# UNIT 9 REGRESSION ANALYSIS

**Structure**



## 9.0 OBJECTIVES

After going through this unit, you should be able to:

- 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.

## 9.1 INTRODUCTION

In the previous Unit 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, to be discussed in this Unit, 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.

## 9.2 THE CONCEPT OF REGRESSION

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.

## 9.3 LINEAR RELATIONSHIP: TWO VARIABLE CASE

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

$$Y_i = a + bX_i \qquad \ldots(9.1)$$

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 (9.1) 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 (9.1). 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 (9.1) 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 (9.1). Thus our stochastic model is

$$Y_i = a + bX_i + e_i \qquad \ldots(9.2)$$

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 (9.2) has two parts, viz., i) deterministic part (that is, $a + bX_i$), and ii) stochastic or randomness part (that is, $e_i$). Equation (9.2) 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 (9.2) on a graph paper the observations will not remain on a straight line.

**Example 9.1**

The amount of rainfall and agricultural production for ten years are given in Table 9.1.

**Table 9.1: 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 |

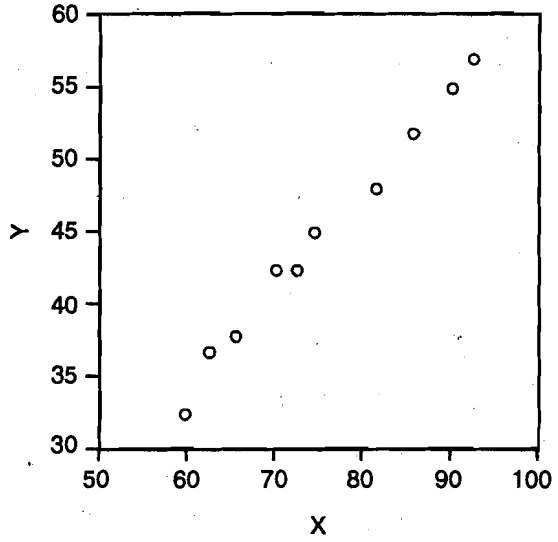


**Fig. 9.1: Scatter Diagram**

We plot the data on a graph paper. The scatter diagram looks something like Fig. 9.1. We observe from Fig. 9.1 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.

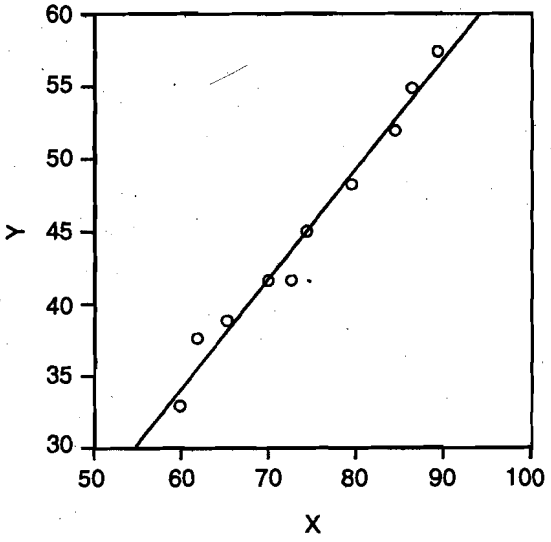


**Fig. 9.2: Regression Line**

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.

## 9.4 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 7). 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. 9.1 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 9.1 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

$$e_i = Y_i - \hat{Y}_i \qquad \ldots (9.3)$$

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, 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.

## 9.5 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 (9.3) 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, $$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - a - bX_i)^2 \qquad \ldots(9.4)$$

The next question is: How do we obtain the values of $a$ and $b$ to minimise (9.3)?

