UNIT 5 CORRELATION AND REGRESSION*

Structure

5.0 OBJECTIVES

After going through this unit you will be in a position to

- plot scatter diagram;
- compute correlation coefficient and state its properties;
- compute rank correlation;
- explain the concept of regression;
- explain the method of least squares;
- identify the limitations of linear regression;
- apply linear regression models to given data; and
- use the regression equation for prediction.

5.1 INTRODUCTION

The word 'bivariate' is used to describe situations in which two character are measured on each individual or item, the character being represented by two variables. For example, the measurement of height (X_i) and weight (Y_i) of students in a school. The subscript i in this case represents the student concerned.

* Prof. Kaustuva Barik, School of Social Sciences, Indira Gandhi National Open University.

Thus, for example, X_5, Y_5 represent the height and weight of the fifth student. Statistical data relating to simultaneous measurement of two variables are called bivariate data. The observation on each individual are paired, one for each variable (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n).

In statistical studies with several variables, there are generally two types of problems. In some problems it is of interest to study how the variables are interrelated; such problems are tackled by using correlation technique. For instance, an economist may be interested in studying the relationship between the stock prices of various companies; for this he may use correlation techniques. In other problems there is a variable y of basic interest and the problem is to find out what information the other variable provides on Y, such problems are tackled using regression techniques. For instance, an economist may be interested in studying what factors determine the pay of an employed person and in particular, he may be interested in exploring what role the factors such as education, experience, market demand, etc. play in determining the pay. In the above situation he may use regression techniques to set up a prediction formula for pay based on education, experience, etc.

5.2 SCATTER DIAGRAM

We first illustrate how the relationship between two variables is studied. A teacher is interested in studying the relationship between the performance in Statistics and Economics of a class of 20 students. For this he compiles the scores on these subjects of the students the last semester examination. Some data of this type are presented in Table 5.1.

Table 5.1: Scores of 20 Students in Statistics and Economics

Serial Number	Score in Statistics	Score in Economics	Serial Number	Score in Statistics	Score in Economics
1	82	64	11	76	58
2	70	40	12	76	66
3	34	35	13	92	72
4	80	48	14	72	46
5	66	54	15	64	44
6	84	56	16	86	76
7	74	62	17	84	52
8	84	66	18	60	40
9	60	58	19	82	60
10	86	82	20	90	60

A representation of data of this type on a graph is a useful device which will help us to understand the nature and form of the relationship between the two variables, whether there is a discernible relationship or not and if so whether it is linear of not. For this let us denote score in Economics by X and the score in Statistics by Y and plot the data of Table 5.1 on the x-y plane. It does not matter which is called X and which Y for this purpose. Such a plot is called *Scatter Plot* or *Scatter Diagram*. For data of Table 5.1 the scatter diagram is given in Fig. 5.1.

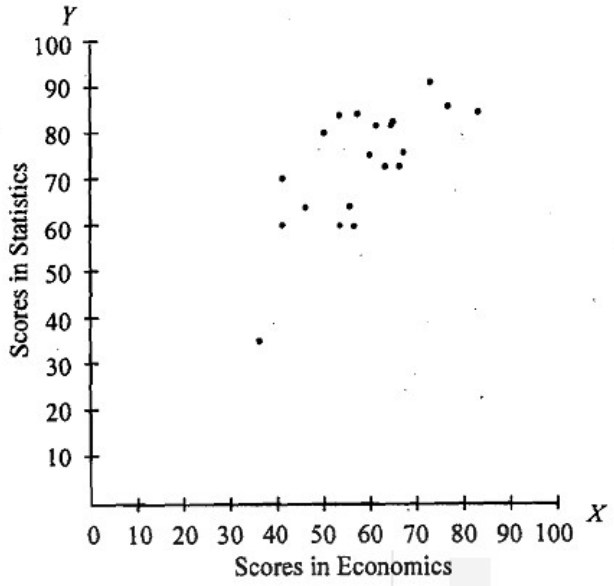

Fig. 5.1: Scatter Diagram of Scores in Statistics and Economics.

An inspection of Table 5.1 and Fig. 5.1 shows that there is a *positive relationship* between *x* and *y*. This means that larger values of *x* associated with larger values of *y* and smaller values of *y*. Further, the points seem to lie scattered around both sides of a straight line. Thus, it appears that a linear relationship exists between *x* and *y*. This relationship, however, in not *perfect* in the sense that there are deviations from such a relationship in the case of certain observations. It would indeed be useful to get a measure of the strength of this linear relationship.

5.3 COVARIANCE

In the case of a single variable we have learnt the concept of variance, which is defined as

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \ldots (5.1)$$

In the above we use a subscript *x* to specify that σ_2^2 represents the variance in *x*. In a similar manner we can represent σ_y^2 as the variance in *y* and σ_x and σ_y as the standard deviation in *x* and *y* respectively.

As you know, variance measures the dispersion from mean. In the case of bivariate data we have to reach a single figure which will present the deviation in both the variables from their respective means. For this purpose we use a concept termed covariance, which is defined as follows:

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \quad \ldots (5.2)$$

You may recall that standard deviation is always positive since it is defined as the positive square root of variance. In the case of covariance there are two terms $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$ which represent the deviations in *x* from $\bar{X}$ and *Y* from $\bar{Y}$.

Moreover, $(X_i - \overline{X})$ can be positive or negative depending on whether x_i is less than or greater than $\overline{X}$. Similarly $(Y_i - \overline{Y})$ can be positive or negative. It is not necessary that whenever $(X_i - \overline{X})$ is positive $(Y_i - \overline{Y})$ will also be positive. Therefore, the product $(X_i - \overline{X})$ $(Y_i - \overline{Y})$ can be either positive or negative. A positive value for $(X_i - \overline{X})$ $(Y_i - \overline{Y})$ implies the whenever $X_i > \overline{X}$, we have $Y_i > \overline{Y}$. Thus a higher value of x_i is associated with a relatively higher value in y_i. On the other hand, $(X_i - \overline{X})(Y_i - \overline{Y}) < 0$ implies that a lower value in X_i is associated with a relatively higher value in y_i. when we sum it over all the observations and ivied by the number of observations, we may obtain a negative or positive value. Therefore, covariance can assume both positive and negative values.

When covariance between x and y is negative $(\sigma_{xy} < 0)$ we can say that the relationship could be inverse. Similarly, $(\sigma_{xy} < 0)$ implies a positive relationship between x and y. A major limitation of covariance is that it is not independent of unit of measurement. It means that if we change the unit of measurement of the variables we will get a difference value for σ_{xy}.

The computation of σ_{xy} as given in (5.2) often involves large numbers. Therefore, it is derived further as

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n}(X_iY_i - \overline{X}Y_i - \overline{XY})$$

By further simplification we find that

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}X_iY_i - \frac{1}{n}\sum_{i=1}^{n}\overline{X}Y_i - \frac{1}{n}\sum_{i=1}^{n}X_i\overline{Y} + \frac{1}{n}\sum_{i=1}^{n}\overline{XY}$$

Since $\frac{1}{n}\sum_{i=1}^{n}\overline{X}Y_i = \frac{1}{n}X_i\overline{Y} = \overline{X}\overline{Y}$ we have

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}X_iY_i - \overline{X}\overline{Y} \qquad \ldots (5.3)$$

5.4 CORRELATION COEFFICIENT

The task before us is to measure the linear relationship between x and y. It is desirable to have this measure of strength of linear relationship independent of the scale chosen for measuring the variables. For instance, if we are measuring the relationship between height and weight, we should get the same measure whether height is measured in inches or centimetres and weight in pounds or kilograms. Similarly, if a variable is temperature, it should not matter whether it is recorded in Celsius or Fahrenheit.

