10

REGRESSION ANALYSIS

10.1 INTRODUCTION

In the previous chapter we have studied the correlation analysis which explains only the causal relationship between two variables. Correlation analysis fails to explain the effect of change in one variable on the value of another variable. In other words, the correlation analysis does not explain the cause and effect relationship between variables. The statistical tool with the help of which we can predict or determine the average change in one variable due to a certain amount of change in other variable is called *regression analysis.*

10.2 MEANING AND DEFINITIONS

Regression indicates the average relationship between two or more variables and from this average relationship the average value of one variable is estimated corresponding to a given value of the other variable. The average value is also termed as *the most probable value* or *the most likely value.* The variable corresponding to whose value, the average value of the other variable is estimated is called the *independent* or *explanatory* or *exogenous variable.* The variable whose average value is estimated is called *dependent* or *explained* or *endogenous variable.*

Literally speaking the term regression means 'going back' or 'moving backward' or 'approaching to the mean value'. The concept of regression analysis was first explained by British biometrician **Sir Francis Galton** in 1877 in paper ***"Regression Towards Mediocrity in Hereditary Stature."*** He studied the relationship between the heights of about one thousand fathers and sons. As per his analysis the tall fathers have tall sons and short fathers have short sons. In addition the mean height of the sons of a group of tall fathers is less than that of the fathers and the mean height of sons of a group of short fathers is more than that of the fathers. Thus, Galton's study revealed that the sons of abnormally tall or short fathers tend to revert or move back to the average height of the population, which Galton described as *Regression to Mediocrity.* Thus, the determination of an appropriate functional relationship between the variables is termed as regression analysis.

Today, regression analysis is one of the most important and widely used statistical tool in almost all natural, social and physical sciences. It is specially used in business and economics to study the relationship between two or more variables that are related causally. ***Regression is a statistical technique with the help of which we study the dependence of one variable on another variable with a view to estimate or predict the value of dependent variable given the value of independent variable.*** For example, we know there is a positive relationship between saving and disposable income of a community. If the level of disposable income is known, the method of regression can be used to estimate the savings that corresponds to given level of disposable income.

According to **M.M. Blair,** ***"Regression is the measure of the average relationship between two or more variables in terms of the original units of data."***

In the words of **Taro Yamane,** ***"One of the most frequently used technique in economics and business research to find a relation between two or more variables that are related causally, is regression analysis."***

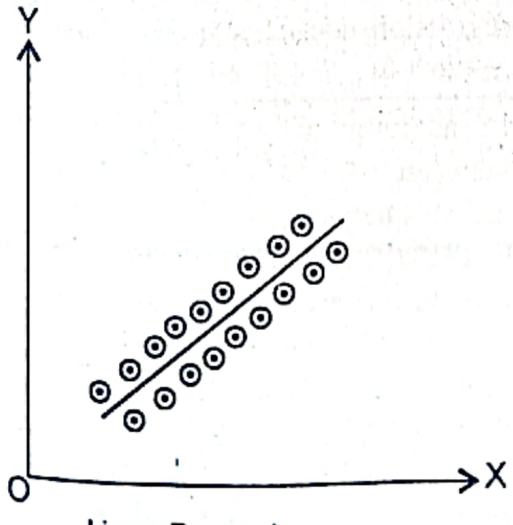
According to **Y.L. Chou,** ***"Regression analysis attempts to establish the nature of the relationship between the variables and thereby provide a mechanism for prediction or forecasting."***

10.3 TYPES OF REGRESSION ANALYSIS

(*i*) Simple and Multiple Regression : When there are only two variables under consideration it is known as a *simple regression.* In simple regression analysis one is dependent variable and other is independent variable. The simple regression function is written as $Y = f(X)$, where Y is dependent and X is independent variable. On the contrary, the study of more than two variables at a time is known as *multiple regression.* Under this, only one variable is taken as a dependent variable and others as independent variables. If we take Z as a function of X and Y we can symbolically write the regression function as $Z = f(X, Y)$. It is a functional relationship between Z and X, Y.

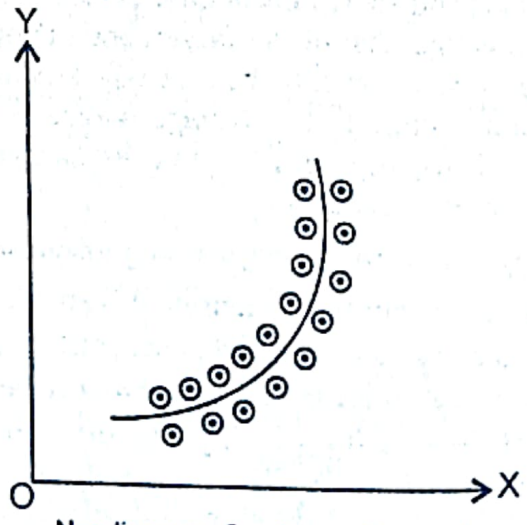
(*ii*) Total and Partial Regression : While considering the *total regression*, all the important variables are taken into consideration at a time. In the present set up most of economic and business phenomenon are affected by multiple factors whose effects have to be studied collectively. In the case of *partial regression* one or two variables are taken into consideration and the others are excluded. Suppose K is a function of X, Y and Z then the total regression is expressed as $K = f(X, Y, Z)$ while the partial regression can be written as $K = f$(X but not Y and Z).

(*iii*) Linear and Non-linear Regression : When the functional relationship between X and Y is expressed as the first degree equation, it is known as *linear regression.* In other words, when the points plotted on a scatter diagram concentrate around a straight line it is the case of linear regression. On the other hand if the line of regression (in scatter diagram) is not a straight line, the regression is termed as *curved* or *non-linear regression.* The regression equations of non-linear regression are represented by equations of higher degree. The following diagrams show the linear and non-linear regressions (See fig. 10.1 and Fig. 10.2).



Linear Regression

Fig. 10.1



Non-linear or Curvilinear Regression

Fig. 10.2

10.4 COMPARISON BETWEEN CORRELATION AND REGRESSION

Although correlation and regression are both concerned with the study of relationship of two or more variables but some basic differences in these two concepts are as under :

Correlation Analysis	Regression Analysis
(*i*) Correlation analysis studies the co-variability between two variables. It tells whether the two variable move in the same or in the opposite directions.	(*ii*) Regression analysis expresses the average relationship between two or more variables.
(*ii*) Correlation analysis does not establish cause and effect relationship between the variables.	(*ii*) Regression analysis is based on cause and effect relationship between the variables. The variable expressing cause is taken as independent variable and that expressing effect is taken as dependent variable.
(*iii*) If there is a mathematical or statistical relation between the variables giving the value of r but logically the relationship does not exist, the correlation is known as *non-sense* or *spurious correlation.*	(*iii*) There is no illusion in regression analysis regarding the relationship between X and Y. For every value of X we get some value of Y and vice-versa.
(*iv*) Correlation coefficient between the variables X and Y or Y and X is always the same. Any one of them can be taken as dependent and other as independent variable in both ways *i.e.* $r_{XY} = r_{YX}$	(*iv*) Regression equations in general deal with the functional relationships between Y and X or X and Y. In these two relations the dependent and independent variables change and as such their regression coefficients also change *i.e.* $b_{XY} \neq b_{YX}$. (See section 10.6.1)
(*v*) Coefficient of correlation r is a relative measure. The linear relationship between the variables X and Y is free from the units of measurement. Its value lies between – 1 and + 1.	(*v*) The regression coefficients are absolute measures representing the changes in the variables. The values of the dependent variables corresponding to independent variables are expressed in the units of measurement already assigned. Value of regression coefficients may not lie in range -1 to $+1$.
(*vi*) Correlation analysis has its application in a limited fields and is confined to the study of relationship of the variables only.	(*vi*) In addition of establishing a relationship between two or more variables, regression analysis helps in forecasting etc.

10.5 OBJECTIVES OF REGRESSION ANALYSIS

Regression analysis is used to achieve the following objectives :

(*i*) **Forecasting or Prediction :** The most important use of the regression analysis is to predict or estimate the value of dependent variable corresponding to any value of independent variable.

(*ii*) **Study of Cause and Effect Relationship :** Since most of our problems are the result of cause and effect relationship, therefore, the regression analysis suits the most.

(iii) **Helpful in Planning Decisions and Policy Making :** It is of an immense use in planning, policy making and other fields of social and economic nature. Once a functional relationship is established, the values of dependent variables can be predicted from the given values of independent variables.