- 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:

$$\frac{\partial \sum_i e_i^2}{\partial a} = \frac{\partial \sum_i (Y_i - a - bX_i)^2}{\partial a} = 2 \cdot (-1) \cdot \sum_i (Y_i - a - bX_i) \qquad \ldots(9.5)$$

$$\frac{\partial \sum_i e_i^2}{\partial b} = \frac{\partial \sum_i (Y_i - a - bX_i)^2}{\partial b} = 2 \cdot (-X_i) \cdot \sum_i (Y_i - a - bX_i) \qquad \ldots(9.6)$$

By equating (9.5) and (9.6) to zero and re-arranging the terms we get the following two equations:

$$\sum_{i=1}^{n} Y_i = na + b\sum_{i=1}^{n} X_i \quad \ldots(9.7)$$

$$\sum_{i=1}^{n} X_i Y_i = a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2 \quad \ldots(9.8)$$

These two equations, (9.7) and (9.8), 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 (9.7) and (9.8) 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 $X_i$)

$\sum X_i Y_i = \sum aX_i + \sum bX_i^2$ or $\sum X_i Y_i = a\sum X_i + b\sum X_i^2$

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

**Example 9.2**: Assume that quantity of agricultural production depends on the amount of rainfall and fit a linear regression to the data given in Example 9.1.

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 (9.7) and (9.8). Next we construct a table as follows:

**Table 9.2: 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 |
| $\sum_i X_i = 750$ | $\sum_i Y_i = 450$ | $\sum_i X_i^2 = 57294$ | $\sum_i X_iY_i = 34526$ | $\sum_i \hat{Y}_i = 450$ | $\sum_i e_i = 0$ |

By substituting values from Table 9.2 in the normal equations (9.7) and (9.8) we get the following:

$$\begin{aligned} 450 &= 10a + 750b \\ 34526 &= 750a + 57294b \end{aligned}$$

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 9.2).

The computation given in Table 9.2 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

$$b = \frac{\sum_{i=1}^{n} xy}{\sum_{i=1}^{n} x^2} \qquad \ldots(9.9)$$

$$a = \bar{Y} - b\bar{X} \qquad \ldots(9.10)$$

You may recall from Unit 8 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 $b = \frac{\sigma_{xy}}{\sigma_x^2}$ $\qquad \ldots(9.11)$

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 9.1 we compute the values of a and b by this method. For this purpose we construct Table 9.3.

**Table 9.3: 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 9.3 we find that

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

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

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

Thus the regression line in this method also $\hat{Y}_i = -10.73 + 0.743X_i$ ...(9.12)

Coefficient $b$ in (9.12) 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 (9.12) the coefficient $b = 0.743$ implies that if rainfall increase 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).

## 9.6 PREDICTION

A major interest in studying regression lies in its ability to forecast. In Example 9.1 in the previous Section 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.

Compare the predicted value with the observed value. From Table 9.1 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 9.2. Thus when rainfall is 60 mm. predicted value is 33.85 thousand tonnes. Thus the error value 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.

## 9.7 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 Example 8.1 in the previous Unit), 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 $\beta$.

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 discussion, 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 $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$.

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 $\sum_{i=1}^{n} (Y_i - \bar{Y})^2$.

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 1

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 $(X)$ | 25 | 22 | 28 | 26 | 35 | 20 | 22 | 40 | 20 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|
| Age of Wife $(Y)$ | 18 | 15 | 20 | 17 | 22 | 14 | 16 | 21 | 15 | 14 |

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

.....

.....

.....

.....

.....

.....

## 17.9 NON-PROBABILITY SAMPLING PROCEDURES

There are different types of non-probability sampling such as:

1) Convenience Sampling
2) Judgment Sampling
3) Quota Sampling
4) Snowball Sampling

We discuss the procedure of drawing a non-probability sampling below.

### 17.9.1 Convenience Sampling

This is one of the most commonly used methods of non-probability sampling. In this method the researcher's convenience forms the basis for selection of the sample. Especially for a exploratory research there is a pressing need for data. In such situations the selection of sampling units is left to the interviewer. The population units are included in the sample simply because they are in the right place at the right time. This method is often used during preliminary research efforts to get a gross estimate of the results, without incurring the cost or time required to select a sample. For example, during the budget session or when the price of a product is increased or a new government is formed, convenience samples are used by the researchers/journalists to reflect public opinion. Convenience samples are extensively used in marketing research.