This can be achieved by standardizing each variable, that is by considering $\frac{X-\overline{X}}{\sigma_x}$ and $\frac{Y-\overline{Y}}{\sigma_y}$ where $\overline{X}$ and $\overline{Y}$ are the means of X and Y respectively and σ_x and σ_y are standard deviations.

Let us denote these standardised variables by u and y respectively. Let us also use the notation (X_i, Y_i) to denote the score i[th] student in Economics and Statistics respectively, i ranging from 1 to n, the number of students, n being 20 in our example. Similarly, let (u_i, v_i) denote the standardised scores of i[th] student. Then recall the following formulae for mean and standard deviation:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n}\overline{X}_i; \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(\overline{X}_i - \overline{X})^2;$$

$$Y = \frac{1}{n}\sum_{i=1}^{n}X_i; \sigma_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

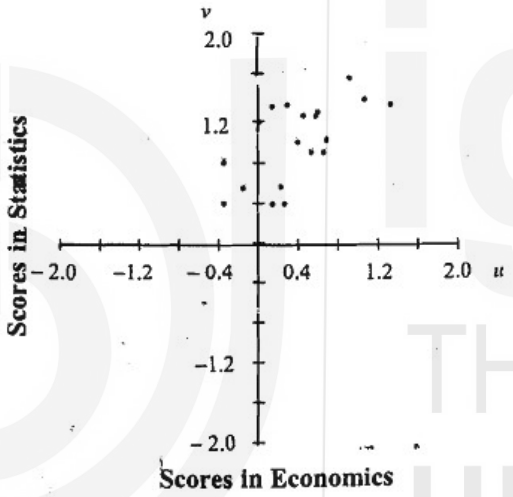

Fig. 5.2: Scatter Diagram of Standardised Scores in Statistics and Economics

Fig. 5.2 is the scatter diagram in terms of standardised variables u and v. Let us observe that in this example there is a positive association between the two scores. The larger one score is, the larger the other score also is; the smaller one score is the smaller the other score is, on the whole. In view of this, most of the points are either in the *first quadrant* or in the *third quadrant.* The first quadrant represents the cases where both scores are above their respective means and third quadrant represents the cases where both scores are below their respective means. There are only a very few points in second and fourth quadrants, which represent the cases where one score is above its mean and the other is below its mean. Thus the product of the u, v values is a suitable indicator of the strength of the relationship; this product is positive in the first and third quadrants and negative in the second and fourth. Thus the product of u, v averaged over all the points may be considered to be suitable measure of the strength of linear relationship between X and Y.

This measure is called the *correlation coefficient* between X and Y and is usually denoted by r_{xy} or simply by r, when it is clear what x and y in the context are. This is also called the *Pearson's Product-Moment Correlation Coefficient* to distinguish it from other types of correlation coefficients.

Thus the formula for r is

$$r = \frac{1}{n}\sum_{i=1}^{n} u_i v_i \qquad \ldots (5.4)$$

If we substitute the variables x and y in (5.4) above

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma_x}\right)\left(\frac{Y_i - \bar{Y}}{\sigma_y}\right) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sigma_x \sigma_y}$$

In the above expression, the term

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

is the *covariance* between x and y (σ_{xy}).

Thus, the formula for correlation coefficient is

$$r = \frac{\sigma_{xy}}{\sigma_x \times \sigma_y} \qquad \ldots (5.5)$$

Incorporating the formulae for $\bar{x}, \bar{y}, \sigma_x, \sigma_y$ it becomes

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \ldots (5.6)$$

Or, alternatively

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\left[\sqrt{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}\right]\left[n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right]} \qquad \ldots (5.7)$$

Let us go back to the data given in Table 5.1 and work out the value of r. You can use any of the formulae (5.4), (5.5) or (5.7) to get the value of r. Since all the formulae are derived from the same concept we obtain the same value for r whichever formulae we use. For the data set in Table 5.1 we have calculated it by using (5.4) and (5.7). We construct Table 5.2 for this purpose.

Table 5.2: Calculation of Correlation Coefficient

Observation No.	X	Y	X^2	Y^2	XY
1	82	64	6724	4096	5248
2	70	40	4900	1600	2800
3	34	35	1156	1225	1190
4	80	48	6400	2304	3840
5	66	54	4356	2916	3564
6	84	56	7056	3136	4704
7	74	62	5476	3844	4588
8	84	66	7056	4356	5544
9	60	52	3600	2704	3120
10	86	82	7396	6724	7052
11	76	58	5776	3364	4408
12	76	66	5776	4356	5016
13	92	72	8464	5184	6624
14	72	46	5184	2116	3312
15	64	44	4096	1936	2816
16	86	76	7396	5776	6536
17	84	52	7056	2704	4368
18	60	40	3600	1600	2400
19	82	60	6724	3600	4920
20	90	60	8100	3600	5400
Total	1502	1133	116292	67141	87450

From Table 5.2 we note that

$$\sum_{i=1}^{20} X_i = 150; \overline{X} = 75.1;$$

$$\sum_{i=1}^{20} Y_i = 1133; \overline{Y} = 56.65;$$

$$\sum_{i=1}^{20} X_i^2 = 116292; \sigma_x^2 = \frac{1}{20}\left[116292 - \frac{1502^2}{20}\right] = 174.59; \sigma_x = 13.21;$$

$$\sum_{i=1}^{20} Y_i^2 = 67141; \sigma_x^2 = \frac{1}{20}\left[67141 - \frac{1133^2}{20}\right] = 147.83; \sigma_y = 12.16;$$

$$\sum X_i Y_i = 87450; \sigma_{xy} = \frac{1}{20}\left[87450 - \frac{1502 \times 1133}{20}\right] = 118.09$$

Thus, using formula given at (5.4), we have

$$r = \frac{118.09}{13.21 \times 12.16} = 0.735$$

Now let us use the formula 5.7. We have

$$r = \frac{20 \times 87450}{\sqrt{(20 \times 116292 - 1502^2)(20 \times 67141 - 1133^2)}} = 0.735$$

Thus we see that both the formulae provide the same value of the correlation coefficient *r.* You can check yourself that the same value of *r* is obtained by using the formula (5.5). For this purpose you will need values on

$\sum(X_i - \bar{X})^2, \sum(Y_i - \bar{Y})^2$ and $\sum(X_i - \bar{X})(Y_i - \bar{Y})$

Hence you can have five columns on

$(X_i - \bar{X}), (Y_i - \bar{Y}), (X_i - \bar{X})^2, (Y_i - \bar{Y})^2$ and $(X_i - \bar{X})(Y_i - \bar{Y})$ in a table and find the totals.

5.5 INTERPRETATION OF CORRELATION COEFFICIENT

It is a mathematical fact that the value of *r* as defined above lies between –1and + 1. The extreme values of –1and + 1are obtained only in situations where there is a *perfect linear relationship* between *X* and *Y.* The –1is obtained when this relationship is perfectly negative (i.e., inverse) and +1 when this is perfect positive (i.e., direct). The value of 0 is obtained when there is no linear relationship between *x* and *y*.

We can make some guess work about the sign and degree of the correlation coefficient from the scatter diagram. Fig. 5.3 gives example of scatter diagrams for various values of *r.* Fig. 5.3 (a) is a scatter diagram for the case $r = 0$; here there is no *linear* relationship between *x* and *y.* Fig. 5.3(b) is also and example of scatter diagram for the case $r = 0$; here there is discernible relationship between *X* and *Y* but it is not of the linear type. Here, initially, *Y* increases with *X* but later *Y* decreases as *X* increases resulting in a definitive quadratic relationship. But the correlation coefficient in the case is zero. Thus the correlation coefficient is only a measure of linear relationship. This sort of scatter diagram is obtained, if we plot, for instance, body weight (*Y)* of individuals against there are (*X*). Fig. 5.3.(c) is an example of a scatter diagram where there is a perfect positive linear relationship between *X* and *Y*. We get this sort of scatter diagram if we plot, for instance, height of individuals in inches (*X*) against their heights in centimetres (*Y*); in that case $Y = 2.54X$, which is a deterministic and perfect linear relationship. Figures 5.3(d) to 5.3(k) are scatter diagrams for other values of *r*. Form these scatter diagrams we get an idea of the nature of relationship and associated values of *r*.

