UNIT 10 CORRELATION AND SIMPLE REGRESSION

STRUCTURE

10.0 Objectives
10.1 Introduction
10.2 Correlation
  10.2.1 Scatter Diagram
10.3 The Correlation Coefficient
  10.3.1 Karl Pearson's Correlation Coefficient
  10.3.2 Testing for the Significance of the Correlation Coefficient
  10.3.3 Spearman's Rank Correlation
10.4 Simple Linear Regression
10.5 Estimating the Linear Regression
  10.5.1 Standard Error of Estimate
  10.5.2 Coefficient of Determination
10.6 Difference Between Correlation and Regression
10.7 Let Us Sum Up
10.8 Key Words
10.9 Answers to Self Assessment Exercises
10.10 Terminal Questions/Exercises
10.11 Further Reading
Appendix Tables

10.0 OBJECTIVES

After studying this unit, you should be able to:

- understand the concept of correlation,
- use scatter diagrams to visualize the relationship between two variables,
- compute the simple and rank correlation coefficients between two variables,
- test for the significance of the correlation coefficient,
- use the regression analysis in estimating the relationship between dependent and independent variables,
- use the least squares criterion to estimate the equation to forecast future values of the dependent variable,
- determine the standard errors of estimate of the forecast and estimated parameters,
- understand the coefficient of determination as a measure of the strength of the association between two variables, and
- distinguish between correlation and simple regression.

10.1 INTRODUCTION

In previous units, so far, we have discussed the statistical treatment of data relating to one variable only. In many other situations researchers and decision-makers need to consider the relationship between two or more variables. For example, the sales manager of a company may observe that the sales are not the same for each month. He/she also knows that the company's advertising expenditure varies from year to year. This manager would be interested in knowing whether a relationship exists between sales and advertising expenditure. If the manager could successfully define the relationship, he/she

might use this result to do a better job of planning and to improve predictions of yearly sales with the help of the regression technique for his/her company. Similarly, a researcher may be interested in studying the effect of research and development expenditure on annual profits of a firm, the relationship that exists between price index and purchasing power etc. The variables are said to be closely related if a relationship exists between them.

The correlation problem considers the joint variation of two measurements neither of which is restricted by the experimenter. The regression problem considers the frequency distribution of one variable (dependent variable) when another variable (independent variable) is held fixed at each of several intervals.

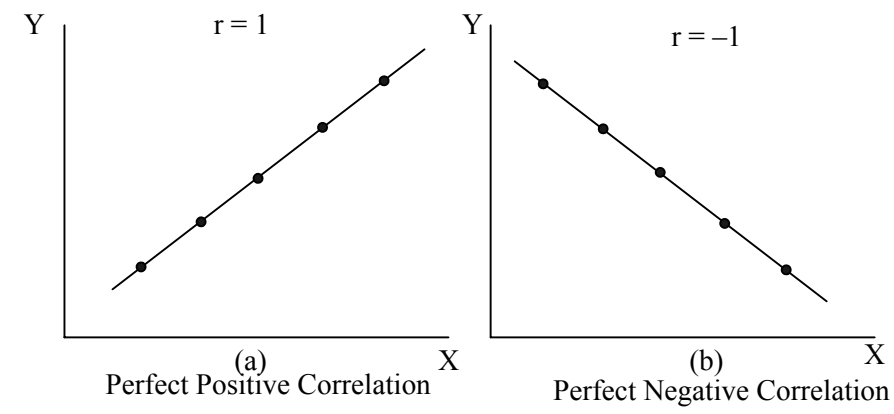
This unit, therefore, introduces the concept of correlation and regression, some statistical techniques of simple correlation and regression analysis. The methods used are important to the researcher(s) and the decision-maker(s) who need to determine the relationship between two variables for drawing conclusions and decision-making.

10.2 CORRELATION

If two variables, say x and y vary or move together in the same or in the opposite directions they are said to be correlated or associated. Thus, correlation refers to the relationship between the variables. Generally, we find the relationship in certain types of variables. For example, a relationship exists between income and expenditure, absenteesim and production, advertisement expenses and sales etc. Existence of the type of relationship may be different from one set of variables to another set of variables. Let us discuss some of the relationships with the help of Scatter Diagrams.

10.2.1 Scatter Diagram

When different sets of data are plotted on a graph, we obtain **scatter diagrams**. A scatter diagram gives two very useful types of information. Firstly, we can observe patterns between variables that indicate whether the variables are related. Secondly, if the variables are related we can get an idea of the type of relationship that exists. The scatter diagram may exhibit different types of relationships. Some typical patterns indicating different correlations between two variables are shown in Figure 10.1.