The advantage of convenience sampling is that it is less expensive and less time-consuming. The limitations of convenience sampling are: (a) it involves sample selection bias, and (b) it does not provide a representative sample of the population and therefore we cannot generalise the results.

### 17.9.2 Judgment Sampling

This is another commonly used non-probability sampling procedure. This procedure is often referred to as *purposive sampling*. In this procedure the researcher selects the sample based on his/her judgment. The researcher believes that the selected sample elements are representative of the population. For example, the calculation of consumer price index is based on judgment sampling. Here the sample consists of a basket of consumer items and other goods and services which are expected to reflect a representative sample. The prices of these items are collected from selected cities that are viewed as typical cities with demographic profiles matching the national profile.

The advantage of judgment sampling is that it is low cost, convenient and quick. The disadvantage is that it does not allow direct generalisations to population. The quality of the sample depends upon the judgment of the researcher.

### 17.9.3 Quota Sampling

In this procedure the population is divided into groups based on some characteristics such as gender, age, education, religion, income group, etc. A quota of units from each group is determined. The quota may be either proportional or non-proportional. The proportional quota sampling is based on the proportion of each characteristic in the population so that the proportion in the sample represents the population proportion. For example, if you know that there are 80% of the households whose income is below say Rs.100000 per annum and 20% households

have income above Rs.100000 per annum in a city. You want to take a sample of size 100 households. Then you include 80 households from below Rs.100000 income and 20 households from above Rs.100000 income. The objective here is to meet the proportional quota of sampling from each characteristic in the population.

The non-proportional quota sampling is a bit less restrictive. In this procedure, you specify the minimum number of sampled units from each group. You are not concerned with having proportions in the population. For instance, in the above example you may simply interview 50 households from each income group instead of 80% and 20%. The interviewer is instructed to fill the quota for each group based on convenience or judgment. The very purpose of quota sampling is that various groups in the population are represented to the extent the investigator desires.

Do not confuse the quota sampling with stratified sampling that you have learned earlier. In stratified sampling you select random samples from each stratum or group whereas in quota sampling the interviewer has a fixed quota. For example, in a city there are five market centres. A company wants to assess the demand for its new product and sends 5 investigators to assess the demand by interviewing 50 prospective customers from each market. It is left to the investigator whom he/she will interview at each market centre. If the product is targeted to women, this way you cannot elicit the information among various groups of women customers like housewives or employed women or young or old. In this sampling you are simply fixing a quota for each investigator.

The quota sampling has the advantage over others if the sample meets the characteristics of the population that you are looking into. In addition, the cost and time involved in collecting the data are greatly reduced. However, there are many disadvantages as well. In quota sampling, the samples are selected according to the convenience of the investigator instead of selecting random samples. Therefore, the selected samples may be biased. If there are a large number of characteristics on the basis of which the quotas are fixed, then it becomes very difficult to fix the quotas/sub-quotas for each group/sub-group. Also the investigators have the tendency to collect information only from those who are willing to provide information and avoid unwilling respondents.

### 17.9.4 Snowball Sampling

In snowball sampling, we begin by identifying someone who meets the criteria for inclusion in our study. We then ask him/her to recommend others who also meets the criteria. Although this method would hardly lead to representative samples, there are times when it may be the best method available. Snowball sampling is especially useful when we are trying to reach populations that are inaccessible or hard to find. For example, if we are studying the homeless, we are not likely to find good lists of homeless people within a specific geographical area. However, if we go to that area and identify one or two, we may find that they know very well who the other homeless people in their vicinity are and how we can find them.