From these is would seem that a value of 0.81 indicates a fair degree of linear relationship between scores in Statistics and Economics of these candidates. Such a quantification of relationship or association between variables is helpful for natural and social scientists to understand the phenomena they are investigating and explore these phenomena further. In an example of this sort, an educational psychologist may compute correlation coefficients between scores in various subjects and by further statistical analysis of the correlation coefficients and using psychological techniques may be able to form a theory as to what mental and other faculties are involved in making students good in various disciplines.

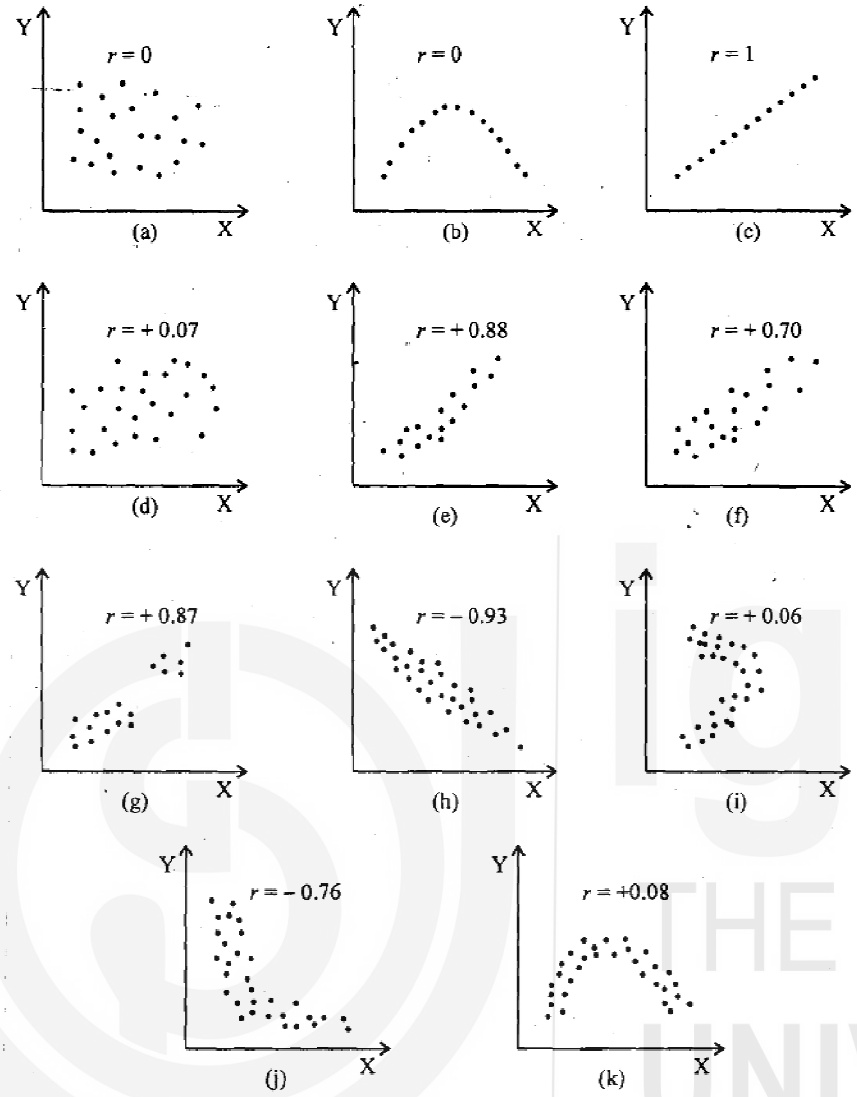

Fig. 5.3: Scatter Plots for Various Values of Correlation Coefficient

You should remember that

- Correlation coefficient shows the *linear relationship* between X and Y. Thus, even if there is a strong non-linear relationship between X and Y, correlation coefficient may be low.
- Correlation coefficient is independent of scale and origin. If we subtract some constant from one (or both) of the variables, correlation coefficient will remain unchanged from one (or both) of the variables by some constant, correlation coefficient will not change.
- Correlation coefficient varies between –1 and +1. It means that r cannot be smaller than –1 and cannot be greater than +1.

The existence of a linear relationship between two variables is not to be interpreted to mean a cause-effect relationship between the two.

For instance, if you work out the correlation between family expenditure on petrol and chocolates, you may find it to be fairly high indicating a fair degree of linear relationship. However, both of these are luxury items and richer families can afford them while poorer ones cannot. Thus the high correlation here is caused by the high correlation of each of the variables with family income. To consider another example, suppose for each of the last twenty years, you work out the average height of an Indian and the average time per week an Indian watches television; you are likely to find a positive correlation. This does not, however, imply that watching television increases one's height or that taller people tend to watch television longer. Both these variables have an increasing trend over time and this is reflected in the high correlation. This kind or correlation between two variables is caused by the effect of a third variable on each of them rather than a direct linear cause-effect of a third variable on each of them rather than a direct linear cause-effect relationship between them is called *spurious correlation.*

Another aspect of the computation of correlation coefficient that we should be aware of is that the correlation coefficient like any other quantity computed from sample, varies from sample to sample and these sample fluctuations should be taken into account in making use of the computed coefficient. We do not discuss these techniques here.

Whether the presence of a linear relationship between two variables and hence a high correlation between them is genuine or spurious, such a situation is helpful to *predict* one variable from the other.

Check Your Progress 1

1) Calculate r from the following given results:

   $n = 10; \sum X = 125; \sum X^2 = 1585; \sum Y = 80; \sum Y^2 = 650; \sum XY = 1007.$

   ……..

   ……..

   ……..

   ……..

   ……..

   ……..

2) Calculate the coefficient of correlation for the ages of husband and wife:

Age of husband	:	23	27	28	29	30	31	33	35	36	39
Age of wife	:	18	22	23	24	25	26	28	29	30	32

   ……..

   ……..

   ……..

   ……..

   ……..

   ……..

3) Specimens of similarly treated alloy steel containing various percentages of nickel are tested for toughness with the following results:

Toughness (arbitrary units)	47	50	52	52	54	56	58	59	60	60	62	64	65	66
Percentage of Nickel	2.7	2.7	2.8	2.8	2.9	3.2	3.2	3.3	3.4	3.5	3.6	3.7	3.7	3.8

Find the correlation coefficient between toughness and nickel content and comment on the result.

……..

……..

……..

……..

……..

……..

4) Determine the correlation coefficient between x and y.

x	:	5	7	9	11	13	15
y	:	1.7	2.4	2.8	3.4	3.7	4.4

……..

……..

……..

……..

……..

……..

5) The following table gives the saving bank deposits in billions of dollars and strikes and lock-outs, in thousands, over a number of years. Compute the correlation coefficient and comment on the result.

Saving deposits	:	5.1	5.4	5.5	5.9	6.4	6.0	7.2
Strikes and lock-outs	:	3.8	4.4	3.3	3.6	3.3	2.3	1.0

……..

……..

……..

……..

……..

……..

……..

5.6 RANK CORRELATION COEFFICIENT

The Pearson's product moment correlation coefficient (or simply, the correlation coefficient) described above is suitable if both the variables involved are measureable (numerical) and the relationship between the variables is linear. However, there are situations where variables are not numerical but various items can be ranked according to the characteristics (i.e., ordinal). Sometimes even when the original variables are measurable, they are converted into ranks and a measure of association is computed. Consider for instance the situation when two examiners are asked to judge ten candidates on the basis of an oral examination. In this case, it may be difficult to assign scores to candidates, but the examiners find it reasonably easy to rank the candidates in order of merit. Before using the resulted it may be advisable to find out if rankings are in reasonable concordance. For this, a measure of association between the ranks assigned by the two examiners may by computed. The Karl Pearson's correlation coefficient is not suitable in this situation. One may use the following called *Spearman's Rank Correlation Coefficient* for this purpose.