(iv) **Testing of Economic Theory :** Regression analysis is very useful in testing the validity of economic theories.

(v) **Useful in Economics and Commerce :** With the help of regression analysis, businessmen can take necessary actions in regards to future production, sale, investment, advertisement etc.

10.6 REGRESSION LINES

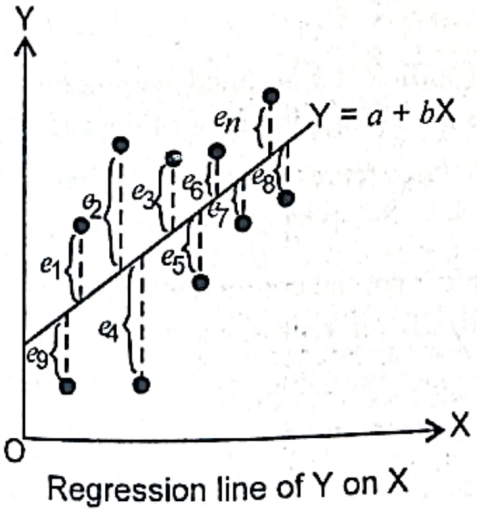
A *regression line* is the line from which we can get the best estimated value of the dependent variable corresponding to a given value of independent variable. For two correlated variables X and Y we have two regression lines. When we treat X as the independent variable and Y as the dependent variable we get the regression line of Y on X and when we treat Y as the independent variable and X as the dependent variable we get the regression line of X on Y.

The *regression line of Y on X* is that line from which we get the best estimated value of Y corresponding to a given value of X. Similarly, the *regression line of X on Y* is that line from which we get the best estimated value of X corresponding to a given value of Y. If there exists either perfect positive or perfect negative correlation between two variables X and Y then the two lines of regression coincide.

A regression line is determined by the *method of least squares*. The method of least squares states that a regression line is fitted to different points of a scatter diagram in such a way that the sum of squares of the deviations of the given observed values from the fitted line should be least or minimum. A line fitted by this method is called the *line of best fit*. This line is also called *the least squares line*.

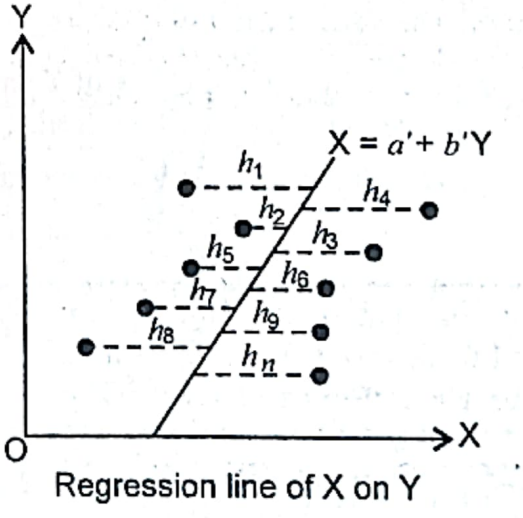
The term *best fit* is interpreted in accordance with the principle of least squares which consists of minimising the sum of squares of the residuals or the errors of estimates *i.e.* the deviations between the given observed values and given by the line of best fit.

The regression lines of Y on X and X on Y are shown with the help of following diagrams :



Regression line of Y on X

Fig. 10.3



Regression line of X on Y

Fig. 10.4

The deviations of the points from the line of best fit can be measured in two ways *i.e.* either parallel to y-axis (in case of Y on X) or parallel to x-axis (in case of X on Y) as shown in above figures 10.3 and 10.4 respectively.

Regression Equations

The mathematical equation of a regression line is called a *regression equation*. In practice, two terms '*regression line*' and '*regression equation*' are used interchangeably. The regression equation of Y on X is expressed in the form $Y = a + bX$ where a and b are called the parameters.

The regression equation X on Y is expressed in the form $X = a' + b'Y$ where a' and b' are called parameters.

However, in extreme case when two variables X and Y are perfectly correlated *i.e.* $r = \pm 1$, two lines of regression coincide each other and we have only one line of regression.

Reason for Two Lines of Regression

When there are two interdependent variables X and Y *i.e.* Y depends on X and X depends on Y *e.g.* in case of price and demand ; sometimes demand depends on price and there are some cases when price depends on demand. In such situations, there must be two regression lines commonly known as Y on X and X on Y.

Regression line of Y on X gives the most probable mean values of Y for given values of X. On the contrary, the regression line of X on Y gives the most probable mean value of X for given values of Y. The former (*i.e.* Y on X) is used to estimate Y (dependent variable) for given value of X (independent variable). Similarly, the line X on Y is used to estimate X (dependent variable) for given value of Y (independent variable). In the least squares method, for estimating the value of Y given the value of X, we have to minimise sum of squares of deviations along Y-axis of the given values from the corresponding values given by the regression line. Similarly, for estimating the value of X given the value of Y, we minimise sum of squares of deviations along x-axis of the given values from the corresponding values given by the regression line. It is not possible for any single line or equation to minimise these two simultaneously.

In other words, the two regression equations are not interchangable because of the reason that the basis and assumptions for the derivation of these equations are quite different as discussed above. Therefore, we have to take two regression lines.

Regression Coefficients

If $Y = a + bX$ is the regression line of Y on X then the coefficient b is called *regression coefficient of Y on X* and is denoted by b_{YX}. Mathematically, it represents the change in value of dependent variable Y corresponding to a unit change in independent variable X *i.e.* b_{YX} represents the rate of change of Y with respect to X.

Similarly, if $X = a' + b'Y$ is the regression line of X on Y then, the coefficient b' is called *regression coefficient of X on Y* and is denoted by b_{XY}. Mathematically, it represents the change in value of dependent variable X corresponding to a unit change in independent variable Y *i.e.* b_{XY} represents the rate of change of X w.r.t. Y.

10.6.1 DETERMINATION OF REGRESSION LINES USING THE METHOD OF LEAST SQUARES

Procedure

Suppose we have n pairs of points $(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)$ corresponding to two variables X and Y under observation.

Let the line of best fit of Y on X be

$$Y = a + b\,X \qquad \ldots(10.1)$$

where a and b are unknowns. For any $P(x_i, y_i)$ in the scatter diagram (see Fig. 10.5), the error of estimate is shown by PQ.

The coordinates of point Q are $(x_i, a + b\,x_i)$

$\therefore$ error of estimate of P is given by

$$PQ = PM - QM$$

i.e. $$e_i = y_i - (a + b\,x_i)$$

i.e. $$e_i = y_i - a - b\,x_i \qquad \ldots(10.2)$$

Clearly the error e_i will be positive or negative according as P lies above or below the regression line.

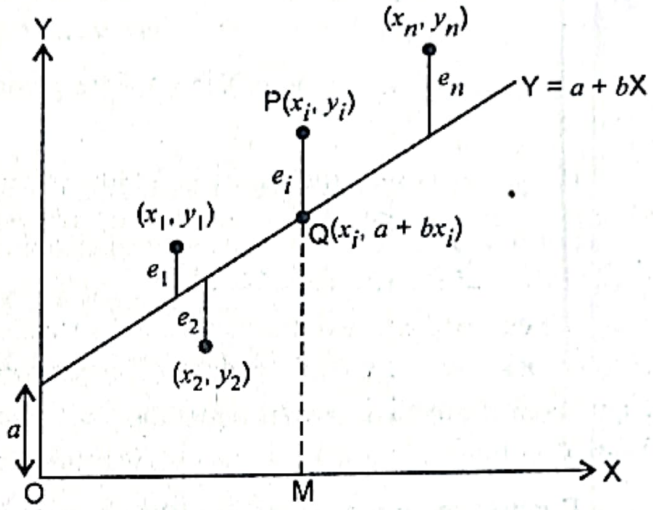


Fig. 10.5

Now our aim is to find the values of the unknowns a and b in equation (10.1) such that sum of squares of errors is least. *i.e.* we have to minimise $\sum_{i=1}^{n} e_i^2$

Let $$S = \sum_{i=1}^{n} e_i^2$$

i.e. $$S = \sum_{i=1}^{n} (y_i - a - b\,x_i)^2$$

Clearly S will be minimum if $\frac{\partial S}{\partial a} = 0$ and $\frac{\partial S}{\partial b} = 0$.