## 17.10 DETERMINING THE SAMPLE SIZE

The use of appropriate sampling procedure is necessary for a representative sample. However, this condition is not sufficient. In addition to the above, we should

determine the sample size. The question of how large a sample should be is a difficult one. Sample size can be determined by various considerations. The following are some of the considerations in determining the sample size:

a) Sampling error
b) Number of comparisons to be made
c) Response rates
d) Funds available

a) **Sampling Error:** In Unit 16 you have learned that smaller samples have greater sampling error than large samples. On the other hand, larger samples have larger non-sampling errors than smaller samples. The sampling error is a number that describes the precision of an estimate of the sample. It is usually expressed as a margin of error associated with a statistical level of confidence. For example, for a prime minister preferential poll you may say that the incumbent is favored by 65% of votes, with a margin of error (precision) of plus or minus 5 percentage points at a 95% confidence level. This means that if the same surveys were conducted with 100 different samples of voters, 95 of the surveys would be expected to show the incumbent favoured by between 60% and 70% of the voters (65% ± 5%). Remember as you increase the precision level of your results you need larger sample size.

b) **Number of Comparisons to Make:** Sometimes we may be interested in making comparisons of two or more groups (strata) in the sample. For example, we may want to make the comparison between male and female respondents or between urban and rural respondents. Or we may want to compare the results for 4 geographical regions of the country say north, south, west and east. Then we need an adequate sample size in each region or stratum of the population. Therefore, the heterogeneity of population characteristics plays a significant role in deciding the sample size.

c) **Response Rates:** In mail surveys, we know that all those questionnaires mailed to the respondents may not reach us back after filling the questionnaires. As per the experiences on mail survey, the response rate ranges between 10% to 50%. Then, if you are expecting a 20% response rate, for example, you will have to mail 5 times the number of sample size required.

d) **Funds Available:** The funds available may influence the sample size. If the funds available for the study are limited then you may not be able to spend more than a certain amount of the total money available with you on collecting the data.

It is even more difficult to decide the sample size, when you use the non-probability sampling procedures. This is because there are no definite rules to be followed in non-probability sampling procedures. It all depends upon on what you want to know, the purpose of inquiry, what will be useful, what will have credibility and what can be done with available time and resources. In purposive sampling, the sample should be judged on the basis of purpose. In non-probability sampling procedures, the validity, meaningfulness, and insights generated have more to do with the information-richness of the sample units selected rather than the sample size.

### Some Formulae to Determine the Sample Size

Technical considerations suggest that the required sample size is a function of the precision of the estimates you wish to achieve, the variance of the population and

the confidence level you wish to use. If you want more precision and confidence level then you may need larger sample size. The more frequently used confidence levels are 95% and 99%. And the more frequently used precision levels are 1% and 5%. There are different formulae used to determine the sample size depending upon various considerations discussed above. In this section we will discuss three of them.

i) If we wish to report the results as percentages (proportions) of the sample responding, we use the following formula:

$$n_i = \frac{P_i(1-P_i)}{\frac{\alpha^2}{z^2} + \frac{P_i(1-P_i)}{N_i}}$$

Where, $n_i$ = sample size of the $i^{th}$ attribute required

$P_i$ = estimated proportion of the population possessing $i^{th}$ attribute of interest (for example, proportion of males, females, urban, rural, etc.)

$\alpha$ = precision required (0.01, 0.05 etc.)

z = standardized value indicating the confidence level (z=1.96 at 95% confidence level and z=2.58 at 99% confidence level)

$N_i$ = population size of the $i^{th}$ attribute (known or estimated)

**Example 17.8**: A population consists 80% rural and 20% urban people. Given that the population size is 50000, determine the sample size required. Assume that the desired precision and confidence levels are 1% and 99% respectively. In this example,

$P_1$ = proportion of rural people = 0.80

$P_2$ = proportion of urban people = 0.20

$N_1$ = rural population size = 50000 × 0.80 = 40000

$N_2$ = urban population size = 50000 × 0.20 = 10000

$\alpha$ = 0.01

z = 2.58 (at 99% confidence level)