Table 5.3: Ranks of 10 Candidates by two Examiners

S. No	Rank given by		Difference	
	Examiner I	*Examiner II*	D_i	D_i^2
1	6.0	6.5	–0.5	0.25
2	2.0	3.0	–1.0	1.00
3	8.5	6.5	2.0	4.00
4	1.0	1.0	0.0	0.00
5	10.0	2.0	8.0	64.00
6	3.0	4.0	–1.0	1.00
7	8.5	9.5	–1.0	1.00
8	4.0	5.0	–1.0	1.00
9	5.0	8.0	–3.0	9.00
10	7.0	9.5	–2.5	6.25
		$\sum D_i = 0$	$\sum D_1^2 = 87.50$	

Let us consider the data of Table 5.3. Here there are some ties; the tied cases are given the same rank in such a way their total is the same as when there is no tie. For example, when there are two cases with rank 6, each is given a rank of 6.5 and there is no case with rank either 6 or 7. Similarly, if there are three cases with rank 5, then each is given a rank of 6 and there is no case with rank 5 or 7. Spearman's rank correlation coefficient, called Spearman's Rho, denoted by ρ, is based on the difference D_i (i for i^{th} observation) between the two rankings. If the two rankings completely coincide, then D_i is zero for every case. The larger the value of D_i, the greater is the difference between the two rankings and smaller is the association. Thus, the association can be measured by considering the magnitudes of D_i. Since the sum of D_i is always zero, to find a single index on the basis of D_i values, we should remove the sign of D_i and consider only the magnitude. In Spearman's ρ, this is done by taking D_i^2.

However, the largeness or smallness of $\sum_{i=1}^{n} D_i^2$, where n is the number of cases, will depend on n. thus, in order to be able to interpret this value, we could create a ratio by dividing this sum by the largest possible value, which depends only on n, which is $\frac{n(n^2-1)}{6}$. However, $\frac{6\times\sum_{i=1}^{n} D_i^2}{n(n^2-1)}$ is zero for perfect association and 2 for lack of association, i.e., perfect negative association, while we would like it to be other way around. So we subtract this ratio from 1. Thus

$$\rho = 1 - \frac{6\times\sum_{i=1}^{n} D_i^2}{n(n^2-1)} \qquad \ldots (5.8)$$

is defined as Spearman's rank correlation.

Let us calculated the value of ρ from the data given in Table 5.3.

$$\rho = -\frac{6\times 87.5}{10(10^2-1)} = 1 - \frac{525}{990} = 1 - 0.53 = 0.47.$$

Like Karl Pearson's coefficient of correlation the Spearman's rank correlation has a value +1 for perfect matching of ranks, –1for perfect mismatching of ranks and 0 for the lack of relation between the ranks.

There are other measures of association suitable for use when the variables are of nominal, ordinal and other types. We do not discuss them here.

Check Your Progress 2

1) In a contest, two judges ranked eight candidates A, B, C, D, E, F, G and H in order of their preference, as shown in the following table. Find the rank correlation coefficient.

	A	B	C	D	E	F	G	H
First Judge	5	2	8	1	4	6	3	7
Second Judge	4	5	7	3	2	8	1	6

…….

…….

……..…

……..…

……..…

……..…

……..

2) Compute the correlation coefficient of the following ranks of a group of students in two examinations. What conclusion do you draw from the results?

Roll Nos.	1	2	3	4	5	6	7	8	9	10
Rank in B. Com. Exam.	1	5	8	6	7	4	2	3	9	10
Rank in M. Com Exam.	2	1	5	7	6	3	4	8	10	9

…….....

…….....…….

…….....…..…

…….....…..…

…….....…..…

…….....…..…

3) Ten competitors in a musical contest were ranked by 3 judges A, B and C in the following order:

Ranks by A :	1	6	5	10	3	2	4	9	7	8
Ranks by B :	3	5	8	4	7	10	2	1	6	9
Ranks by C :	6	4	9	8	1	2	3	10	5	7

Using the rank correlation method, discuss which pair of judges has the nearest approach to common liking in music.

…….....…….....

…….....……..

…….....…..…

…….....…..…

…….....…..…

4) Ten students obtained the following marks in Mathematics and Statistics. Calculate the rank correlation coefficient.

Student (Roll No.)	1	2	3	4	5	6	7	8	9	10
Marks in Mathematics	78	36	98	25	75	82	90	62	65	39
Marks in Statistics	84	51	91	60	68	62	86	58	53	47

…….....

…….....….

…….....…

…….....…

…….....…

5.7 THE CONCEPT OF REGRESSION

In the previous section we noted that correlation coefficient does not reflect cause and effect relationship between two variables. Thus we cannot predict the value of one variable for a given value of the other variable. This limitation is removed by regression analysis. In regression analysis, the relationship between variables are expressed in the form of a mathematical equation. It is assumed that one variable is the cause and the other is the effect. You should remember that regression is a statistical tool which helps understand the relationship between variables and predicts the unknown values of the dependent variable from known values of the independent variable.

In regression analysis we have two types of variables: i) dependent (or explained) variable, and ii) independent (or explanatory) variable. As the name (explained and explanatory) suggests the dependent variable is explained by the independent variable.

In the simplest case of regression analysis there is one dependent variable and one independent variable. Let us assume that consumption expenditure of a household is related to the household income. For example, it can be postulated that as household income increases, expenditure also increases. Here consumption expenditure is the dependent variable and household income is the independent variable.

Usually we denote the dependent variable as Y and the independent variable as X. Suppose we took up a household survey and collected *n* pairs of observations in X and Y. The next step is to find out the nature of relationship between X and Y.

The relationship between X and Y can take many forms. The general practice is to express the relationship in terms of some mathematical equation. The simplest of these equations is the linear equation. This means that the relationship between X and Y is in the form of a straight line and is termed linear regression. When the equation represents curves (not a straight line) the regression is called non-linear or curvilinear.

Now the question arises, 'How do we identify the equation form?' There is no hard and fast rule as such. The form of the equation depends upon the reasoning and assumptions made by us. However, we may plot the X and Y variables on a graph paper to prepare a scatter diagram. From the scatter diagram, the location of the points on the graph paper helps in identifying the type of equation to be fitted. If the points are more or less in a straight line, then linear equation is assumed. On the other hand, if the points are not in a straight line and are in the form of a curve, a suitable non-linear equation (which resembles the scatter) is assumed.

We have to take another decision, that is, the identification of dependent and independent variables. This again depends on the logic put forth and purpose of analysis: whether 'Y depends on X' or 'X depends on Y'. Thus there can be two regression equations from the same set of data. These are i) Y is assumed to be

dependent on X (this is termed 'Y on X' line), and ii) X is assumed to be dependent on Y (this is termed 'X on Y' line).

Regression analysis can be extended to cases where one dependent variable is explained by a number of independent variables. Such a case is termed multiple regression. In advanced regression models there can be a number of both dependent as well as independent variables.

You may by now be wondering why the term 'regression', which means 'reduce'. This name is associated with a phenomenon that was observed in a study on the relationship between the stature of father (x) and son (y). It was observed that the average stature of sons of the tallest fathers has a tendency to be less than the average stature of these fathers. On the other hand, the average stature of sons of the shortest fathers has a tendency to be more than the average stature of these fathers. This phenomenon was called *regression towards the mean*. Although this appeared somewhat strange at that time, it was found later that this is due to natural variation within subgroups of a group and the same phenomenon occurred in most problems and data sets. The explanation is that many tall men come from families with average stature due to vagaries of natural variation and they produce sons who are shorter than them on the whole. A similar phenomenon takes place at the lower end of the scale.