Now $\frac{\partial S}{\partial a} = 0 \Rightarrow \sum_{i=1}^{n} 2(y_i - a - b\,x_i)(-1) = 0$

$\Rightarrow$ $$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} a + \sum_{i=1}^{n} b\,x_i$$

$\Rightarrow$ $$\sum_{i=1}^{n} y_i = n\,a + b\sum_{i=1}^{n} x_i \qquad \ldots(10.3)$$

and $\frac{\partial S}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} 2(y_i - a - b\,x_i)(-x_i) = 0$

$$\Rightarrow \quad \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} a x_i + \sum_{i=1}^{n} b x_i^2$$

$$\Rightarrow \quad \sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 \qquad \ldots(10.4)$$

Equations (10.3) and (10.4) are called *normal equations.*

For notational convenience, we can also write these normal equations as

$$\Sigma Y = n a + b \Sigma X \qquad \ldots(10.5)$$

and $$\Sigma XY = a \Sigma X + b \Sigma X^2 \qquad \ldots(10.6)$$

From these two normal equations, we can obtain the values a and b hence the required equation of regression line by simultaneously solving these two normal equations.

For this, rewriting equation (10.5) as

$$a = \frac{\Sigma Y}{n} - \frac{b \Sigma X}{n} \qquad \ldots(10.7)$$

Substituting the value of a in equation (10.6), we get

$$\Sigma XY = \left(\frac{\Sigma Y}{n} - \frac{b \Sigma X}{n} \right) \Sigma X + b \Sigma X^2$$

$$\Rightarrow \quad \Sigma XY = \frac{\Sigma X \Sigma Y}{n} - \frac{b (\Sigma X)^2}{n} + b \Sigma X^2$$

$$\Rightarrow \quad \Sigma XY - \frac{\Sigma X \Sigma Y}{n} = b \left(\Sigma X^2 - \frac{(\Sigma X)^2}{n} \right)$$

$$\Rightarrow \quad \frac{n \Sigma XY - \Sigma X \Sigma Y}{n} = \frac{b [n \Sigma X^2 - (\Sigma X)^2]}{n}$$

$$\Rightarrow \quad b = \frac{n \Sigma XY - \Sigma X \Sigma Y}{n \Sigma X^2 - (\Sigma X)^2} \qquad \ldots(10.8)$$

Hence the regression line of Y on X is given by

$$Y = a + b X$$

where $$b = \frac{n \Sigma XY - \Sigma X \Sigma Y}{n \Sigma X^2 - (\Sigma X)^2} \quad \text{and} \quad a = \frac{\Sigma Y}{n} - \frac{b \Sigma X}{n}$$

Notes (*i*) Since the regression coefficient of Y on X is denoted by b_{YX}.

So $$b_{YX} = \frac{n \Sigma XY - \Sigma X \Sigma Y}{n \Sigma X^2 - (\Sigma X)^2} \qquad \ldots(10.9)$$

(*ii*) Since $\Sigma(X-\bar{X})(Y-\bar{Y}) = \Sigma(XY - X\bar{Y} - \bar{X}Y + \bar{X}\bar{Y})$

$= \Sigma XY - \bar{Y}\Sigma X - \bar{X}\Sigma Y + n\bar{X}\bar{Y}$

$= \Sigma XY - \bar{Y}(n\bar{X}) - \bar{X}(n\bar{Y}) + n\bar{X}\bar{Y}$

$[\because \Sigma X = n\bar{X} \text{ and } \Sigma Y = n\bar{Y}]$

$= \Sigma XY - n\bar{X}\bar{Y} = \Sigma XY - n\frac{\Sigma X}{n}\frac{\Sigma Y}{n}$

i.e. $\Sigma(X-\bar{X})(Y-\bar{Y}) = \Sigma XY - \frac{\Sigma X \Sigma Y}{n}$

or $n\Sigma(X-\bar{X})(Y-\bar{Y}) = n\Sigma XY - \Sigma X \Sigma Y$ ...(10.10)

and $\Sigma(X-\bar{X})^2 = \Sigma(X^2 + \bar{X}^2 - 2X\bar{X})$

$= \Sigma X^2 + n\bar{X}^2 - 2\bar{X}\Sigma X$

$= \Sigma X^2 + n\bar{X}^2 - 2\bar{X}(n\bar{X})$ $[\because \Sigma X = n\bar{X}]$

$= \Sigma X^2 - n\bar{X}^2 = \Sigma X^2 - n\left(\frac{\Sigma X}{n}\right)^2$

i.e. $n\Sigma(X-\bar{X})^2 = n\Sigma X^2 - (\Sigma X)^2$ ...(10.11)

Using equations (10.10) and (10.11) in equation (10.9), we get,

$$b_{YX} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(X-\bar{X})^2} \quad \text{...(10.12)}$$

$$= \frac{\frac{1}{n}\Sigma(X-\bar{X})(Y-\bar{Y})}{\frac{1}{n}\Sigma(X-\bar{X})^2}$$

$\Rightarrow$ $$b_{YX} = \frac{\text{Cov}(X, Y)}{\sigma_X^2} \quad \text{...(10.13)}$$

Further $$b_{YX} = \frac{\text{Cov}(X, Y)}{\sigma_X^2} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \cdot \frac{\sigma_Y}{\sigma_X}$$

Since $$r = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

So $$b_{YX} = r\frac{\sigma_Y}{\sigma_X} \quad \text{...(10.14)}$$

Hence summarizing all the formulae for b_{YX}, we have

$$b \text{ or } b_{YX} = \frac{n\Sigma XY - \Sigma X\,\Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(X-\bar{X})^2} = \frac{\text{Cov}(X, Y)}{\sigma_X{}^2} = r\frac{\sigma_Y}{\sigma_X}$$

(*iii*) To find the regression line of X on Y, we assume that the regression line of X on Y is

$$X = a' + b'Y \qquad \ldots(10.15)$$

Continuing in the same way as above, we can obtain the two normal equations as

$$\Sigma X = n\,a' + b'\,\Sigma Y \qquad \ldots(10.16)$$

and

$$\Sigma XY = a'\,\Sigma Y + b'\,\Sigma Y^2 \qquad \ldots(10.17)$$

which can be solved for a' and b'.

Also the regression coefficient of X on Y is given by

$$b' \text{ or } b_{XY} = \frac{n\Sigma XY - \Sigma X\,\Sigma Y}{n\Sigma Y^2 - (\Sigma Y)^2} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(Y-\bar{Y})^2} = \frac{\text{Cov}(X, Y)}{\sigma_Y{}^2} = r\frac{\sigma_X}{\sigma_Y}$$

CHECKPOINTS

1. Write a note on regression analysis. (G.N.D.U. B.Sc. C.Sc. April 2004)
2. Define regression analysis. What is the difference between correlation analysis and regression analysis ? What is the relationship between the two ?
   (G.N.D.U. B.C.A. April 2002, Sept. 2007, 2008, P.U. B.C.A. Sept. 2006)
3. Explain the concept of linear regression. (G.N.D.U. B.C.A. April 2008)
4. Write a note on nonlinear regression. (P.U. B.C.A. April 2005, Sept. 2006)
5. Distinguish between
   (*i*) Simple and multiple regression (G.N.D.U. B.Sc. C.Sc. Sept. 2007)
   (*ii*) Linear and Nonlinear regression
   (G.N.D.U. B.Sc. C.Sc. Sept. 2007; P.U. B.C.A. Sept. 2007, 2008)
6. Why we have, in general, two lines of regression ?
   (P.U. B.C.A. Sept. 2004, 2006; G.N.D.U. B.C.A. Sept. 2008)
7. Discuss the least square method of fitting regression. (G.N.D.U. B.C.A. April 2002)
8. Explain the method of least squares to fit a straight line which is a best fit.
   (P.U. B.C.A. April 2007, Sept. 2007, 2008; G.N.D.U. B.Sc. I.T. 2009)
9. Write an algorithm to fit a regression line of Y on X for a set of n paired observations (X_i, Y_i); $i = 1, 2, 3, \ldots\ldots n$. (P.U. B.C.A. April 2004)

ILLUSTRATIVE EXAMPLES

Example 1. By the method of least squares, find a straight line that best fits the following data :

X :	1	2	3	4	5
Y :	14	27	40	55	68

(P.U. B.C.A. April 2008)

Sol. Let the straight line which is best fit to the given data be

$$Y = a + b\,X \qquad \ldots(i)$$

where a and b are unknowns.