The required sample size is

$$n_1 = \text{rural sample} = \frac{P_1(1-P_1)}{\frac{\alpha^2}{z^2} + \frac{P_1(1-P_1)}{N_1}}$$

$$= \frac{0.80(1-0.80)}{\frac{0.01^2}{2.58^2} + \frac{0.80(1-0.80)}{40000}}$$

$$= \frac{0.80(0.20)}{\frac{0.0001}{6.6564} + \frac{0.80(0.20)}{40000}}$$

$$= \frac{0.16}{0.000019 + \frac{0.16}{40000}}$$

$$= \frac{0.16}{0.000019 + 0.000004}$$

$$= \frac{0.16}{0.000023} = 8410.8 \text{ or say } 8411$$

$n_2$ = urban sample = $\dfrac{P_2(1-P_2)}{\dfrac{\alpha^2}{z^2} + \dfrac{P_2(1-P_2)}{N_2}}$

$$= \frac{0.20(1-0.20)}{\dfrac{0.01^2}{2.58^2} + \dfrac{0.20(1-0.20)}{10000}}$$

$$= \frac{0.20(0.80)}{\dfrac{0.0001}{6.6564} + \dfrac{0.20(0.80)}{10000}}$$

$$= \frac{0.16}{0.000019 + \dfrac{0.16}{10000}}$$

$$= \frac{0.16}{0.000019 + 0.000016}$$

$$= \frac{0.16}{0.000035} = 4568.4 \text{ or say } 4568$$

Therefore we need to have a sample of size 8411 + 4568 = 12979 units.

ii) If we wish to report the results as means (averages) of the sample responding, we use the following formula:

$$n_i = \frac{P_i^2}{\dfrac{\alpha^2}{z^2} + \dfrac{P_i^2}{N_i}}$$

Where, $n_i$ = sample size of the $i^{th}$ attribute required

$P_i$ = estimated standard deviation of the $i^{th}$ attribute of interest (for example, average income of high income group, low income group etc.)

$\alpha$ = precision required (0.01 or 0.05 as the case may be)

$z$ = standardized value indicating the confidence level (z=1.96 at 95% confidence level and z=2.58 at 99% confidence level)

$N_i$ = population size of the $i^{th}$ attribute (known or estimated)

**Example 17.9**: It is planned to conduct a study to know the average income of households. Given that the standard deviation of households is 2.5 and the population size is 10000, determine the sample size required. Assume that the desired precision and confidence levels are 5% and 95% respectively.

In this example,

$P_1$ = standard deviation of income = 2.5

$N_1$ = number of households = 10000

$\alpha$ = 0.05
z = 1.96 (at 95% confidence level)

The required sample size is

$$n_1 = \frac{P_1^2}{\frac{\alpha^2}{z^2} + \frac{P_1^2}{N_1}}$$

$$= \frac{2.5^2}{\frac{0.05^2}{1.96^2} + \frac{2.5^2}{10000}}$$

$$= \frac{6.25}{\frac{0.0025}{3.8416} + \frac{6.25}{10000}}$$

$$= \frac{6.25}{0.000651 + 0.000625}$$

$$= \frac{6.25}{0.001276} = 4898$$

iii) If we wish to report the results in a variety of ways or we have the difficulty in estimating the proportion or standard deviation of the attribute of interest, we use the following formula:

$$n = \frac{0.25}{\frac{\alpha^2}{z^2} + \frac{0.25}{N}}$$

Where, $n$ = sample size required
$\alpha$ = precision required (0.01 or 0.05 as the case may be)
z = standardized value indicating the confidence level (z=1.96 at 95% confidence level and z=2.58 at 99% confidence level)
$N$ = population size (known or estimated)

**Example 17.10**: Given that the population size is 10000, determine the sample size required when desired precision and confidence levels are 5% and 99% respectively.