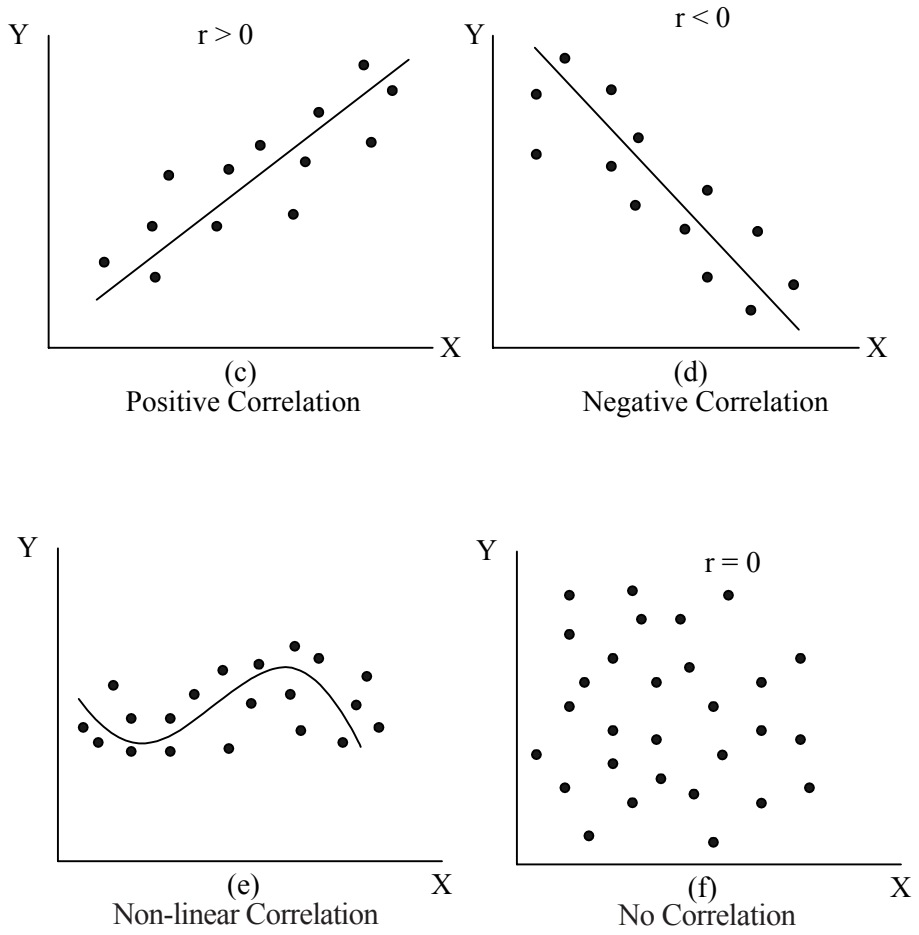


Figure 10.1 : Possible Relationships Between Two Variables, X and Y

If X and Y variables move in the same direction (i.e., either both of them increase or both decrease) the relationship between them is said to be **positive correlation** [Fig. 10.1 (a) and (c)]. On the other hand, if X and Y variables move in the opposite directions (i.e., if variable X increases and variable Y decreases or vice-versa) the relationship between them is said to be **negative correlation** [Fig. 10.1 (b) and (d)]. If Y is unaffected by any change in X variable, then the relationship between them is said to be **un-correlated** [Fig. 10.1 (f)]. If the amount of variations in variable X bears a constant ratio to the corresponding amount of variations in Y, then the relationship between them is said to be **linear-correlation** [Fig. 10.1 (a) to (d)], otherwise it is **non-linear or curvilinear correlation** [Fig. 10.1 (e)]. Since measuring non-linear correlation for data analysis is far more complicated, we therefore, generally make an assumption that the association between two variables is of the linear type.

If the relationship is confined to two variables only, it is called **simple correlation**. The concept of simple correlation can be best understood with the help of the following illustration which relates advertisment expenditure to sales of a company.

Illustration 1

Table 10.1 : A Company's Advertising Expenses and Sales Data (Rs. in crore)

Years :	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Advertise-ment expenses (X)	6	5	5	4	3	2	2	1.5	1.0	0.5
Sales (Y)	60	55	50	40	35	30	20	15	11	10

The company's sales manager claims the sales variability occurs because the marketing department constantly changes its advertisment expenditure. He/she is quite certain that there is a relationship between sales and advertising, but does not know what the relationship is.

The different situations shown in Figure 10.1 are all possibilities for describing the relationships between sales and advertising expenditure for the company. To determine the appropriate relationship, we have to construct a scatter diagram shown in Figure 10.2, considering the values shown in Table 10.1.

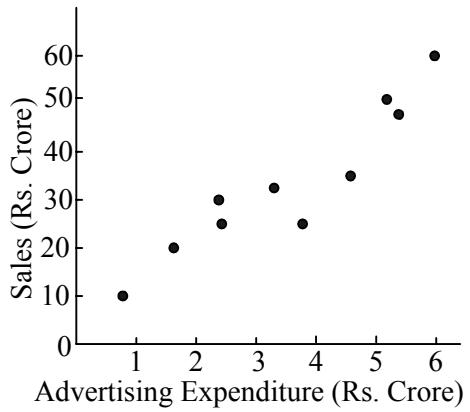


Figure 10.2 : Scatter Diagram of Sales and Advertising Expenditure for a Company.

Figure 10.2 indicates that advertising expenditure and sales seem to be linearly (positively) related. However, the strength of this relationship is not known, that is, how close do the points come to fall on a straight line is yet to be determined. The quantitative measure of strength of the linear relationship between two variables (here sales and advertising expenditure) is called the **correlation coefficient**. In the next section, therefore, we shall study the methods for determining the coefficient of correlation.

Self Assessment Exercise A

1) Suggest eight pairs of variables, four in each, which you expect to be positively correlated and negatively correlated

.....

.....

.....

2) How does a scatter diagram approach help in studying the correlation between two variables?

.....

.....

.....

10.3 THE CORRELATION COEFFICIENT

As explained above, the coefficient of correlation helps in measuring the degree of relationship between two variables, X and Y. The methods which are used to measure the degree of relationship will be discussed below.

10.3.1 Karl Pearson's Correlation Coefficient

Karl Pearson's coefficient of correlation (r) is one of the mathematical methods of measuring the degree of correlation between any two variables X and Y is given as:

$$r = \frac{\sum (X - \overline{X})\ (Y - \overline{Y})/n}{\sigma_X\ \sigma_Y}$$

The simplified formulae (which are algebraic equivalent to the above formula) are:

1) $r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}}$, where $x = X - \overline{X}, \quad y = Y - \overline{Y}$

Note: This formula is used when $\overline{X}$ and $\overline{Y}$ are integers.

2) $$r = \frac{\sum XY - \frac{\sum X . \sum Y}{n}}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{n}}\sqrt{\sum Y^2 - \frac{(\sum Y)^2}{n}}}$$

Before we proceed to take up an illustration for measuring the degree of correlation, it is worthwhile to note some of the following important points.

i) 'r' is a dimensionless number whose numerical value lies between +1 to –1. The value +1 represents a perfect positive correlation, while the value –1 represents a perfect negative correlation. The value 0 (zero) represents lack of correlation. Figure 10.1 shows a number of scatter plots with corresponding values for correlation coefficient.

ii) The coefficient of correlation is a pure number and is independent of the units of measurement of the variables.

iii) The correlation coefficient is independent of any change in the origin and scale of X and Y values.