By the method of least squares, the values of a and b are obtained by solving the following normal equations :

$$\Sigma Y = n\,a + b\,\Sigma X \qquad \ldots(ii)$$

and

$$\Sigma XY = a\,\Sigma X + b\,\Sigma X^2 \qquad \ldots(iii)$$

Now we prepare the following table :

X	Y	X^2	XY
1	14	1	14
2	27	4	54
3	40	9	120
4	55	16	220
5	68	25	340
$\Sigma X = 15$	$\Sigma Y = 204$	$\Sigma X^2 = 55$	$\Sigma XY = 748$

Substituting these values in equations (*ii*) and (*iii*), we get

$$5\,a + 15\,b = 204$$

and

$$15\,a + 55\,b = 748$$

On solving these two equations for a and b, we get , $a = 0$ and $b = 13.6$

$\therefore$ from (*i*), the straight line which is best fit for the data is given by

$$Y = 0 + 13.6\,X$$

i.e.

$$Y = 13.6\,X$$

Example 2. Calculate the regression equation of X on Y and Y on X from the following data :

X :	10	12	13	17	18
Y :	5	6	7	9	13

Sol. Let the regression equation of Y on X be $Y = a + bX$ ...(*i*)

and the regression equation of X on Y be $X = a' + b'Y$ ...(*ii*)

where a, b, a' and b' are unknowns.

By method of least squares, the values of a and b are calculated by using the normal equations

$$\Sigma Y = n\,a + b\,\Sigma X \quad \ldots(iii)$$

and $$\Sigma XY = a\,\Sigma X + b\,\Sigma X^2 \quad \ldots(iv)$$

and the values of a' and b' are calculated by using the normal equations

$$\Sigma X = n\,a' + b'\,\Sigma Y \quad \ldots(v)$$

and $$\Sigma XY = a'\,\Sigma Y + b'\,\Sigma Y^2 \quad \ldots(vi)$$

Now we construct the following table :

X	Y	X^2	Y^2	XY
10	5	100	25	50
12	6	144	36	72
13	7	169	49	91
17	9	289	81	153
18	13	324	169	234
$\Sigma X = 70$	$\Sigma Y = 40$	$\Sigma X^2 = 1026$	$\Sigma Y^2 = 360$	$\Sigma XY = 600$

Substituting these values from the table in equations (*iii*) and (*iv*), we get

$$5\,a + 70\,b = 40$$

and $$70\,a + 1026\,b = 600$$

On solving these two equations for a and b, we get , $a = -4.18$ and $b = 0.87$

$\therefore$ from (*i*), the regression equation of Y on X is given by

$$Y = -4.18 + 0.87X$$

Now substituting the values from the table in equations (*v*) and (*vi*), we get

$$5\,a' + 40\,b' = 70$$

and $$40\,a' + 360\,b' = 600$$

On solving these two equations for a' and b', we get $a' = 6$ and $b' = 1$

$\therefore$ from (*ii*), the regression equation of X on Y is given by $X = 6 + Y$

EXERCISE 10.1

1. The following table gives the age of cars of a certain make and annual maintenance costs. Obtain the regression equation for costs related to age.

Age of Cars (in years) :	2	4	6	8
Maintenance Cost (in hundred of Rs.) :	10	20	25	30

2. The result of measurement of electric resistance R of a copper bar at various temperatures t(°C) is given as :

t :	19	25	30	36	40	45	50
R :	76	77	79	80	82	83	85

Find a relation $R = a + b\,t$ where a and b are constants to be determined using method of least squares.

(P.U. B.C.A. April 2002)

3. From the following data estimate the regression equation of Y on X. Also estimate the value of Y when X = 30.

X :	25	22	28	26	35	20	22	40	20	18	19	25
Y :	18	15	20	17	22	14	15	21	15	14	16	17

4. Determine two regression equations by the method of least squares from the following data :

X :	5	8	7	6	4
Y :	3	4	5	2	1

5. From the following data, find the two regression equations :

X :	2	4	6	8	10
Y :	12	6	14	12	16

ANSWERS

1. $Y = 5 + 3.25\,X$ **2.** $R = 70.14 + 0.29\,t$ **3.** $Y = 7.625 + 0.375\,X$, 18.875
4. $Y = -1.8 + 0.8\,X$, $X = 3.6 + 0.8\,Y$ **5.** $Y = 12.8 + 0.2\,X$, $X = -1 + 0.5\,Y$

10.6.2 AN ALTERNATIVE WAY TO OBTAIN TWO REGRESSION LINES

The calculations of normal equations discussed above become difficult and complex when values of X and Y are large. The work can be simplified by an alternative approach which is discussed as below :

The regression equation of Y on X is

$$Y = a + b\,X \qquad \ldots(10.18)$$

Taking summation over the n values and dividing both sides of equation (10.18) by n, we have

$$\overline{Y} = a + b\,\overline{X} \qquad \ldots(10.19)$$

Now, subtracting equation (10.19) from (10.18), we have

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

or

$$Y - \overline{Y} = b_{YX}(X - \overline{X}) \qquad \ldots(10.20)$$

as $b = b_{YX}$. This is the regression equation of Y on X.

Similarly, the regression equation of X on Y can be written as

$$X - \overline{X} = b_{XY}(Y - \overline{Y}), \qquad \ldots(10.21)$$

where b_{XY} is regression coefficient of X on Y.

The regression coefficients b_{YX} and b_{XY} can be calculated by following methods :

(i) Direct Method

(i) If the arithmetic means of two series are integers *i.e.* $\overline{X}$ and $\overline{Y}$ are integers then we shall use the formula (10.12)

i.e. $$b_{YX} = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{\Sigma (X - \bar{X})^2}$$

or $$b_{YX} = \frac{\Sigma xy}{\Sigma x^2}$$

where $x = X - \bar{X}$ and $y = Y - \bar{Y}$

Similarly $$b_{XY} = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{\Sigma (Y - \bar{Y})^2}$$

or $$b_{YX} = \frac{\Sigma xy}{\Sigma y^2}$$

where $x = X - \bar{X}$ and $y = Y - \bar{Y}$

(*ii*) If the arithmetic means of two series are in fractions, then we shall use the formula (10.9)

i.e. $$b_{YX} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}$$

Similarly $$b_{XY} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma Y^2 - (\Sigma Y)^2}$$

(*ii*) Short Cut Method

If $\bar{X}$ and $\bar{Y}$ are in fractions and X and Y assume large values then we can change the origin *i.e.* we can take the deviations of the variables X and Y from assumed means A and B as the regression coefficients are independent of change of origin (see Sec. 10.6.3 Property V).

i.e. $$b_{YX} = \frac{n\Sigma dx\, dy - \Sigma dx \Sigma dy}{n\Sigma dx^2 - (\Sigma dx)^2}$$

$dx = X - A$ and $dy = Y - B$

Similarly $$b_{XY} = \frac{n\Sigma dx\, dy - \Sigma dx \Sigma dy}{n\Sigma dy^2 - (\Sigma dy)^2}$$

(*iii*) Step Deviation Method

This method is applied when some common factor can be taken from dx or dy or both.

In other words if we change both the origin and scale and transform the variables X and Y to new variables

$$U = \frac{X - A}{h} \quad \text{and} \quad V = \frac{Y - B}{k}$$

then $$b_{YX} = \frac{k}{h} b_{VU}$$ (See Sec. 10.6.3 Property V)

i.e. $$b_{YX} = \frac{n\Sigma d'x\, d'y - \Sigma d'x \Sigma d'y}{n\Sigma d'x^2 - (\Sigma d'x)^2} \times \frac{k}{h}$$

where $$d'x = \frac{dx}{h} \text{ and } d'y = \frac{dy}{k}$$

Similarly $$b_{XY} = \frac{n\Sigma d'x\, d'y - \Sigma d'x \Sigma d'y}{n\Sigma d'y^2 - (\Sigma d'y)^2} \times \frac{h}{k}$$

where $$d'x = \frac{dx}{h} \text{ and } d'y = \frac{dy}{k}$$

ILLUSTRATIVE EXAMPLES

Example 1. Obtain the regression line equation of X on Y and Y on X for given data :

$\overline{X} = 25$, $\overline{Y} = 40$, $r = 0.8$, $\sigma_X = 3$ and $\sigma_Y = 6$. (G.N.D.U. B.Sc. C.Sc. Sept. 2007)

Sol. Given $\overline{X} = 25$, $\overline{Y} = 40$, $r = 0.8$, $\sigma_X = 3$ and $\sigma_Y = 6$

Now regression coefficient of Y on X is given by

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X} = 0.8 \times \frac{6}{3} = 1.6$$

and regression coefficient of X on Y is given by

$$b_{XY} = r\frac{\sigma_X}{\sigma_Y} = 0.8 \times \frac{3}{6} = 0.4$$

$\therefore$ regression line equation of Y on X is given by

$$Y - \overline{Y} = b_{YX}(X - \overline{X})$$

i.e. $$Y - 40 = 1.6(X - 25)$$

i.e. $$Y = 1.6X$$

and regression line equation of X on Y is given by

$$X - \overline{X} = b_{XY}(Y - \overline{Y})$$

i.e. $$X - 25 = 0.4(Y - 40)$$

i.e. $$X = 0.4Y + 9$$

which are the required equations.

