 80th day after their birth. Examine the data to decide if the gain in weight depends on the initial weight. Compute the linear regression equations of gain in weight on initial weight. Present your results on a graph taking initial weight on the X-axis and gain in weight on the Y-axis.

	Rat Number											
	1	2	3	4	5	6	7	8	9	10	11	12
Initial weight (gm)	45	50	69	68	76	46	59	74	54	62	64	48
Gain in weight (gm)	93	128	106	126	154	79	112	128	120	118	120	98

E 5) The data below give the percentage of fruits affected by worms and the corresponding number fruits on a sample of 14 apple fruit trees. Examine through fitting a linear regression to this data, if the percentage of fruits affected by worms is dependent on the number of fruits on the tree. Present the data on a graph paper.

	Tree number in the sample trees													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Number of fruits (in hundred)	5	8	9	15	12	6	10	8	7	16	4	11	13	14
Per cent affected by worms	60	44	41	11	28	54	37	46	52	4	65	32	24	17

Regression of Y on X and Regression of X on Y

In our earlier discussion, we have considered one of the variables, Y as the outcome variable and X as the predictor variable. What happens if we reverse the variables, i.e., what happens if we consider X as the outcome variable and Y as the predictor variable? We get a regression of the form

$$X = a' + b' Y$$

where a′ and b′ are different from a and b of the regression equation Y = a + bx. In other words, the regression of Y on X on the regression of X and Y are two different lines. It is easy to see why they are so. Recall that for fitting regression of Y on X we minimised $(Y - a - d X)^2$, Y being the observed X value, that is, we minimised the sum of squares of the deviations of the observed values from those predicted by the "best" regression line. These deviations, we noted, were measured along the Y axis or, in other words, the deviations were the vertical deviations from the line along the Y axis. If we fit the regression line of X on Y, then we will have to minimise **(X—a′—b′ Y)? (X being the observed Y value) or, in other words, minimise the deviations from the "best" line, measuring the deviations along the X-axis. We consider horizontal instead of vertical deviations which means the interchange of X-axis and Y-axis. That is, we get two different lines for regression of Y on X and X on Y.**

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

17.6 SUMMARY

In this unit, we have

1) described **scatter diagram** in studying the nature and strength of relationship between two quantitative variables,

2) described **positive correlation** and **negative correlation,**

3) defined the Pearson's correlation coefficient,

4) described the concept of statistical regression and distinguished it from the concept of correlation,

5) described the method of least squares for fitting a linear regression line to a given set of data and defined the regression coefficient,

6) distinguished between the two regression lines of regression of Y on X and regression of X on Y.

17.7 SOLUTIONS/ANSWERS

E 1) Set I : There is no relationship between X and Y
Set II : The relationship is linear and negative

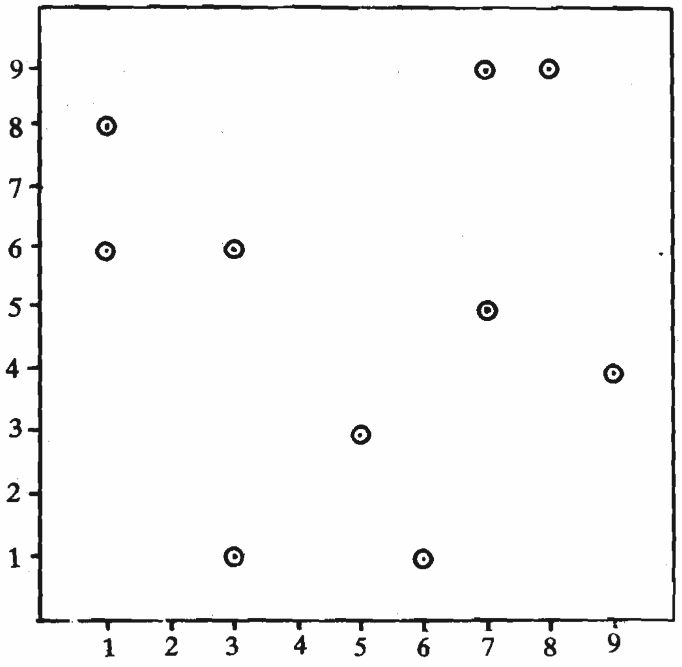


Fig. 5 (a)