Remark: Care should be taken when interpreting the correlation results. Although a change in advertising may, in fact, cause sales to change, the fact that the two variables are correlated does not guarantee a cause and effect relationship. Two seemingly unconnected variables may often be highly correlated. For example, we may observe a high degree of correlation: (i) between the height and the income of individuals or (ii) between the size of the

shoes and the marks secured by a group of persons, even though it is not possible to conceive them to be casually related. When correlation exists between such two seemingly unrelated variables, it is called **spurious or non-sense correlation**. Therefore we must avoid basing conclusions on spurious correlation.

Illustration 2

Taking as an illustration, the data of advertisement expenditure (X) and sales (Y) of a company for 10 years shown in Table 10.1, we proceed to determine the correlation coefficient between these variables.

Solution: **Table 10.2 : Calculation of Correlation Coefficient**

(Rs. in crore)

Advertisement expenditure Rs. (X)	Sales Rs. (Y)	XY	X^2	Y^2
6	60	360.0	36	3600
5	55	275.0	25	3025
5	50	250.0	25	2500
4	40	160.0	16	1600
3	35	105.0	9	1225
2	30	60.0	4	900
2	20	40.0	4	400
1.5	15	22.5	2.25	225
1.0	11	11.0	1	121
0.5	10	5.0	0.25	100
$\Sigma X = 30$	$\Sigma Y = 326$	$\Sigma XY = 1288.5$	$\Sigma X^2 = 122.50$	$\Sigma Y^2 = 13696$

We know that

$$r = \frac{\sum XY - \frac{\sum(X)\sum(Y)}{n}}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{n}}\sqrt{\sum Y^2 - \frac{(\sum Y)^2}{n}}}$$

$$= \frac{\frac{1288.5 - (30)(326)}{10}}{\sqrt{122.5 - \frac{(30)^2}{10}}\sqrt{13696 - \frac{(326)^2}{10}}} = \frac{310.5}{315.7}$$

$$= 0.9835$$

The calculated coefficient of correlation r = 0.9835 shows that there is a high degree of association between the sales and advertisement expenditure. For this particular problem, it indicates that an increase in advertisement expenditure is likely to yield higher sales. If the results of the calculation show a strong correlation for the data, either negative or positive, then the line of best fit to

that data will be useful for forecasting (it is discussed in Section 10.4 on 'Simple Linear Regression').

You may notice that the manual calculations will be cumbersome for real life research work. Therefore, statistical packages like minitab, SPSS, SAS, etc., may be used to calculate 'r' and other devices as well.

10.3.2 Testing for the Significance of the Correlation Coefficient

Once the coefficient of correlation has been obtained from sample data one is normally interested in asking the questions: Is there an association between the two variables? Or with what confidence can we make a statement about the association between the two variables? Such questions are best answered statistically by using the following procedure.

Testing of the null hypothesis (testing hypothesis and t-test are discussed in detail in Units 15 and 16 of this course) that population correlation coefficient equals zero (variables in the population are uncorrelated) versus alternative hypothesis that it does not equal zero, is carried out by using t-statistic formula.

$t = r\sqrt{\frac{n-2}{1-r^2}}$, where, r is the correlation coefficient from sample.

Referring to the table of t-distribution for (n–2) degree of freedom, we can find the critical value for t at any desired level of significance (5% level of significance is commonly used). If the calculated value of t (as obtained by the above formula) is less than or equal to the table value of t, we accept the null hypothesis (H_0), meaning that the correlation between the two variables is not significantly different from zero.

The following example will illustrate the use of this test.

Illustration 3

Suppose, a random sample of 12 pairs of observations from a normal population gives a correlation coefficient of 0.55. Is it likely that the two variables in the population are uncorrelated?

Solution: Let us take the null hypothesis (H_0) that the variables in the population are uncorrelated.

Applying t-test,

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.55\sqrt{\frac{12-2}{1-0.55^2}}$$

$$= 0.55 \times 3.786 = 2.082$$

From the t-distribution (refer the table given at the end of this unit) with 10 degrees of freedom for a 5% level of significance, we see that the table value of $t_{0.05/2,\,(10-2)} = 2.228$. The calculated value of t is less than the table value of t. Therefore, we can conclude that this r of 0.55 for n = 12 is not significantly different from zero. Hence our hypothesis (H_0) holds true, i.e., the sample variables in the population are uncorrelated.

Let us take another illustration to test the significance.

Illustration 4

A random sample of 100 pairs of observations from a normal population gives a correlation coefficient of 0.55. Do you accept that the variables in the population are correlated?

Solution: Let us take the hypothesis that the variables in the population are uncorrelated. Apply the t-test:

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.55\sqrt{\frac{100-2}{1-0.55^2}}$$

$$= 6.52$$

Referring to the table of the t-distribution for n–2 = 98 degrees of freedom, the critical value for t at a 5% level of significance $[t_{0.05/2,\,(10-2)}] = 1.99$ (approximately). Since the calculated value of t (6.52) exceeds the table value of t (1.99), we can conclude that there is statistically significant association between the variables. Hence, our hypothesis does not hold true.

10.3.3 Spearman's Rank Correlation

The Karl Pearson's correlation coefficient, discussed above, is not applicable in cases where the direct quantitative measurement of a phenomenon under study is not possible. Sometimes we are required to examine the extent of association between two ordinally scaled variables such as two rank orderings. For example, we can study efficiency, performance, competitive events, attitudinal surveys etc. In such cases, a measure to ascertain the degree of association between the ranks of two variables, X and Y, is called **Rank Correlation**. It was developed by Edward Spearman, its coefficient (R) is expressed by the following formula:

$$R = 1 - \frac{6\sum d^2}{N^3 - N}$$ where, N = Number of pairs of ranks, and Σd^2 = squares of difference between the ranks of two variables.