Example 2. You are given the following data :

Series	X	Y
Mean	18	100
S.D.	14	20

Coefficient of correlation between X and Y is 0.8. Find

(*i*) two regression coefficients (*ii*) two lines of regression

(*iii*) the estimate for value of Y when X = 70 (*iv*) the estimate for value of X when Y = 90.

(P.U. B.C.A. Sept. 2002)

Sol. Given $\overline{X} = 18$, $\overline{Y} = 100$, $\sigma_X = 14$, $\sigma_Y = 20$ and $r = 0.8$.

(*i*) The regression coefficient of Y on X is given by

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X} = 0.8 \times \frac{20}{14} = 1.14$$

and the regression coefficient of X on Y is given by

$$b_{XY} = r\frac{\sigma_X}{\sigma_Y} = 0.8 \times \frac{14}{20} = 0.56$$

(*ii*) The regression line of Y on X is given by

$$Y - \overline{Y} = b_{YX}(X - \overline{X})$$

i.e. $$Y - 100 = 1.14\,(X - 18)$$

i.e. $$Y = 1.14\,X + 79.48 \quad \ldots(1)$$

and regression line of X on Y is given by

$$X - \overline{X} = b_{XY}(Y - \overline{Y})$$

i.e. $$X - 18 = 0.56\,(Y - 100)$$

i.e. $$X = 0.56\,Y - 38 \quad \ldots(2)$$

(*iii*) When X = 70 then from equation (1), we have

$$Y = 1.14\,(70) + 79.48 = 159.28$$

(*iv*) When Y = 90 then from equation (2), we have

$$X = 0.56\,(90) - 38 = 12.4$$

Example 3. The following calculations have been made for prices of twelve stocks (X) on the Calcutta Stock Exchange in a certain day along with the volume of sales of shares (Y). From these calculations, find the regression equation of prices of stocks on the volume of sales of shares and volume of sales of shares on prices of stocks. Also find correlation coefficient.

$$\Sigma X = 580,\ \Sigma Y = 370,\ \Sigma XY = 11494,\ \Sigma X^2 = 41658,\ \Sigma Y^2 = 17206$$

Sol. Given $\Sigma X = 580, \Sigma Y = 370, \Sigma XY = 11494, \Sigma X^2 = 41658, \Sigma Y^2 = 17206$

$$\overline{X} = \frac{\Sigma X}{n} = \frac{580}{12} = 48.33, \quad \overline{Y} = \frac{\Sigma Y}{n} = \frac{370}{12} = 30.83$$

Regression Coefficient of Y on X is given by

$$b_{YX} = \frac{n\,\Sigma XY - (\Sigma X)(\Sigma Y)}{n\,\Sigma X^2 - (\Sigma X)^2} = \frac{12 \times 11494 - (580)(370)}{12 \times 41658 - (580)^2} = -\frac{76672}{163496} = -0.47$$

$\therefore$ Regression equation of Y on X is given by

$$Y - \overline{Y} = b_{YX}(X - \overline{X})$$

$$Y - 30.83 = -0.47\,(X - 48.33)$$

i.e. $$Y = -0.47\,X + 53.54 \qquad \ldots(i)$$

i.e.

Regression Coefficient of X on Y is given by

$$b_{XY} = \frac{n\,\Sigma XY - (\Sigma X)(\Sigma Y)}{n\,\Sigma Y^2 - (\Sigma Y)^2} = \frac{12 \times 11494 - (580)(370)}{12 \times 17206 - (370)^2} = -\frac{76672}{69572} = -1.102$$

$\therefore$ Regression equation of X on Y is given by

$$X - \overline{X} = b_{XY}(Y - \overline{Y})$$

i.e. $$X - 48.33 = -1.102\,(Y - 30.83)$$

i.e. $$X = -1.102\,Y + 82.305 \qquad \ldots(ii)$$

Further, $$r^2 = b_{YX}\,b_{XY} = (-1.102)(-0.47) = 0.51794$$

$\therefore$ $$r = -\sqrt{0.51794} = -0.72$$

[Taking r as – ve because b_{YX} and b_{XY} are both – ve]

ample 4. Find the regression lines of Y on X and X on Y from the following data :

X :	1	2	3
Y :	2	4	5

l. First, we prepare the following table :

X	Y	X^2	Y^2	XY
1	2	1	4	2
2	4	4	16	8
3	5	9	25	15
$\Sigma X = 6$	$\Sigma Y = 11$	$\Sigma X^2 = 14$	$\Sigma Y^2 = 45$	$\Sigma XY = 25$

Now, $$\overline{X} = \frac{\Sigma X}{n} = \frac{6}{3} = 2 \quad \text{and} \quad \overline{Y} = \frac{\Sigma Y}{n} = \frac{11}{3} = 3.67$$

Regression coefficient of Y on X is given by

$$b_{YX} = \frac{n.\Sigma XY - \Sigma X.\Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} = \frac{3(25) - (6)(11)}{3(14) - (6)^2} = \frac{75 - 66}{42 - 36} = \frac{9}{6} = 1.5$$

$\therefore$ Regression equation of Y on X is given by

$$Y - \overline{Y} = b_{YX}(X - \overline{X})$$

i.e. $$Y - 3.67 = 1.5\,(X - 2)$$

i.e. $Y - 3.67 = 1.5\,X - 3$

$$Y = 0.67 + 1.5X$$

Regression coefficient of X on Y is given by

$$b_{XY} = \frac{n\Sigma XY - \Sigma X.\Sigma Y}{n\Sigma Y^2 - (\Sigma Y)^2} = \frac{3(25) - (6)(11)}{3(45) - (11)^2} = \frac{75 - 66}{135 - 121} = \frac{9}{14} = 0.643$$

$\therefore$ Regression equation of X on Y is given by

$$X - \overline{X} = b_{XY}(Y - \overline{Y})$$

$$X - 2 = 0.643\,(Y - 3.67)$$

$$X - 2 = 0.643\,Y - 2.36$$

$$X = -0.36 + 0.643Y$$

Example 5. Find the estimation of two lines of regression for the data :

X :	1	2	3	4	5
Y :	7	6	5	4	3

and hence find an estimate of Y for X = 3.5 from the approximate line of regression.

(G.N.D.U. B.Sc. C.Sc. April 2007)

Sol. First we shall prepare the following table :

X	Y	$\overline{X} = 3$ $x = X - \overline{X}$	$\overline{Y} = 5$ $y = Y - \overline{Y}$	x^2	y^2	xy
1	7	−2	2	4	4	−4
2	6	−1	1	1	1	−1
3	5	0	0	0	0	0
4	4	1	−1	1	1	−1
5	3	2	−2	4	4	−4
$\Sigma X = 15$	$\Sigma Y = 25$			$\Sigma x^2 = 10$	$\Sigma y^2 = 10$	$\Sigma xy = -10$

$$\overline{X} = \frac{\Sigma X}{n} = \frac{15}{5} = 3 \quad \text{and} \quad \overline{Y} = \frac{\Sigma Y}{n} = \frac{25}{5} = 5$$

Now regression coefficient of Y on X is given by

$$b_{YX} = \frac{\Sigma xy}{\Sigma x^2} = \frac{-10}{10} = -1$$

and regression coefficient of X on Y is given by

$$b_{XY} = \frac{\Sigma xy}{\Sigma y^2} = \frac{-10}{10} = -1$$

$\therefore$ regression equation of Y on X is given by

$$Y - \overline{Y} = b_{YX}\,(X - \overline{X})$$

i.e. $$Y - 5 = -1(X - 3)$$

i.e. $$Y = -X + 8 \quad \ldots(i)$$

and regression equation of X on Y is given by

$$X - \overline{X} = b_{XY}\,(Y - \overline{Y})$$

i.e. $$X - 3 = -1(Y - 5) \quad \textit{i.e.} \quad X = -Y + 8 \quad \ldots(ii)$$

Further, when X = 3.5 then from (*i*), we have

$$Y = -3.5 + 8 = 4.5$$

Example 6. Calculate regression coefficients for the following data :