5.8 LINEAR RELATIONSHIP: TWO-VARIABLES CASE

The simplest relationship between X and Y could perhaps be a linear *deterministic* function given by

$$Y_i = a + bX_i \qquad \ldots(5.9)$$

In the above equation X is the independent variable or explanatory variable and Y is the dependent variable or explained variable. You may recall that the subscript i represents the observation number, i ranges from 1 to n. Thus Y_1 is the first observation of the dependent variable, X_5 is the fifth observation of the independent variable, and so on.

Equation (5.9) implies that Y is completely determined by X and the parameters a and b. Suppose we have parameter values $a = 3$ and $b = 0.75$, then our linear equation is $Y = 3 + 0.75\ X$. From this equation we can find out the value of Y for given values of X. For example, when $X = 8$, we find that $Y = 9$. Thus if we have different values of X then we obtain corresponding Y values on the basis of (5.9). Again, if X_i is the same for two observations, then the value of Y_i will also be identical for both the observations. A plot of Y on X will show no deviation from the straight line with intercept 'a' and slope 'b'.

If we look into the deterministic model given by (5.9) we find that it may not be appropriate for describing economic interrelationship between variables. For example, let Y = consumption and X = income of households. Suppose you record your income and consumption for successive months.

For the months when your income is the same, do your consumption remain the same? The point we are trying to make is that economic relationship involves certain randomness.

Therefore, we assume the relationship between Y and X to be *stochastic* and add one error term in (5.9). Thus our stochastic model is

$$Y_i = a + bX_i + e_i \qquad \ldots(5.10)$$

where e_i is the error term. In real life situations e_i represents randomness in human behaviour and excluded variables, if any, in the model. Remember that the right hand side of (5.10) has two parts, viz., i) deterministic part (that is, $a + bX_i$), and ii) stochastic or randomness part (that is, e_i). Equation (5.10) implies that even if X_i remains the same for two observations, Y_i need not be the same because of different e_i. Thus, if we plot (5.10) on a graph paper the observations will not remain on a straight line.

Example 5.1

The amount of rainfall and agricultural production for ten years are given in Table 5.4.

Table 5.4: Rainfall and Agricultural Production

Rainfall (in mm.)	Agricultural production (in tonne)
60	33
62	37
65	38
71	42
73	42
75	45
81	49
85	52
88	55
90	57

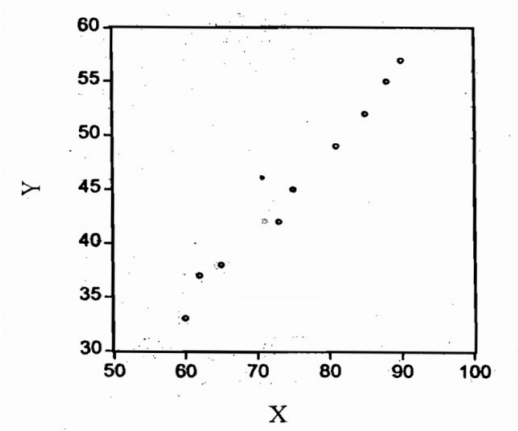

Fig. 5.4: Scatter Diagram

We plot the data on a graph paper. The scatter diagram looks something like Fig. 5.4. We observe from Fig. 5.4 that the points do not lie strictly on a straight line. But they show an upward rising tendency where a straight line can be fitted. Let us draw the regression line along with the scatter plot.

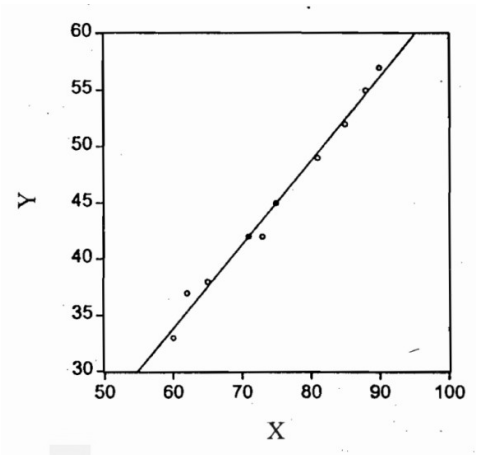

Fig. 5.5: Regression Line

The vertical difference between the regression line and the observations is the error e_i. The value corresponding to the regression line is called the predicted value or the expected value. On the other hand, the actual value of the dependent variable corresponding to a particular value of the independent variable is called the observed value. Thus 'error' is the difference between predicted value and observed value.

A question that arises is, 'How do we obtain the regression line? The procedure of fitting a straight line to the data is explained below.

5.9 MINIMISATION OF ERRORS

As mentioned earlier, a straight line can be represented by

$$Y_i = a + bX_i$$

where *b* is the *slope* and *a* is the *intercept* on y-axis. The location of a straight line depends on the value of *a* and *b*, called *parameters*. Therefore, the task before us is to *estimate* these parameters from the collected data. (You will learn more about the concept of estimation in Block 4). In order to obtain the line of best fit to the data we should find estimates of *a* and *b* in such a way that the error e_i is minimum.

In Fig. 5.4 these differences between observed and predicted values of Y are marked with straight lines from the observed points, parallel to y-axis, meeting the regression line. The lengths of these segments are the errors at the observed points.

Let us denote the *n* observations as before by (X_i, Y_i), $i = 1, 2, ..., n$. In Example 5.1 on agricultural production and rainfall, n=10.

Let us denote the predicted value of Y_i at X_i by $\hat{Y}_i$ (the notation $\hat{Y}_i$ is pronounced as 'Y_i-cap' or 'Y_i-hat'). Thus

$\hat{Y}_i = a + bX_i$, i = 1, 2, ..., n.

The error at the i^{th} point will then be

$$e_i = Y_i - \hat{Y}_i \qquad \ldots\ldots(5.11)$$

It would be nice if we can determine a and b in such a way that each of the e_i, i = 1, 2, ..., n is zero. But this is impossible unless it so happens that all the n points lie on a straight line, which is very unlikely. Thus we have to be content with minimising a combination of e_i, i = 1, 2, ..., n. What are the options before us?

- It is tempting to think that the total of all the e_i, i = 1, 2, ….., n, that is, $\sum_{i=1}^{n} e_i$ is a suitable choice. But it is not. Because, e_i for points above the line are positive and below the line are negative. Thus by having a combination of large positive and large negative errors, it is possible for $\sum_{i=1}^{n} e_i$ to be very small.
- A second possibility is that if we take $a = \bar{y}$ (the arithmetic mean of the Y_i's) and b = 0, $\sum_{i=1}^{n} e_i$ could be made zero. In this case, however, we do not need the value of X at all for prediction! The predicted value is the same irrespective of the observed value of X. This evidently is wrong.
- What then is wrong with the criterion $\sum_{i=1}^{n} e_i$? It takes into account the sign of e_i. What matters is the magnitude of the error and whether the error is on the positive side or negative side is really immaterial. Thus, the criterion $\sum_{i=1}^{n} |e_i|$ is a suitable criterion to minimise. Remember that $|e_i|$ means the absolute value of e_i. Thus, if e_i = 5 then $|e_i|$ = 5 and also if e_i = –5 then $|e_i|$=5. However, this option poses some computational problems.
- For theoretical and computational reasons, the criterion of *least squares* is preferred to the absolute value criterion. While in the absolute value criterion the sign of e_i is removed by taking its absolute value, in the *least squares criterion* it is done by squaring it. Remember that the squares of both 5 and –5 are 25. This device has been found to be mathematically and computationally more attractive.

We explain in detail the least squares method in the following section.

5.10 METHOD OF LEAST SQUARES

In the least squares method we minimise the sum of squares of the error terms, that is, $\sum_{i=1}^{n} e_i^2$.