The following example illustrates the computation of rank correlation coefficient.

Illustration 5

Salesmen employed by a company were given one month training. At the end of the training, they conducted a test on 10 salesmen on a sample basis who were ranked on the basis of their performance in the test. They were then posted to their respective areas. After six months, they were rated in terms of their sales performance. Find the degree of association between them.

Salesmen:	1	2	3	4	5	6	7	8	9	10
Ranks in training (X):	7	1	10	5	6	8	9	2	3	4
Ranks on sales Peformance (Y):	6	3	9	4	8	10	7	2	1	5

Solution: **Table 10.3: Calculation of Coefficient of Rank Correlation.**

Salesmen	Ranks Secured in Training X	Ranks Secured on Sales Y	Difference in Ranks D = (X–Y)	D^2
1	7	6	1	1
2	1	3	–2	4
3	10	9	1	1
4	5	4	1	1
5	6	8	–2	4
6	8	10	–2	4
7	9	7	2	4
8	2	2	0	0
9	3	1	2	4
10	4	5	–1	1
				$\Sigma D_2 = 24$

Using the Spearman's formula, we obtain

$$R = 1 - \frac{6\sum D^2}{N^3 - N} = 1 - \frac{6 \times 24}{10^3 - 10}$$

$$= 1 - \frac{144}{990} = 0.855$$

we can say that there is a high degree of positive correlation between the training and sales performance of the salesmen.

Now we proceed to test the significance of the results obtained. We are interested in testing the null hypothesis (H_0) that the two sets of ranks are not associated in the population and that the observed value of R differs from zero only by chance. The test which is used is t-statistic.

$$t = R\sqrt{\frac{n-2}{1-R^2}} = 0.855\sqrt{\frac{10-2}{1-0.855^2}}$$

$$= 0.855\sqrt{29.74} = 4.663$$

Referring to the t-distribution table for 8 d.f (n–2), the critical value for t at a 5% level of significance [$t_{0.05/2, (10-2)}$] is 2.306. The calculated value of t is greater than the table value. Hence, we reject the null hypothesis concluding that the performance in training and on sales are closely associated.

Sometimes the data, relating to qualitative phenomenon, may not be available in ranks, but values. In such a situation the researcher must assign the ranks to the values. Ranks may be assigned by taking either the highest value as 1 or the lowest value as 1. But the same method must be followed in case of both the variables.

Tied Ranks

Sometimes there is a tie between two or more ranks in the first and/or second series. For example, there are two items with the same 4th rank, then instead of awarding 4th rank to the respective two observations, we award 4.5 (4+5/2) for each of the two observations and the mean of the ranks is unaffected. In such cases, an adjustment in the Spearman's formula is made. For this, Σd^2 is increased by $\frac{(t^3 - t)}{12}$ for each tie, where t stands for the number of observations in each tie. The formula can thus be expressed as:

$$r = 1 - \frac{6\left(\sum d^2 + \frac{t^3 - t}{12} + \frac{t^3 - t}{12} + \ldots\right)}{N^3 - N}$$

Self Assessment Exercise B

1) Compute the degree of relationship between price of share (X) and price of debentures over a period of 8 years by using Karl Pearson's formula and test the significance (5% level) of the association. Comment on the result.

Years:	1996	1997	1998	1999	2000	2001	2002	2003
Price of shares:	42	43	41	53	54	49	41	55
Price of debentures:	98	99	98	102	97	93	95	94

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

2) Consider the above exercise and assign the ranks to price of shares and price of debentures. Find the degree of association by applying Spearman's formula and test its significance.

.....

.....

.....

.....

.....

.....

.....

10.4 SIMPLE LINEAR REGRESSION

When we identify the fact that the correlation exists between two variables, we shall develop an estimating equation, known as regression equation or estimating line, i.e., a methodological formula, which helps us to estimate or predict the unknown value of one variable from known value of another variable. In the words of Ya-Lun-Chou, "regression analysis attempts to establish the nature of the relationship between variables, that is, to study the functional relationship between the variables and thereby provide a mechanism for prediction, or forecasting." For example, if we confirmed that advertisment expenditure (independent variable), and sales (dependent variable) are correlated, we can predict the required amount of advertising expenses for a given amount of sales or vice-versa. Thus, the statistical method which is used for prediction is called regression analysis. And, when the relationship between the variables is linear, the technique is called **simple linear regression**.

Hence, the technique of regression goes one step further from correlation and is about relationships that have been true in the past as a guide to what may happen in the future. To do this, we need the regression equation and the correlation coefficient. The latter is used to determine that the variables are really moving together.

The objective of simple linear regression is to represent the relationship between two variables with a model of the form shown below:

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

wherein

Y_i = value of the dependent variable,

β_0 = Y-intercept,

β_1 = slope of the regression line,

X_i = value of the independent variable,

e_i = error term (i.e., the difference between the actual Y value and the value of Y predicted by the model.

10.5 ESTIMATING THE LINEAR REGRESSION

If we consider the two variables (X variable and Y variable), we shall have two regression lines. They are:

i) Regression of Y on X

ii) Regression of X on Y.

The first regression line (Y on X) estimates value of Y for given value of X. The second regression line (X on Y) estimates the value of X for given value of Y. These two regression lines will coincide, if correlation between the variable is either perfect positive or perfect negative.

When we draw the regression lines with the help of a scatter diagram as shown earlier in Fig. 10.1, we may get an infinite number of possible regression lines for a set of data points. We must, therefore, establish a criterion for selecting the best line. The criterion used is the **Least Squares Method**. According to the least squares criterion, the best regression line is the one that minimizes the sum of squared vertical distances between the observed (X, Y) points and the regression line, i.e., $\sum(Y - \hat{Y})^2$ is the least value and the sum of the positive and negative deviations is zero, i.e., $\sum(Y - \hat{Y}) = 0$. It is important to note that the distance between (X, Y) points and the regression line is called the 'error'.