X:	1	2	3	4	5	6	7	8
Y:	3	7	10	12	14	17	20	24

(P.U. B.C.A. April 2008)

Sol. First, we prepare the following table :

X	Y	A = 4 $dx = X - A$	B = 14 $dy = Y - B$	dx^2	dy^2	$dx\,dy$
1	3	−3	−11	9	121	33
2	7	−2	−7	4	49	14
3	10	−1	−4	1	16	4
4	12	0	−2	0	4	0
5	14	1	0	1	0	0
6	17	2	3	4	9	6
7	20	3	6	9	36	18
8	24	4	10	16	100	40
		$\Sigma dx = 4$	$\Sigma dy = -5$	$\Sigma dx^2 = 44$	$\Sigma dy^2 = 335$	$\Sigma dx\,dy = 115$

Now, regression coefficient of Y on X is given by

$$b_{YX} = \frac{n\Sigma dx\,dy - \Sigma dx\,\Sigma dy}{n\Sigma dx^2 - (\Sigma dx)^2} = \frac{8(115) - 4(-5)}{8(44) - (4)^2} = \frac{940}{336} = 2.80$$

and regression coefficient of X on Y is given by

$$b_{XY} = \frac{n\Sigma dx\,dy - \Sigma dx\,\Sigma dy}{n\Sigma dy^2 - (\Sigma dy)^2} = \frac{8(115) - 4(-5)}{8(335) - (-5)^2} = \frac{940}{2655} = 0.35$$

Example 7. The observations from an experiment are obtained for Y by varying X as shown in following table :

X :	1.0	1.5	2.0	2.5	3.0	3.5
Y :	6.2	7.5	9.0	10.5	11.5	12.0

Determine the suitable regression line.

(G.N.D.U. B.Sc. I.T. April 2005)

Sol. Since the given observations are obtained for Y by varying X so we shall take X as independent variable and Y as dependent variable. Thus we have to find regression line of Y on X.

First, we prepare the following table :

X	Y	$dx = X - 2.5$	$dy = Y - 10.5$	dx^2	$dx\,dy$
1.0	6.2	– 1.5	– 4.3	2.25	6.45
1.5	7.5	– 1.0	– 3.0	1.00	3.00
2.0	9.0	– 0.5	– 1.5	0.25	0.75
2.5	10.5	0	0	0	0
3.0	11.5	0.5	1.0	0.25	0.5
3.5	12.0	1.0	1.5	1.00	1.5
$\Sigma X = 13.5$	$\Sigma Y = 56.7$	$\Sigma\, dx = -1.5$	$\Sigma\, dy = -6.3$	$\Sigma\, dx^2 = 4.75$	$\Sigma\, dx\, dy = 12.2$

Now $$\overline{X} = \frac{\Sigma X}{n} = \frac{13.5}{6} = 2.25$$

and $$\overline{Y} = \frac{\Sigma Y}{n} = \frac{56.7}{6} = 9.45$$

The regression coefficient of Y on X is given by

$$b_{YX} = \frac{n \Sigma\, dx\, dy - \Sigma\, dx\, \Sigma\, dy}{n \Sigma\, dx^2 - (\Sigma\, dx)^2}$$

$$= \frac{6\,(12.2) - (-1.5)\,(-6.3)}{6\,(4.75) - (-1.5)^2}$$

$$= \frac{73.2 - 9.45}{28.5 - 2.25} = \frac{63.75}{26.25} = 2.43$$

$\therefore$ regression equation of Y on X is given by

$$Y - \overline{Y} = b_{YX}\,(X - \overline{X})$$

$$Y - 9.45 = 2.43\,(X - 2.25)$$

i.e.

$$Y = 2.43\,X + 3.98$$

i.e.

Example 8. Obtain the two regression lines and estimate the blood pressure when the age is 45 years from the following data :

Age (X) :	56	42	72	36	63	47	35	49	38	42	48	60
Blood Pressure (Y):	147	125	160	118	149	128	150	145	155	140	152	155

(G.N.D.U. B.C.A. April 2004; P.U. B.C.A. Sept. 2004)

Sol. First we prepare the following table :

X	Y	$dx = X - 49$	$dy = Y - 144$	dx^2	dy^2	$dx\ dy$
56	147	7	3	49	9	21
42	125	−7	−19	49	361	133
72	160	23	16	529	256	368
36	118	−13	−26	169	676	338
63	149	14	5	196	25	70
47	128	−2	−16	4	256	32
35	150	−14	6	196	36	−84
49	145	0	1	0	1	0
38	155	−11	11	121	121	−121
42	140	−7	−4	49	16	28
48	152	−1	8	1	64	−8
60	155	11	11	121	121	121
$\Sigma X = 588$	$\Sigma Y = 1724$	$\Sigma dx = 0$	$\Sigma dy = -4$	$\Sigma dx^2 = 1484$	$\Sigma dy^2 = 1942$	$\Sigma dx\ dy = 898$

$$\overline{X} = \frac{\Sigma X}{n} = \frac{588}{12} = 49 \quad \text{and} \quad \overline{Y} = \frac{\Sigma Y}{n} = \frac{1724}{12} = 143.67$$

Regression coefficient of Y on X is given by

$$b_{YX} = \frac{n.\Sigma dx\,dy - \Sigma dx \Sigma dy}{n.\Sigma dx^2 - (\Sigma dx)^2} = \frac{12(898) - 0(-4)}{12(1484) - 0} = 0.605$$

and regression coefficient of X on Y is given by

$$b_{XY} = \frac{n.\Sigma dx\,dy - \Sigma dx \Sigma dy}{n.\Sigma dy^2 - (\Sigma dy)^2} = \frac{12(898) - 0}{12(1942) - (-4)^2} = \frac{10776}{23288} = 0.463$$

$\therefore$ Regression equation of Y on X is given by

$$Y - \overline{Y} = b_{YX}\,(X - \overline{X})$$

$$Y - 143.67 = 0.605\,(X - 49)$$

$$Y = 0.605\,X + 114.025 \quad ...(i)$$

and regression equation of X on Y is given by

$$X - \overline{X} = b_{XY}\,(Y - \overline{Y})$$

$$X - 49 = 0.463\,(Y - 143.67)$$

$$X = 0.463\,Y - 17.52 \quad ...(ii)$$

Further, when X = 45, then from (*i*)

$$Y = 0.605\,(45) + 114.025 = 141.25$$

EXERCISE 10.2

1. From the following data, find the most likely value of Y when X = 24.

	X	Y
Mean	18.1	985.8
S.D.	2.0	36.4
	$r = 0.58$	

(P.U. B.C.A. April 2008)

2. You are given below the following information about advertisement and sales

	Advertisement Expenditure (X) (Rs. in crores)	Sales (Y) (Rs. in crores)
Mean	20	120
S.D.	5	25

Correlation co-efficient $r = 0.8$

(*i*) Calculate the two regression equations

(*ii*) Find the likely sales when advertisement expenditure is Rs. 25 crores

(*iii*) What should be the advertisement budget if the company wants to attain sales target of Rs. 150 crores ?

(P.U. B.C.A. Sept. 2005; G.N.D.U. B.C.A. Sept. 2006)

3. Find the most likely price in Bombay corresponding to the price of Rs. 70 at Calcutta from the following :

	Calcutta	Bombay
Average Price	65	67
Standard Deviation	2.5	3.5

Correlation coefficient between the prices of commodities in the two cities is 0.8.

(G.N.D.U. B.C.A. April 2005)

4. You are given the following information about advertising and sales :

	Adv. Expenses (Rs. Lakh)	Sales (Rs. Lakh)
Mean	10	90
S.D.	3	12

Correlation coefficient = 0.8

(*a*) Calculate the two regression lines.

(*b*) Find the likely sales when advertisement expenditure is Rs. 15 lakh.

(*c*) What should be the advertisement expenditure if the company wants to attain sales target of Rs. 120 lakh ?