From (5.9) we find that $e_i = Y_i - \hat{Y}_i$

which implies $e_i = Y_i - (a + bX_i) = Y_i - a - bX_i$.

Hence, $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - a - bX_i)^2$ …(5.12)

The next question is: How do we obtain the values of *a* and *b* to minimise (5.12)?

- Those of you who are familiar with the concept of differentiation will remember that the value of a function is minimum when the first derivative of the function is zero and second derivative is positive. Here we have to choose the value of *a* and *b*. Hence, $\sum_{i=1}^{n} e_i^2$ will be minimum when its partial derivatives with respect to *a* and *b* are zero. The partial derivatives of $\sum_{i=1}^{n} e_i^2$ are obtained as follows:

$$\frac{\partial \sum_i e_i^2}{\partial a} = \frac{\partial \sum_i (Y_i - a - bX_i)^2}{\partial a} = 2 \cdot (-1) \cdot \sum_i (Y_i - a - bX_i) \quad \ldots(5.13)$$

$$\frac{\partial \sum_i e_i^2}{\partial b} = \frac{\partial \sum_i (Y_i - a - bX_i)^2}{\partial b} = 2 \cdot (-X_i) \cdot \sum_i (Y_i - a - bX_i) \quad \ldots(5.14)$$

  By equating (5.13) and (5.14) to zero and re-arranging the terms we get the following two equations:

$$\sum_{i=1}^{n} Y_i = na + b\sum_{i=1}^{n} X_i \quad \ldots(5.15)$$

$$\sum_{i=1}^{n} X_i Y_i = a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2 \quad \ldots(5.16)$$

  These two equations, (5.15) and (5.16), are called the *normal equations* of least squares. These are two simultaneous linear equations in two unknowns. These can be solved to obtain the values of *a* and *b*.

- Those of you who are not familiar with the concept of differentiation can use a rule of thumb (We suggest that you should learn the concept of differentiation, which is so much useful in Economics). We can say that the normal equations given at (5.15) and (5.16) are derived by multiplying the coefficients of *a* and *b* to the linear equation and summing over all observations. Here the linear equation is $Y_i = a + bX_i$. The first normal equation is simply the linear equation $Y_i = a + bX_i$ summed over all observations (since the coefficient of *a* is 1).

  $\sum Y_i = \sum a + \sum bX_i$ or $\sum Y_i = na + b\sum X_i$

The second normal equation is the linear equation multiplied by X_i (since the coefficient of b is X_i)

$$\sum X_i Y_i = \sum aX_i + \sum bX_i^2 \quad \text{or} \quad \sum X_i Y_i = a\sum X_i + b\sum X_i^2$$

After obtaining the normal equations we calculate the values of a and b from the set of data we have.

Example 5.2: Assume that quantity of agricultural production depends on the amount of rainfall and fit a linear regression to the data given in Example 5.1.

In this case dependent variable (Y) is quantity of agricultural production and independent variable (X) is amount of rainfall. The regression equation to be fitted is

$$Y_i = a + bX_i + e_i$$

For the above equation we find out the normal equations by the method of least squares. These equations are given at (5.15) and (5.16). Next we construct a table as follows:

Table 5.5: Computation of Regression Line

X_i	Y_i	X_i^2	X_iY_i	$\hat{Y}_i$	e_i
60	33	3600	1980	33.85	-0.85
62	37	3844	2294	35.34	1.66
65	38	4225	2470	37.57	0.43
71	42	5041	2982	42.03	-0.03
73	42	5329	3066	43.51	-1.51
75	45	5625	3375	45.00	0.00
81	49	6561	3969	49.46	-0.46
85	52	7225	4420	52.43	-0.43
88	55	7744	4840	54.66	0.34
90	57	8100	5130	56.15	0.85
Total $\sum_i X_i$ = 750	$\sum_i Y_i$ =450	$\sum_i X_i^2$ =57294	$\sum_i X_iY_i$ =34526	$\sum_i \hat{Y}_i$ =450	$\sum_i e_i$ =0

By substituting values from Table 5.5 in the normal equations (5.15) and (5.16) we get the following:

$$450 = 10a + 750b$$

$$34526 = 750a + 57294b$$

By solving these two equations we obtain $a = -10.73$ and $b = 0.743$.

So the regression line is $\hat{Y}_i = -10.73 + 0.743X_i$.

Notice that the sum of errors $\sum_i e_i$ for the estimated regression equation in zero (see the last column of Table 5.5).

The computation given in Table 5.5 often involves large numbers and poses difficulty. Hence we have a short-cut method for calculating the values of *a* and *b* from the normal equations.

Let us take

$x = X - \bar{X}$ and $y = Y - \bar{Y}$ where $\bar{X}$ and $\bar{Y}$ are the arithmetic means of X and Y respectively.

Hence $xy = (X - \bar{X})(Y - \bar{Y})$

By re-arranging terms in the normal equations we find that

$$b = \frac{\sum_{i=1}^{n} xy}{\sum_{i=1}^{n} x^2} \qquad \ldots(5.17)$$

$$a = \bar{Y} - b\bar{X} \qquad \ldots(5.18)$$

You may recall that *covariance* is given by $\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$.

Moreover, variance of X is given by $\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$

Since $b = \frac{\sum_{i=1}^{n} xy}{\sum_{i=1}^{n} x^2}$ we can say that $b = \frac{\sigma_{xy}}{\sigma_x^2}$ …(5.19)

Since these formulae are derived from the normal equations we get the same values for *a* and *b* in this method also. For the data given in Table 5.4 we compute the values of a and b by this method. For this purpose we construct Table 5.6.

Table 5.6: Computation of Regression Line (short-cut method)

X_i	Y_i	x_i	y_i	x_i^2	$x_i\ y_i$
60	33	-15	-12	225	180
62	37	-13	-8	169	104
65	38	-10	-7	100	70
71	42	-4	-3	16	12
73	42	-2	-3	4	6
75	45	0	0	0	0
81	49	6	4	36	24
85	52	10	7	100	70
88	55	13	10	169	130
90	57	15	12	225	180
Total = 750	450	0	0	1044	776

On the basis of Table 5.6 we find that

$$\overline{X} = \frac{750}{10} = 75 \quad \text{and} \quad \overline{Y} = \frac{450}{10} = 45$$

$$b = \frac{\sum_{i=1}^{n} xy}{\sum_{i=1}^{n} x^2} = \frac{776}{1044} = 0.743$$

$$a = \overline{Y} - b\overline{X} = 45 - 0.743 \times 10 = -10.73$$

Thus the regression line in this method also $\hat{Y}_i = -10.73 + 0.743X_i$ …(5.20)

Coefficient *b* in (5.20) is called the regression coefficient. This coefficient reflects the amount of increase in Y when there is a unit increase in X. In regression equation (5.20) the coefficient *b* = 0.743 implies that if rainfall increases by 1 mm., agricultural production will increase 0.743 thousand tonne.

Regression coefficient is widely used. It is also an important tool of analysis. For example, if Y is aggregate consumption and X is aggregate income, *b* represents marginal propensity to consume (MPC).

5.11 PREDICTION

A major interest in studying regression lies in its ability to forecast. In Example 5.1 we assumed that the quantity of agricultural production is dependent on the amount of rainfall. We fitted a linear equation to the observed data and got the relationship

$$\hat{Y}_i = -10.73 + 0.743X_i$$

From this equation we can predict the quantity of agricultural output given the amount of rainfall. Thus when rainfall is 60 mm. agricultural production is $(-10.73 + 0.74 \times 60) = 33.85$ thousand tonnes. This figure is the *predicted value* on the basis of regression equation. In a similar manner we can find the predicted values of Y for different values of X.