Regression Equations

As we discussed above, there are two regression equations, also called estimating equations, for the two regression lines (Y on X, and X on Y). These equations are, algebraic expressions of the regression lines, expressed as follows:

Regression Equation of Y on X

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

where, $\hat{Y}$ is the computed values of Y (dependent variable) from the relationship for a given X, 'a' and 'b' are constants (fixed values), 'a' determines the level of the fitted line at Y-axis (Y-intercept), 'b' determines the slope of the regression line, X represents a given value of independent variable.

The alternative simplified expression for the above equation is:

$$\hat{Y} - \overline{Y} = byx\ (X - \overline{X})$$

$$byx = r\frac{\sigma_y}{\sigma_x} = \frac{(\Sigma XY) - \frac{(\Sigma X)(\Sigma Y)}{N}}{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}$$

Regression equation of X on Y

$$\hat{X} = a + by$$

Alternative simplified expression is:

$$\hat{X} - \overline{X} = bxy\ (Y - \overline{Y})$$

$$bxy = r\frac{\sigma x}{\sigma y} = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}}$$

It is worthwhile to note that the estimated simple regression line always passes through $\overline{X}$ and $\overline{Y}$ (which is shown in Figure 10.3). The following illustration shows how the estimated regression equations are obtained, and hence how they are used to estimate the value of Y for given X value.

Illustration 6

From the following 12 months sample data of a company, estimate the regression lines and also estimate the value of sales when the company decided to spend Rs. 2,50,000 on advertising during the next quarter.

(Rs. in lakh)

Advertisement Expenditure:	0.8	1.0	1.6	2.0	2.2	2.6	3.0	3.0	4.0	4.0	4.0	4.6
Sales:	22	28	22	26	34	18	30	38	30	40	50	46

Solution:

Table 10.4: Calculations for Least Square Estimates of a Company.

(Rs. in lakh)

Advertising (X)	(Y)	Sales X^2	Y^2	XY
0.8	22	0.64	484	17.6
1.0	28	1.00	784	28.0
1.6	22	2.56	484	35.2
2.0	26	4.00	676	52.0
2.2	34	4.84	1156	74.8
2.6	18	6.76	324	46.8
3.0	30	9.00	900	90.0
3.0	38	9.00	1,444	114.0
4.0	30	16.00	900	120.0
4.0	40	16.00	1600	160.0
4.0	50	16.00	2,500	200.0
4.6	46	21.16	2,116	211.6
$\Sigma X=32.8$	$\Sigma Y=384$	$\Sigma X^2=106.96$	$\Sigma Y_2=13368$	$\Sigma XY=1,150.0$

Now we establish the best regression line (estimated by the least square method).

We know the regression equation of Y on X is:

$$\hat{Y} - \overline{Y} = byx\,(X - \overline{X})$$

$$\overline{Y} = \frac{384}{12} = 32; \quad \overline{X} = \frac{32.8}{12} = 2.733 \qquad \hat{Y} - \overline{Y} = byx\,(X - \overline{X})$$

$$byx = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}$$

$$= \frac{1{,}150 - \frac{(32.8)(384)}{12}}{106.96 - \frac{(32.8)^2}{12}} = 5.801$$

$$\hat{Y} - 32 = 5.801\,(X - 2.733)$$

$$\hat{Y} = 5.801X - 15.854 + 32 = 5.801X + 16.146$$

or $\hat{Y} = 16.146 + 5.801X$

which is shown in Figure 10.3. Note that, as said earlier, this line passes through $\overline{X}$ (2.733) and $\overline{Y}$ (32).

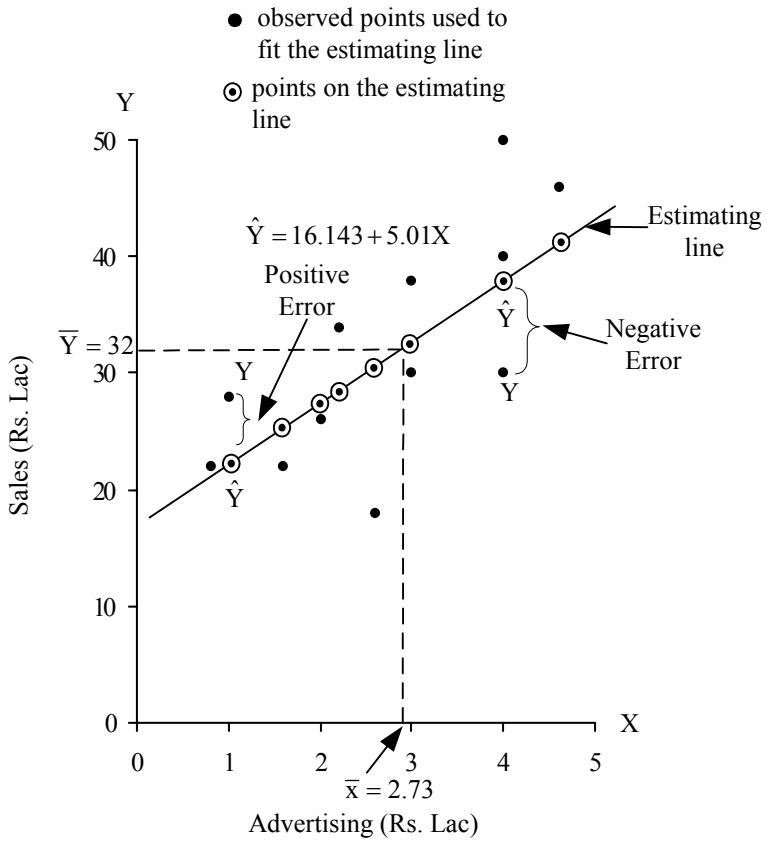


Figure 10.3: Least Squares Regression Line of a Company's Advertising Expenditure and Sales.

It is worthwhile to note that the relationship displayed by the scatter diagram may not be the same if the estimating equation is extended beyond the data points (values) considered in computing the regression equation.