(G.N.D.U. B.C.A. Sept. 2007)

5. Given the following information find the estimated price in Amritsar of a commodity whose price in Chandigarh is Rs. 57.

	Average Price	Standard Deviation
Chandigarh	52	1.5
Amritsar	56	2.0

Correlation coefficient between the prices of the commodity in the two cities is 0.75.

(G.N.D.U. B.C.A. Sept. 2008)

6. For a group of 500 students the data relating to marks in Statistics and marks in Business Administration are given below :

	Statistics	Business Administration
Mean	72	60
S.D.	16	12

Sum of product of deviation about actual mean = 61440.

Find (*i*) Coefficient of correlation (*ii*) Two regression coefficients

(*iii*) Two regression equations

(*iv*) Estimated marks in Business Administration who obtained 75 marks in Statistics.

(P.U. B.C.A. April 2002)

7. The following data, based on 450 students, are given for marks in Statistics and Economics at a certain examination.

Mean Marks in Statistics (X) = 40

Mean Marks in Economics (Y) = 48

S.D. of Marks in Statistics = 12

Variance of marks in Economics = 256

Sum of the products of deviations of marks from respective means = 42075

Give the equations of the two lines of regression and estimate the average marks in Economics of candidate who obtained 50 marks in Statistics.

8. By using the following data, find out the lines of regression and compute the value of Y when X = 100

$\Sigma X = 250, \Sigma Y = 300, \Sigma XY = 7900, \Sigma X^2 = 6500, \Sigma Y^2 = 10000, N = 10.$

9. For 10 observation on price (P) and supply (S) the following data were obtained (in appropriate units).

$\Sigma P = 130, \Sigma S = 220, \Sigma P^2 = 2288, \Sigma S^2 = 5506, \Sigma PS = 3467, N = 10$

Obtain the line of regression of S on P and estimate the supply when the price is 16 units.

10. Calculate the two Regression Co-efficient from the following data ;

X :	1	2	3	4	5
Y :	2	5	3	8	7

11. Fit a least square line to the data in the following table using X as an independent variable and Y as a dependent variable :

X :	3	5	6	8	9	11
Y :	2	3	4	6	5	8

12. Find the two regression equations from the following data :

X :	1	2	3	4	5
Y :	2	3	4	5	6

If X = 2.5, what will be the value of Y ?

13. Obtain the lines of regression from the following data :

X :	35	25	29	31	27	24	33	36
Y :	23	27	26	21	24	20	29	30

(G.N.D.U. B.C.A. Sept. 2006)

14. The following table gives the data relating to purchases and sales. Obtain the two regression equations by the method of least squares and estimate the likely sales when purchases equal 100 :

Purchase	62	72	98	76	81	56	76	92	88	49
Sales	112	124	131	117	132	96	120	136	97	85

(G.N.D.U. B.C.A. April 2003)

15. The following data relates to the scores obtained by 9 salesmen of a company in an intelligence test and their weekly sales in thousand rupees :

Salesman :	A	B	C	D	E	F	G	H	I
Intelligence Test Score :	50	60	50	60	80	50	80	40	70
Weekly Sales :	30	60	40	50	60	30	70	50	60

(*a*) Obtain the regression equation of sales on intelligence test scores of the salesmen.

(*b*) If the intelligence test score of a salesman is 65, what would be his expected weekly sales ?

(G.N.D.U. B.C.A. April 2007, 2008)

16. Obtain the regression equations for the following :

X :	15	27	27	30	34	38	46
Y :	120	140	150	170	180	200	250

17. Write down two regular equations that may be associated with the following pair of values :

x :	152	114	138	158	144	153	141	117	136
y :	198	300	414	594	676	549	320	483	481

(G.N.D.U. B.C.A. April 2009)

18. Estimate the height of son if his father is 60 inches tall on the basis of following data :

Height of Father (in inches) :	65	66	67	67	68	69	71	72
Height of Son (in inches) :	67	68	64	68	72	74	76	78

(P.U. B.C.A. April 2005)

19. Data on the amount of fertilizer (X) and yield of wheat (Y) are given in the following table. Find the regression line assuming wheat as a function of fertilizer.

Fertilizer (Kgs.) (X) :	2	4	5	7	10	11	12	15
Yield of wheat (Kgs.) (Y) :	8	9	11	11	12	14	15	16

ANSWERS

1. 1048.1
2. (*i*) $Y = 40 + 4X$, $X = 0.8 + 0.16Y$ (*ii*) Rs. 140 crores (*iii*) Rs. 24.8 crores
3. Rs. 72.60
4. (*a*) $Y = 3.2X + 58$, $X = 0.2Y - 8$ (*b*) Rs. 106 lakh (*c*) Rs. 16 lakh
5. Rs. 61
6. (*i*) 0.64 (*ii*) $b_{YX} = 0.48$, $b_{XY} = 0.85$
   (*iii*) $Y = 0.48X + 25.44$, $X = 0.85Y + 21$ (*iv*) 61.44
7. $Y = 0.649X + 22.04$, $X = 0.365Y + 22.48$, 54.49
8. $Y = 1.6X - 10$, $X = 0.4Y + 13$, 150
9. $S = 1.015P + 8.805$, 25.045
10. $b_{YX} = 1.3$, $b_{XY} = 0.5$
11. $Y = -0.3 + 0.71X$
12. $Y = X + 1$, $X = Y - 1$, 3.5
13. $Y = 14.5 + 0.35X$, $X = 16.5 + 0.54Y$
14. $Y = 0.78X + 56.5$, $X = 0.65Y + 0.25$, 134.5
15. $Y = 0.75X + 5$, 53.75 thousand rupees
16. $Y = 4.26X + 40.80$, $X = 0.22Y - 7.03$
17. $y = 2.50x + 98.06$, $x = 0.025y + 128.07$
18. 56.17 inches
19. $Y = 0.62X + 6.885$

10.6.3 PROPERTIES OF REGRESSION COEFFICIENTS

If 'r' is the coefficient of correlation and b_{XY} and b_{YX} are the two regression coefficients in two series X and Y then there exist following properties based upon r, b_{XY} and b_{YX}.

Property I. The signs of both of the regression coefficients and the coefficient of correlation must be same.

Proof : We know that

$$b_{YX} = \frac{\text{cov}(X, Y)}{\sigma_X^2}, \quad b_{XY} = \frac{\text{cov}(X, Y)}{\sigma_Y^2} \quad \text{and} \quad r = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

Since $\sigma_X \geq 0$ and $\sigma_Y \geq 0$

$\therefore$ Signs of b_{YX}, b_{XY} and r should be the same as that of cov (X, Y).

Property II. The geometric mean of the two regression coefficients is equal to coefficient of correlation.

Proof : We know $b_{YX} = r \cdot \frac{\sigma_Y}{\sigma_X}$ and $b_{XY} = r \cdot \frac{\sigma_X}{\sigma_Y}$

$$\therefore \quad b_{YX} \cdot b_{XY} = r \cdot \frac{\sigma_Y}{\sigma_X} \times r \cdot \frac{\sigma_X}{\sigma_Y} = r^2$$

$$\therefore \quad r = \pm \sqrt{b_{YX} \cdot b_{XY}}$$

i.e. Coefficient of correlation is geometric mean of two regression coefficients.

Property III. If one of the regression coefficient is greater than unity, the other must be less than unity.

OR

The product of two regression coefficient is always less than or equal to unity.

Proof : Let one of the regression coefficient (say) b_{YX} be greater than unity then we have to show that $b_{XY} < 1$.

Since $r^2 = b_{YX}\, b_{XY}$

and $-1 \leq r \leq +1$ *i.e.* $r^2 \leq 1$

So $b_{YX}\, b_{XY} \leq 1$

$\Rightarrow$ $b_{XY} \leq \frac{1}{b_{YX}} < 1$ $\left[\because b_{YX} > 1 \Rightarrow \frac{1}{b_{YX}} < 1\right]$

Hence $b_{YX} > 1 \Rightarrow b_{XY} < 1$

Similarly $b_{XY} > 1 \Rightarrow b_{YX} < 1$

Property IV. The arithmetic mean of absolute value of two regression coefficients is equal to or greater than absolute value of coefficient of correlation.

Proof: We know that for any two positive real numbers,

$$\text{A.M.} \geq \text{G.M.}$$

In particular, for two positive real numbers $|b_{YX}|$ and $|b_{XY}|$,

$$\frac{|b_{YX}|+|b_{XY}|}{2} \geq \sqrt{|b_{YX}||b_{XY}|}$$

i.e.

$$\frac{|b_{YX}|+|b_{XY}|}{2} \geq \sqrt{|b_{YX}b_{XY}|} = |r|$$

So the mean of absolute value of two regression coefficients shall be equal to or greater than absolute value of coefficient of correlation.