In this example,
$N$ = 10000
$\alpha$ = 0.05
z = 2.58 (at 99% confidence level)

The required sample size is

$$n = \frac{0.25}{\frac{0.05^2}{2.58^2} + \frac{0.25}{10000}}$$

$$n = \frac{0.25}{\frac{0.0025}{6.6564} + \frac{0.25}{10000}}$$

$$n = \frac{0.25}{0.0003756 + 0.000025} = \frac{0.25}{0.000401} = 624$$

**Check Your Progress 2**

1) Say whether the following statements are true or false.
   a) When the units included in the sample are based on judgment of the investigator, the sampling is said to be random.
   b) With increasing sample size the sampling error decreases.
   c) Convenience sampling has the disadvantage that it may not be representative sample.
2) One of the major disadvantage of judgment sampling is
   a) The procedure is very cumbersome
   b) The sample selection depends on the individual judgment of the investigator
   c) It gives small sample size
   d) It is very expensive.

## 17.11 LET US SUM UP

The most commonly used probability sampling procedure is the simple random sampling which allows a chance to all population units to be included in the sample. The sample units are chosen using random number tables. A systematic random sample uses the first sample unit at random as a starting point and the subsequent sample units are chosen systematically. A stratified sample guarantees inclusion of units from each stratum. A cluster sample involves complete enumeration of one or more randomly selected clusters.

The non-probability sampling procedures include convenience sampling, judgment sampling, quota sampling and snowball sampling. These sampling procedures are not independent from sampling bias but still popular in some situations particularly marketing research.

A number of factors decide the sample size. It may be the number of groups in the population, the heterogeneity of population, funds and time available, etc.

Using a sample saves a lot of money, time and manpower. If a suitable sampling procedure is used in selecting units, appropriate sample size is selected and necessary precautions are taken to reduce sampling errors, then a sample should yield a valid and reliable information about the population.

## 17.12 KEY WORDS

**Cluster Sampling** : It is a sampling procedure where the entire population is divided into groups called clusters and then a random number of clusters are selected. All observations in the selected clusters are included in the sampling.

**Convenience Sampling** : It refers to the method of obtaining a sample that is most conveniently available to the researcher.

**Judgment Sampling** : In this sampling procedure the selection of sample is based on the researcher's judgment about some appropriate characteristic required of the sample units.

**Multistage Sampling** : The sample selection is done in a number of stages.

**Quota Sampling** : In this sampling procedure the samples are selected on the basis of some parameters such as age, gender, geographical region, education, income, religion, etc.

**Random Sampling** Random sampling is a sampling technique where we select sample from a population. Here, each unit of the population has a chance of being included in the sample.

**Simple Random Sampling** : It is the basic sampling procedure when we select samples using lottery method or using random number tables.

**Snowball Sampling** : Snowball sampling relies on referrals from initial sampling units to generate additional sampling units.

**Stratified Sampling** : In this sampling procedure the population is divided into groups called strata and then the samples are selected from each stratum using a random sampling method.

**Systematic Sampling** : A sampling procedure in which units are selected from the population at uniform interval that is measured in time, order or space.

## 17.13 SOME USEFUL BOOKS

Kothari, C.R.(1985) *Research Methodology : Methods and Techniques*, Wiley Eastern, New Delhi.

Levin, R.I. and D.S. Rubin. (1999) *Statistics for Management*, Prentice-Hall of India, New Delhi

Mustafi, C.K.(1981) *Statistical Methods in Managerial Decisions*, Macmillan, New Delhi.

Plane, D.R. and E.B. Oppermann. (1986) *Business and Economic Statistics*, Business Publications, Inc: Plano.

Zikmund, William G. (1988) *Business Research Methods*, The Dryden Press, New York.

## 17.14 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

**Check Your Progress 1**

1) d
2) c
3) a) False
   b) True
4) c)

**Check Your Progress 2**

1) a) False
   b) True
   c) True
2) b)