Using Regression for Prediction

Regression, a statistical technique, is used for predictive purposes in applications ranging from predicting demand sales to predicting production and output levels. In the above illustration 6, we obtained the regression model of the company for predicting sales which is:

$$\hat{Y} = 16.146 + 5.801X$$

wherein $\hat{Y}$ = estimated sales for given value of X, and

X = level of advertising expenditure.

To find $\hat{Y}$, the estimate of expected sales, we substitute the specified

advertising level into the regression model. For example, if we know that the company's marketing department has decided to spend Rs. 2,50,000/- (X = 2.5) on advertisement during the next quarter, the most likely estimate of sales ($\hat{Y}$) is :

$$\hat{Y} = 16.1436 + 5.801(2.5) = 30.6455$$

= Rs. 30,64,850

Thus, an advertising expenditure of Rs. 2.5 lakh is estimated to generate sales for the company to the tune of Rs. 30,64,850.

Similarly, we can also establish the best regression line of X on Y as follows:

Regression Equation of X on Y

$$\hat{X} - \overline{X} = bxy\,(Y - \overline{Y})$$

$$bxy = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}} = \frac{1{,}150 - \frac{(32.8)(384)}{12}}{13368 - \frac{(384)^2}{12}} = 0.093$$

$$\hat{X} - 2.733 = 0.093\,(Y - 32)$$

$$\hat{X} - 2.733 = 0.093\,Y - 2.976$$

$$\hat{X} = 2.733 - 2976 + 0.093Y$$

$$\hat{X} = -0.243 + 0.093Y$$

The following points about the regression should be noted:

1) The geometric mean of the two regression coefficients (byx and bxy) gives coefficient of correlation.

That is, $r = \pm\sqrt{(bxy)\ (byx)}$

Consider the values of regression coefficients from the previous illustration to know the degree of correlation between advertising expenditure and sales.

$$r = \pm\sqrt{0.093 \times 5.801} = 0.734$$

2) Both the regression coefficients will always have the same sign (+ or –).

3) Coefficient of correlation will have the same sign as that of regression coefficients. If both are positive, then r is positive. In case both are negative, r is also negative. For example, bxy = –1.3 and byx = –0.65, then r is:

$\pm\sqrt{-1.3x - 0.65} = -0.919$ but not +0.919.

4) Regression coefficients are independent of change of origin, but not of scale.

10.5.1 Standard Error of Estimate

Once the line of best fit is drawn, the next process in the study of regression analysis is how to measure the reliability of the estimated regression equation. Statisticians have developed a technique to measure the reliability of the estimated regression equation called "Standard Error of Estimate (S_e)." This S_e is similar to the standard deviation which we discussed in Unit-9 of this course. We will recall that the standard deviation is used to measure the variability of a distribution about its mean. Similarly, **the standard error of estimate measures the variability, or spread, of the observed values around the regression line.** We would say that both are measures of variability. The larger the value of S_e, the greater the spread of data points around the regression line. If S_e is zero, then all data points would lie exactly on the regression line. In that case the estimated equation is said to be a perfect estimator. The formula to measure S_e is expressed as:

$$S_e = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n}}$$

where, S_e is standard error of estimate, Y is values of the dependent variable, $\hat{Y}$ is estimated values from the estimating equation that corresponds to each Y value, and n is the number of observations (sample size).

Let us take up an illustration to calculate S_e in a given situation.

Illustration 7

Consider the following data relating to the relationships between expenditure on research and development, and annual profits of a firm during 1998–2004.

Years:	1998	1999	2000	2001	2002	2003	2004
R&D (Rs. lakh):	2.5	3.0	4.2	3.0	5.0	7.8	6.5
Profit (Rs. lakh):	23	26	32	30	38	46	44

The estimated regression equation in this situation is found to be $\hat{Y} = 14.44 + 4.31x$. Calculate the standard error of estimate.

Note: Before proceeding to compute S_e you may calculate the regression equation of Y on X on your own to ensure whether the given equation for the above data is correct or not.

Solution: To calculate S_e for this problem, we must first obtain the value of $\sum(Y - \hat{Y})^2$. We have done this in Table 10.5.

Table 10.5: Calculation of $\Sigma (Y- \hat{Y})^2$

(Rs. in lakh)

Years	Expenditure on R&D X	Profit Y	$\hat{y}$ Estimating values (14.44 + 4.31X)	Individual error $(y - \hat{y})$	$(y - \hat{y})^2$
1998	2.5	23	14.44 + 4.31(2.5) = 25.21	–2.21	4.88
1999	3.0	26	14.44 + 4.31(3) = 27.37	–1.37	1.88
2000	4.2	32	14.44 + 4.31(4.2) = 32.54	–0.54	0.29
2001	3.0	30	14.44 + 4.31(3) = 27.37	2.63	6.92
2002	5.0	38	14.44 + 4.31(5) = 35.99	2.01	4.04
2003	7.8	46	14.44 + 4.31(7.8) = 48.06	–2.06	4.24
2004	6.5	44	14.44 + 4.31(6.5) = 42.46	1.54	2.37

$$\Sigma (Y - \hat{Y})^2 = 24.62$$

We can, now, find the standard error of estimate as follows.

$$S_e = \sqrt{\frac{\Sigma (Y - \hat{Y})^2}{n}}$$

$$\sqrt{\frac{24.62}{7}} = 1.875$$

Standard error of estimate of annual profit is Rs. 1.875 lakh.

We also notice, as discussed in Section 10.5, that $\Sigma (Y - \hat{Y}) = 0$. This is one way to verify the accuracy of the regression line fitted by the least square method.