Let us compare the predicted value with the observed value. From Table 5.4, where observed values are given, we find that when rainfall is 60 mm. agricultural production is 33 thousand tonnes. In fact, the predicted values $\hat{Y}_i$ for observed values of X are given in the fifth column of Table 5.5. Thus when rainfall is 60 mm. Predicted value is 33.85 thousand tonnes. Thus the error value e_i is –0.85 thousand tonne.

Now a question arises, 'Which one, between observed and predicted values, should we believe?' In other words, what will be the quantity of agricultural production if there is a rainfall of 60 mm. in future? On the basis of our regression line it is given to be 33.85 tonnes. And we accept this value because it is based on the overall data. The error of –0.85 is considered as a random fluctuation which may not be repeated.

The second question that comes to our mind is, 'Is the prediction valid for any value of X?' For example, we find from the regression equation that when rainfall is zero, agricultural production is –10.73 thousand tonne. But common sense tells us that agricultural production cannot be negative! Is there anything wrong with our regression equation? In fact, the regression equation here is estimated on the basis of rainfall data in the range of 60-90 mm. Thus prediction is be valid in this range of X. Our prediction should not be for far off values of X.

A third, question that arises here is, 'Will the predicted value come true?' This depends upon the *coefficient of determination*. If the coefficient of determination is closer to one, there is greater likelihood that the prediction will be realised. However, the predicted value is constrained by elements of randomness involved with human behaviour and other unforeseen factors.

5.12 RELATIONSHIP BETWEEN REGRESSION AND CORRELATION

In regression analysis the status of the two variables (X, Y) are different such that Y is the variable to be predicted and X is the variable, information on which is to be used. In the rainfall-agricultural production problem, it makes sense to predict agricultural production on the basis of rainfall and it would not make sense to try and predict rainfall on the basis of agricultural production. However, in the case of scores in Economics and Statistics (see Table 5.1), either one could be X and the other Y. Hence we consider the two prediction problems: (i) predicting Economics score (Y) from Statistics score (X); and (ii) predicting Statistics score (X) from Economics score (Y).

Thus, we can have two regression coefficients from a given set of data depending upon the choice of dependent and independent variables. These are:

a) Y on X line, $Y_i = a + bX_i$

b) X on Y line, $X_i = \alpha + \beta Y_i$

You may ask, 'What is the need for having two different lines? By rearrangement of terms of the Y on X line we obtain $X_i = -\frac{a}{b} + \frac{1}{b}Y_i$. Thus we should have $\alpha = -\frac{a}{b}$ and $\beta = \frac{1}{b}$. However, the observations are not on a straight line and the relation between X and Y is not a mathematical one. You may recall that estimates of the parameters are obtained by the method of least squares. Thus the regression line $\hat{Y}_i = a + bX_i$ is obtained by minimising $\sum_i (Y_i - a - bX_i)^2$ whereas the regression line $\hat{X}_i = \alpha + \beta Y_i$ is obtained by minimising $\sum_i (X_i - \alpha - \beta Y_i)^2$.

However, there is a relationship between the two regression coefficients b and β. We have noted earlier that $b = \frac{\sigma_{xy}}{\sigma_x^2}$. By a similar formula by interchanging the roles of X and Y we find $\beta = \frac{\sigma_{xy}}{\sigma_y^2}$. But by definition we notice that $\sigma_{xy} = \sigma_{yx}$.

Thus $b \times \beta = \frac{\sigma_{xy}^2}{\sigma_x^2 \times \sigma_y^2}$, which is the same as r^2.

This r^2 is called the *coefficient of determination*. Thus the product of the two regression coefficients of Y on X and X on Y is the square of the correlation coefficient. This gives a relationship between correlation and regression. Notice, however, that the coefficient of determination of either regression is the same, i.e., r^2; this means that although the two regression lines are different, their predictive powers are the same. Note that the coefficient of determination r^2 ranges between 0 and 1, i.e., the maximum value it can assume is unity and the minimum value is zero; it cannot be negative.

From the previous discussions, two points emerge clearly:

1) If the points in the scatter lie close to a straight line, then there is a strong relationship between X and Y and the correlation coefficient is high.
2) If the points in the scatter diagram lie close to a straight line, then the observed values and predicted values of Y by least squares are very close and the prediction errors $(Y_i - \hat{Y}_i)$ are small.

Thus, the prediction errors by least squares seem to be related to the correlation coefficient. We explain this relationship here. The sum of squares of errors at the various points upon using the least squares linear regression is $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$.

On the other hand, if we had not used the value of observed X to predict Y, then the prediction would be a constant, say, a. The best value of a by least squares criterion is such an a that minimises $\sum_{i=1}^{n}(Y_i - a)^2$; the solution to this a is seen to be $\bar{Y}$. Thus the sum of squares of errors of prediction at various points without using X is $\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$.

The ratio, $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \Big/ \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ can then be used as an index of how much has been gained by the use of X. In fact, this ratio is the coefficient of determination and same as r^2 mentioned above. Since both the numerator and denominator of this ratio are non-negative, the ratio is greater than or equal to zero.

Check Your Progress 3

1) From the following data find the coefficient of linear correlation between X and Y. Determine also the regression line of Y on X, and then make an estimate of the value of Y when $X = 12$.

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

.....

……

……

……

……

……

2) Obtain the lines of regression for the following data:

(X)	1	2	3	4	5	6	7	8	9
(Y)	9	8	10	12	11	13	14	16	15

……

……

……

……

……

3) Find the two lines of regression from the following data:

Age of Husband (X)	2	2	2	2	3	2	2	4	2	1
	5	2	8	6	5	0	2	0	0	8
Age of Wife (Y)	1	1	2	1	2	1	1	2	1	1
	8	5	0	7	2	4	6	1	5	4

Hence estimate (i) age of husband when the age of wife is 19, (ii) the age of wife when the age of husband is 30.

……

……

……

……

……

4) From the following data, obtain the two regression equations :

Sales	:	91	97	108	121	67	124	51	73	111	57
Purchases	:	71	75	69	97	70	91	39	61	80	47

……

……

……

……

……

5) Obtain the equation of the line of regression of yield of rice (y) on water (x) from the data given in the following table :

Water in inches (x)	12	18	24	30	36	42	48
Yield in tons (y)	5.27	5.68	6.25	7.21	8.02	8.71	8.42

Estimate the most probable yield of rice for 40 inches of water.

……

……

……

……

……

5.13 MULTIPLE REGRESSION

So far we have considered the case of the dependent variable being explained by one independent variable. However, there are many cases where the dependent variable is explained by two or more independent variables. For example, yield of crops (Y) being explained by application of fertilizer (X_1) and irrigation water(X_2). This sort of models is termed multiple regression. Here, the equation that we consider is

$$Y = \alpha + \beta X_1 + \gamma X_2 + e \qquad \ldots(5.21)$$

Where Y is the explained variable, X_1 and X_2 are explanatory variables, and e is the error term. In order to make the presentation simple we have dropped the subscripts. A regression equation can be fitted to (5.21) by applying the method of least squares. Here also we minimise $\sum e^2$ and obtain the normal equations as follows:

$$\begin{aligned} \Sigma Y &= n\alpha + \beta\Sigma X_1 + \gamma\Sigma X_2 \\ \Sigma X_1 Y &= \alpha\Sigma X_1 + \beta\Sigma X_1^2 + \gamma\Sigma X_1 X_2 \\ \Sigma X_2 Y &= \alpha\Sigma X_2 + \beta\Sigma X_1 X_2 + \gamma\Sigma X_2^2 \end{aligned} \qquad \ldots (5.22)$$

By solving the above equations we obtain estimates for α, β and γ. The regression equation that we obtain is

$$\hat{Y} = \alpha + \beta X_1 + \gamma X_2 \qquad \ldots(5.23)$$

Remember that we obtain predicted or forecast values of Y (that is $\hat{Y}$) through (5.23) by applying various values for X_1 and X_2.