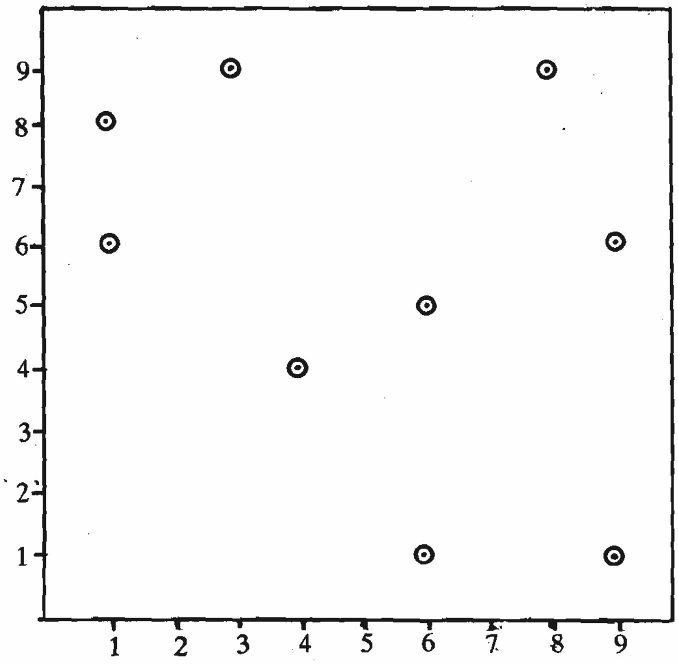


Fig. 5 (b)

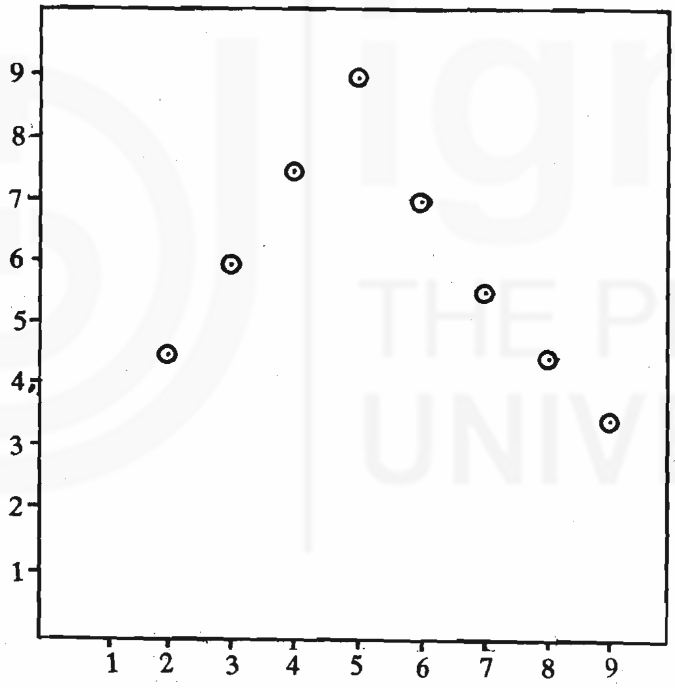


Fig. 5 (c)

E 2) The correlation coefficient for set II of E 1) is

$$r = \frac{\Sigma xy - \frac{\Sigma x . \Sigma y}{n}}{\sqrt{[\Sigma x^2 - \frac{(\Sigma x)^2}{n}].[\Sigma y^2 - \frac{(\Sigma y)^2}{n}]}}$$

$$= \frac{[185 - \frac{50 \times 51}{10}]}{\sqrt{[324 - \frac{50^2}{10}]\quad[333 - \frac{60^2}{10}]}}$$

= − 0.953. The correlation is negative.

E 3) The relationship between X and Y is non-linear and hence, Pearson's correlation coefficient is not applicable here.

E 4) $\Sigma x = 758$, $\Sigma y = 1359$, $\Sigma xy = 88914$, $\dfrac{\Sigma x . \Sigma y}{n} = 85843.5$

$\Sigma x^2 = 50068$, $\dfrac{(\Sigma x)^2}{n} = 47880.3333$, $n = 12$

$$b = \frac{\Sigma\, xy - \dfrac{\Sigma x.\ \Sigma y}{n}}{\sqrt{[\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}].\ [\Sigma y^2 - \dfrac{(\Sigma y)^2}{n}]}} = \frac{3070.5}{2187.6667} = 1.4036$$

$$a = \bar{y} - b\bar{x} = [\frac{\Sigma y}{n} - b\ \frac{\Sigma x}{n}] = 24.5924$$

The regression equation is $Y = 24.5924 + 14030X$

E 5) $\Sigma x = 138$, $\Sigma y = 515$, $\Sigma xy = 4174$, $\dfrac{\Sigma x\ \Sigma y}{n} = 5076.4286$

$\Sigma x^2 = 1546$, $\dfrac{(\Sigma x)^2}{n} = n = 14.$

$$b = \frac{\Sigma\, xy - \dfrac{\Sigma x . \Sigma y}{n}}{\sqrt{[\ \Sigma x^2 - \dfrac{(\Sigma x)^2}{n}].\ [\ \Sigma y^2 - \dfrac{(\Sigma y)^2}{n}\ ]}} = \frac{-702.4296}{185.7143}$$

$$= -\ 4.8572$$

$a = \bar{y} - b\bar{x} = 84.6838$

Regression equation $Y = 84.6838 - 4.8592\ X$

Note that the regression coefficient b is negative.