10.5.2 Coefficient of Determination

Coefficient of determination (R^2) measures the percentage of variation in the dependent variable which is explained by the independent variable. R^2 can be any value between 0 and 1. It is used by many decision-makers to indicate how well the estimated regression line fits the given (X, Y) data points. If R^2 is closer to 1, the better the fit which in turn implies greater explanatory power of the estimated regression equation and, therefore, better prediction of the dependent variable. On the other hand, if R^2 is closer to 0 (zero), it indicates a very weak linear relationship. Thus prediction should not be based on such weak estimated regression. R^2 is given by:

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} \quad \text{or,} \quad 1 - \frac{\Sigma (Y - \hat{Y})^2}{\Sigma (Y - \bar{Y})^2}$$

$$\Sigma(Y-\overline{Y})^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n}$$

Note: When we employ simple regression, there is an alternative way of computing R^2, as shown in the equation below:

$$R^2 = r^2$$

where, R^2 is the coefficient of determination and r is the simple coefficient of correlation.

Refer to Illustration 6, where we have computed 'r' with the help of regression coefficients (bxy and byx), as an example for R^2

$$r = 0.734$$

$$R^2 = r^2 = 0.734^2 = 0.5388$$

This means that 53.88 per cent of the variation in the sales (Y) can be explained by the level of advertising expenditure (X) for the company.

Self Assessment Exercise C

You are given the following data relating to age of Autos and their maintenance costs. Obtain the two regression equations by the method of least squares and estimate the likely maintenance cost when the age of Auto is 5 years and also compute the standard error of estimate.

Age of Autos (yrs.):	2	4	6	8
Maintenance costs (Rs.00):	10	20	25	30

.....

.....

.....

.....

.....

.....

10.6 DIFFRENCE BETWEEN CORRELATION AND REGRESSION

After having an understanding about the concept and application of simple correlation and simple regression, we can draw the difference between them. They are:

1) Correlation coefficient 'r' between two variables (X and Y) is a measure of the direction and degree of the linear relationship between them, which is mutual. It is symmetric (i.e., $r_{xy} = r_{yx}$) and it is inconsiderable which, of X and Y, is dependent variable and which is independent variable. Whereas regression analysis aims at establishing the functional relationship between the two variables under study, and then using this relationship to predict the value of the dependent variable for any given value of the independent variable. It also

reflects upon the nature of the variables (i.e., which is the dependent variable and which is independent variable). Regression coefficients, therefore, are not symmetric in X and Y (i.e., byx ≠ bxy).

2) Correlation need not imply cause and effect relationship between the variables under study. But regression analysis clearly indicates the cause and effect relationship between the variables. The variable corresponding to cause is taken as independent variable and the variable corresponding to effect is taken as dependent variable.

3) Correlation coefficient 'r' is a relative measure of the linear relationship between X and Y variables and is independent of the units of measurement. It is a number lying between ±1. Whereas the regression coefficient byx (or bxy) is an absolute measure representing the change in the value of the variable Y (or X) for a unit change in the value of the variable X (or Y). Once the functional form of the regression curve is known, by susbstituting the value of the dependent variable we can obtain the value of the independent variable which will be in the unit of measurement of the variable.

4) There may be spurious (non-sense) correlation between two variables which is due to pure chance and has no practical relevance. For example, the correlation between the size of shoe and the income of a group of individuals. There is no such thing as spurious regression.

5) Correlation analysis is confined only to study of linear relationship between the variables and, therefore, has limited applications. Whereas regression analysis has much wider applications as it studies linear as well as non-linear relationships between the variables.

10.7 LET US SUM UP

In this unit, fundamental concepts and techniques of correlation (or association) and simple linear regression have been discussed. Scatter diagrams, which exhibit some typical patterns indicating different kinds of relationships have been illustrated. A scatter plot of the variables may suggest that the two variables are related but the value of the Karl Pearson's correlation coefficient (r) quantifies the degree of this association. The closer the correlation coefficient is to ±1.0, the stronger the linear relationship between the two variables. Test for significance of the correlation coefficient has been described. Spearman's rank correlation for data with ranks is outlined.

Once it is identified that correlation exists between the variables, an estimating equation known as regression equation could be developed by the least squares method for prediction. It also explained a statistical test called Standard Error of Estimate, to measure the accuracy of the estimated regression equation. Finally, the conceptual differences between correlation and regression have been highlighted. The techniques of correlation and regression analysis are widely used in business decision making and data analysis.

10.8 KEY WORDS

Coefficient of Determination: The square of the correlation coefficient. A measure that defines the proportion of variation in the dependent variable explained by the independent variable in the regression model.

Correlation Coefficient: A quantitative measure of the linear relationship between two variables.

Linear Relationship: The relationship between two variables described by a straight line.

Least Squares Criterion: The criterion for determining a regression line that minimizes the sum of squared errors.

Simple Regression Analysis: A regression model that uses one independent variable to explain the variation in the dependent variable.

Spurious Correlation: Correlation between two variables that have no known cause and effect connection.

Standard Error of Estimate: A measure of the dispersion of the actual Y values around the estimated regression line.

10.9 ANSWERS TO SELF ASSESSMENT EXERCISES

B) 1. $r_K = -0.071$

$t = -0.1743$

For 6 degrees of freedom, the critical value for t at a 5% level of significance is = 2.4469

2. $R = -0.185$

$t = -1.149$

table value of t for d.f at a 5% level of significance is 2.4469.

C) Y on X : $\hat{Y} = 5 + 3.25x$

X on Y : $\hat{X} = -3 + 0.297y$

10.10 TERMINAL QUESTIONS/EXERCISES

1) What do you understand by the term Correlation? Distinguish between different kinds of correlation with the help of scatter diagrams.

2) Explain the difference between Karl Pearson's correlation coefficient and Spearman's rank correlation coefficient. Under what situations, is the latter preferred to the former?

3) What do you mean by Spurious Correlation?

4) What do you understand by the term regression? Explain its significance in decision-making.

5) Distinguish between correlation and regression.

6) A personal manager of a firm is interested in studying as to how the number of worker absent on a given day is related to the average temperature on that day. A random sample of 12 days was used for the study. The data is given below:

No. of workers absent:	6	4	8	9	3	8	5	2	4	10	7	6
Average temperature (^oC):	12	30	15	18	40	30	45	35	23	15	25	35

a) State the independent variable and dependent variable.

b) Draw a scatter diagram.

c) What type of relationship appears between the variables?

d) What is the logical explanation for the observed relationship?