In the bivariate case (Y,X) we could plot the regression line on a graph paper. However, it is quite complex to plot the three variable case (Y, X_1, X_2) on graph paper because it will require three dimensions. However, the intuitive idea remains the same and we have to minimise the sum of errors. In fact when we add all the error terms ($e_1, e_2, \ldots\ldots e_n$) it sum up to zero.

In many cases the number of explanatory variables may be more than two. In such cases we have to follow the basic principle of least squares: minimize Σe^2. Thus if $Y = a_0 + a_1 X_1 + a_2 X_2 + \ldots\ldots\ldots\ldots + a_n X_n + e$ then we have to minimize

$$\Sigma e^2 = \Sigma (Y - a_0 - a_1 X_1 - a_2 X_2 \ldots\ldots - a_n X_n)^2$$

and find out the normal equations.

Now a question arises, 'How many variables should be added in a regression equation?' It depends on our logic and what variables are considered to be important. Whether a variable is important or not can be identified on the basis of statistical tests also. These tests will be discussed later in Block 4.

We present a numerical example of multiple regression below.

Example 5.3

A student tries to explain the rent charged for housing near the University. She collects data on monthly rent, area of the house and distance of the house from the university campus and fits a linear regression model.

Rent (in Rs.'000) Y	Area (in sq.mt.) X_1	distance(in Km.) X_2
20	65	5.7
25	66	3.2
26	70	7.5
28	70	6.5
30	75	5.0
31	76	4.0
32	72	6.0
33	75	6.2
35	78	3.5
40	103	2.4

In the above example rent charged (Y) is the dependent variable while area of the house (X_1) and distance of the house from the university campus (X_2) are independent variables.

The steps involved in estimation of regression line are:

i) Find out the regression equation to be estimated. In this case it is given by $Y = \alpha + \beta X_1 + \gamma X_2 + e$.

ii) Find out the normal equations for the regression equation to be estimated. In this case the normal equations are

$$\Sigma Y = n\alpha + \beta \Sigma X_1 + \gamma \Sigma X_2$$
$$\Sigma X_1 Y = \alpha \Sigma X_1 + \beta \Sigma X_1^2 + \gamma \Sigma X_1 X_2$$
$$\Sigma X_2 Y = \alpha \Sigma X_2 + \beta \Sigma X_1 X_2 + \gamma \Sigma X_2^2$$

iii) Construct a table as given in Table 9.4.

iv) Put the values from the table in the normal equations.

v) Solve for the estimates of α, β and γ.

Table 5.7: Computation of Multiple Regression

Y	X_1	X_2	X_1Y	X_2Y	X_1^2	X_2^2	X_1X_2	$\hat{Y}$	e_i
20	65	5.7	1300	114	4225	32.49	370.5	25.49	-5.49
25	66	3.2	1650	80	4356	10.24	211.2	25.71	-0.71
26	70	7.5	1820	195	4900	56.25	525	27.94	-1.94
28	70	6.5	1960	182	4900	42.25	455	27.85	0.15
30	75	5	2250	150	5625	25	375	30.00	0.00
31	76	4	2356	124	5776	16	304	30.37	0.63
32	72	6	2304	192	5184	36	432	28.72	3.28
33	75	6.2	2475	204.6	5625	38.44	465	30.11	2.89
35	78	3.5	2730	122.5	6084	12.25	273	31.24	3.76
40	103	2.4	4120	96	10609	5.76	247.2	42.58	-2.58
300	750	50	225000	15000	562500	2500	37500	300	0

By applying the above mentioned steps we obtain the estimated regression line as

$\hat{Y} = -4.80 + 0.45X_1 + 0.09X_2$.

5.14 NON-LINEAR REGRESSION

The equation fitted in regression can be non-linear or curvilinear also. In fact, it can take numerous forms. A simpler form involving two variables is the quadratic form. The equation is

$Y = a + bX + cX^2$

There are three parameters here viz., a, b and c and the normal equations are:

$$\Sigma Y = n\alpha + b\Sigma X + c\Sigma X^2$$
$$\Sigma XY = \alpha\Sigma X + b\Sigma X^2 + c\Sigma X^3$$
$$\Sigma X^2 Y = \alpha\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4$$

By solving for these equation we obtain the values of a, b and c.

Certain non-linear equations can be transformed into linear equations by taking logarithms. Finding out the optimum values of the parameters from the transformed linear equations is the same as the process discussed in the previous section. We give below some of the frequently used non-linear equations and the respective transformed linear equations.

1) $Y = a\,c^{bx}$

By taking natural log (ln), it can be written as

$\ln Y = \ln a + bX$

or $Y' = \alpha + \beta X'$

Where, $Y' = \ln Y$, $\alpha = \ln a$, $X' = X$ and $\beta = b$

2) $Y = aX^b$

By taking logarithm (log), the equation can be transformed into

$\log Y = \log a + b \log X$

or $Y' = \alpha + \beta X'$

where, $Y' = \log Y$, $\alpha = \log a$, $\beta = b$ and $X' = \log X$

3) $Y = \dfrac{1}{a + bX}$

If we take $Y' = \dfrac{1}{Y}$ then

$Y' = a + bX$

4) $Y = a + b\sqrt{X}$

If we take $X' = \sqrt{X}$ then

$Y = a + bX'$

Once the non-linear equation is transformed, the fitting of a regression line is as per the method discussed in the beginning of this Unit.

We derive the normal equations and substitute the values calculated from the observed data. From the transformed parameters, the actual parameters can be obtained by making the reverse transformation.

Check Your Progress 4

1) Using the data on scores in Statistics and Economics of Table 5.1, compute the regression of y on x and x on y and check that the two lines are different. On the scatter diagram, plot both these regression lines. Check that the product of the regression coefficients is the square of the correlation coefficient.

……

……

……

……

……

2) Suppose that the least squares linear regression of family expenditure on clothing (Rs. y) on family annual income (Rs. x) has been found to be $y = 100 + 0.09x$, in the range $1000 < x < 100000$. Interpret this regression line. Predict the expenditure on the clothing of a family with an annual income of Rs. 10,000. What about families with annual income of Rs. 100 and Rs. 10,00,000?

……

……

……

……

……

5.15 LET US SUM UP

In this Unit we discussed an important statistical tool, that is, regression. In regression analysis we have two types of variables: dependent and independent. The dependent variable is explained by independent variables. The relationship between variable takes the form of a mathematical equation. Based on our logic, understanding and purpose of analysis we categorise variables and identify the equation form.

The regression coefficient enables us to make predictions for the dependent variable given the values of the independent variable. However, prediction remains more or less valid within the range of data used for analysis. If we attempt to predict for far off values of the independent variable we may get insensible values for the dependent variable.

5.16 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) + 0.47

2) + 0.996

3) + 0.98

4) + 0.995

5) – 0.84

Check Your Progress 2

1) 2/3

2) + 0.64

3) – 0.21, + 0.64, –0.30

4) + 0.82

Check Your Progress 3

1) + 0.98 ; $y = 0.64x + 0.54$; 8.2

2) $x = 0.95y - 6.4$; $y = 0.95x + 7.25$

3) $x = 2.23y - 12.70$; $y = 0.39x + 7.33$

(i) 29.6 (ii) 18.9

4) $y = 0.613x + 14.81$; $x = 1.360y - 5.2$

5) $y = 3.99 + 0.103x$; 8.11 tons

Check Your Progress 4

1) (i) $y = a + bx = 5.856 + 0.676x$

(ii) $x = \alpha + \beta y = 29.848 + 0.799y$

(iii) $r = 0.73$

(iv) $0.676 \times 0.799 = 0.54$

2) Expenditure on clothing, when family income is Rs. 10,000, is Rs. 1,000. In case of income below 1,000 or above 1,00,000 the regression line may not hold good. In between both these figures, one rupee increase in income increases expenditure on clothes by 9 paise.