7) The following table gives the demand and price for a commodity for 6 days.

Price (Rs.):	4	3	6	9	12	10
Demand (mds):	46	65	50	30	15	25

a) Obtain the value of correlation coefficient and test its significance at 5% level.

b) Develop the estimating regression equations.

c) Compute the standard error of estimate.

d) Predict Demand for price (Rs.) = 5, 8, and 11.

f) Compute coefficient of determination and give your comment on the distribution.

8) Two judges have ranked 10 students in order of their merit in a competition.

Students:	A	B	C	D	E	F	G	H	I	J
Rank by Ist judge:	5	2	4	1	8	9	7	6	3	10
Rank by IInd judge:	1	9	7	8	10	2	4	5	3	6

Find out whether the judges are in agreement with each other or not and apply the t-test for significance at 5% level.

9) A sales manager of a soft drink company is studying the effect of its latest advertising campaign. People chosen at random were called and asked how many bottles they had bought in the past week and how many advertisements of this product they had seen in the past week.

No. of ads (X):	4	0	2	7	3	4	2	6
Bottles purchased (Y):	6	5	4	16	10	9	6	14

a) Develop the estimating equation that best fits the data and test its accuracy.

b) Calculate correlation coefficient and coefficient of determination.

c) Predict Y value when X = 5.

Note: These questions/exercises will help you to understand the unit better. Try to write answers for them. But do not submit your answers to the university for assessment. These are for your practice only.

10.11 FURTHER READING

A number of good text books are available for the topics dealt with in this unit. The following books may be used for more indepth study.

Richard I. Levin and David S. Rubin, 1996, *Statistics for Management*. Prentice Hall of India Pvt. Ltd., New Delhi.

Peters, W.S. and G.W. Summers, 1968, *Statistical Analysis for Business Decisions*, Prentice Hall, Englewood-cliffs.

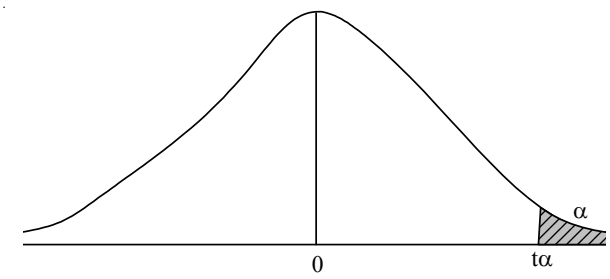
Hooda, R.P., 2000, *Statistics for Business and Economics*, MacMillan India Ltd., New Delhi.

Gupta, S.P. 1989, *Elementary Statistical Methods*, Sultan Chand & Sons : New Delhi.

Chandan, J.S. - *Statistics for Business and Economics*, Vikas Publishing House Pvt. Ltd., New Delhi.

APPENDIX : TABLE OF t-DISTRIBUTION AREA

The table gives points of t- distribution corresponding to degrees of freedom and the upper tail area (suitable for use n one tail test).



Values of $t_{a, m}$

m \ α	0.1	0.05	0.025	0.01	0.005
1	3.078	6.3138	12.706	31.821	63.657
2	1.886	2.9200	4.3027	6.965	9.9248
3	1.638	2.3534	3.1825	4.541	5.8409
4	1.533	2.1318	2.7764	3.747	4.6041
5	1.476	2.0150	2.5706	3.365	4.0321
6	1.440	1.9432	2.4469	3.143	3.7074
7	1.415	1.8946	2.3646	2.998	3.4995
8	1.397	1.8595	2.3060	2.896	3.3554
9	1.383	1.8331	2.2622	2.821	3.2498
10	1.372	1.8125	2.2281	2.764	3.1693
11	1.363	1.7959	2.2010	2.718	3.1058
12	1.356	1.7823	2.1788	2.681	3.0545
13	1.350	1.7709	2.1604	2.650	3.0123
14	1.345	1.7613	2.1448	2.624	2.9768
15	1.341	1.7530	2.1315	2.602	2.9467
16	1.337	1.7459	2.1199	2.583	2.9208
17	1.333	1.7396	2.1098	2.567	2.8982
18	1.330	1.7341	2.1009	2.552	2.8784
19	1.328	1.7291	2.0930	2.539	2.8609
20	1.325	1.7247	2.0860	2.528	2.8453
21	1.323	1.7207	2.0796	2.518	2.8314
22	1.321	1.7171	2.0739	2.508	2.8188
23	1.319	1.7139	2.0687	2.500	2.8073
24	1.318	1.7109	2.0639	2.492	2.7969
25	1.316	1.7081	2.0595	2.485	2.7874

(Contd…)

m \ α	0.10	0.05	0.025	0.01	0.005
26	1.315	1.7056	2.0555	2.479	2.7787
27	1.314	1.7033	2.0518	2.473	2.7707
28	1.313	1.7011	2.0484	2.467	2.7633
29	1.311	1.6991	2.0452	2.462	2.7564
30	1.310	1.6973	2.0423	2.457	2.7500
35	1.3062	1.6896	2.0301	2.438	2.7239
40	1.3031	1.6839	2.0211	2.423	2.7045
45	1.3007	1.6794	2.0141	2.412	2.6896
50	1.2987	1.6759	2.0086	2.403	2.6778
60	1.2959	1.6707	2.0003	2.390	2.6603
70	1.2938	1.6669	1.994	2.381	2.6480
80	1.2922	1.6641	1.9945	2.374	2.6388
90	1.2910	1.6620	1.9901	2.364	2.6316
100	1.2901	1.6602	1.9867	2.364	2.6260
120	1.2887	1.6577	1.9840	2.358	2.6175
140	1.2876	1.6658	1.9799	2.353	2.6114
160	1.2869	1.6545	1.9771	2.350	2.6070
180	1.2863	1.6534	1.9749	2.347	2.6035
200	1.2858	1.6525	1.9733	2.345	2.6006
∞	1.282	1.645	1.96	2.326	2.576