Property V. Regression coefficients are independent of change of origin but not of scale.

Proof: Let X and Y are two random variables. After the shift of origin and scale, they are converted to new random variable say U and V as

$$U = \frac{X-a}{h}, \quad V = \frac{Y-b}{k}$$

$$\Rightarrow \qquad X = a + hU, \quad Y = b + kV$$

where a, b, h and k are constants, with $h > 0, k > 0$

Since $r_{YX} = r_{VU}$, $\sigma_X = h\,\sigma_U$ and $\sigma_Y = k\,\sigma_V$

$$\therefore \qquad b_{YX} = r_{YX}\frac{\sigma_Y}{\sigma_X} = r_{VU}\frac{k\,\sigma_V}{h\,\sigma_U}$$

$$= \frac{k}{h} r_{VU}\frac{\sigma_V}{\sigma_U} = \frac{k}{h} b_{VU}$$

$$\therefore \qquad b_{YX} = \frac{k}{h} b_{VU}$$

Similarly, $$b_{XY} = \frac{h}{k} b_{UV}$$

Hence regression coefficients are independent of change of origin but not of scale.

Property VI. If coefficient of correlation is zero, then the value of two regression coefficient must be equal to zero.

Proof: Given $r = 0$

$$\therefore \quad b_{YX} = r \cdot \frac{\sigma_Y}{\sigma_X} = 0$$

and $$b_{XY} = r \cdot \frac{\sigma_X}{\sigma_Y} = 0$$

Property VII. The two lines of regression intersect at the point $(\overline{X}, \overline{Y})$.

Proof : We know that the regression line of Y on X is given by

$$Y - \overline{Y} = b_{YX}(X - \overline{X})$$

and the regression line of X on Y is given by

$$X - \overline{X} = b_{XY}(Y - \overline{Y})$$

Clearly the coordinates $(\overline{X}, \overline{Y})$ satisfies both of these equations. So both of these equations pass through the point $(\overline{X}, \overline{Y})$ and hence these two lines intersect at the point $(\overline{X}, \overline{Y})$ (See fig. 10.6).

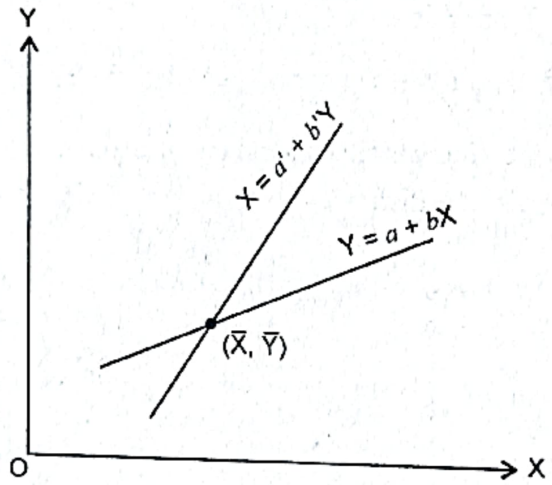


Fig. 10.6

Property VIII. If there is a perfect correlation between the two variables then both the regression lines coincide and if the two variables are uncorrelated then the regression lines will be perpendicular to each other.

Proof : If there is a perfect correlation between two variables say X and Y then $r = +1$ or -1.

Let $r = +1$

$$\therefore \quad b_{YX} = r\frac{\sigma_Y}{\sigma_X} = \frac{\sigma_Y}{\sigma_X} \quad \text{and} \quad b_{XY} = r\frac{\sigma_X}{\sigma_Y} = \frac{\sigma_X}{\sigma_Y}$$

So the regression lines of Y on X and X on Y are

$$Y - \overline{Y} = b_{YX}(X - \overline{X}) \quad \text{and} \quad X - \overline{X} = b_{XY}(Y - \overline{Y}) \text{ respectively.}$$

$$i.e. \quad Y-\overline{Y} = \frac{\sigma_Y}{\sigma_X}(X-\overline{X}) \text{ and } X-\overline{X} = \frac{\sigma_X}{\sigma_Y}(Y-\overline{Y}) \text{ respectively.}$$

$$i.e. \quad Y-\overline{Y} = \frac{\sigma_Y}{\sigma_X}(X-\overline{X}) \text{ and } Y-\overline{Y} = \frac{\sigma_Y}{\sigma_X}(X-\overline{X}) \text{ respectively.}$$

Clearly both of these two equations are identical.

Hence if $r = +1$ then the two regression lines coincide (See fig. 10.7).

Similarly if $r = -1$ then the two regression lines coincide (See fig. 10.8).

Further if the two variables say X and Y are uncorrelated *i.e.* $r = 0$ then

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X} = 0 \text{ and } b_{XY} = r\frac{\sigma_X}{\sigma_Y} = 0$$

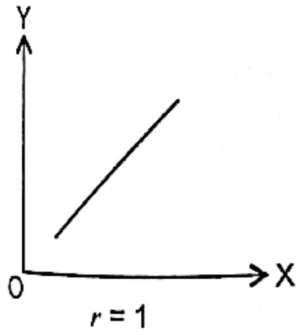
The regression equation of Y on X is

$$Y-\overline{Y} = b_{YX}(X-\overline{X}) \; i.e. \; Y-\overline{Y} = 0(X-\overline{X}) \; i.e. \; Y-\overline{Y} = 0 \; i.e. \; Y = \overline{Y}$$

and regression equation of X on Y is

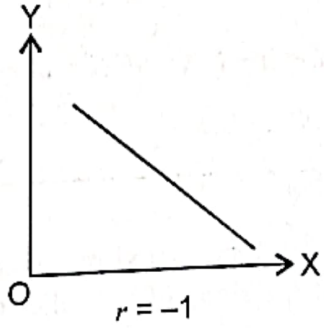
$$X-\overline{X} = b_{XY}(Y-\overline{Y}) \quad i.e. \; X-\overline{X} = 0(Y-\overline{Y}) \; i.e. \; X-\overline{X} = 0 \; i.e. \; X = \overline{X}$$

Since $Y = \overline{Y}$ represents the line parallel to x-axis at a distance $\overline{Y}$ from the x-axis and $X = \overline{X}$ represents the line parallel to y-axis at a distance $\overline{X}$ from the y-axis. So both these lines are perpendicular to each other as shown in fig. 10.9.



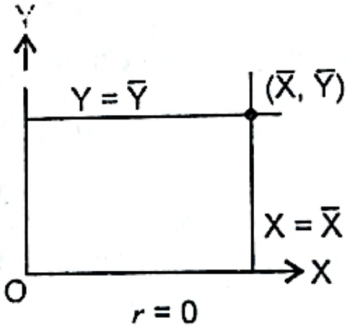
(Two lines coincide)

Fig. 10.7



(Two lines coincide)

Fig. 10.8



(Two lines are perpendicular)

Fig. 10.9

CHECKPOINTS

1. Are regression coefficients independent of scale and origin ?
2. If the coefficient of correlation is zero, what are the values of regression coefficients ?
3. At what point, the two regression equations intersect ?

ILLUSTRATIVE EXAMPLES

Example 1. If two regression coefficients are 0.8 and 1.2 then what would be the value of the coefficient of correlation ?

(P.U. B.C.A. April, 2007)

Sol. Let $b_{XY} = 0.8$ and $b_{YX} = 1.2$.

Coefficient of correlation is given by

$$r = \pm \sqrt{b_{XY} b_{YX}}$$

Since both b_{YX} and b_{XY} are positive so r should also be positive.

$$\therefore \quad r = \sqrt{b_{XY} b_{YX}} = \sqrt{0.8 \times 1.2} = 0.98.$$

Example 2. Can $Y = 5 + 2.8\,X$ and $X = 3 - 0.5\,Y$ be estimated regression equation of Y on X and X on Y respectively ? Explain your answer with suitable theoretical arguments.

(G.N.D.U. B.Sc. I.T. April 2007)

Sol. Given two regression equations as

$$Y = 5 + 2.8\,X \qquad ...(i)$$

and

$$X = 3 - 0.5\,Y \qquad ...(ii)$$

Comparing equations (*i*) and (*ii*) with $Y = a + b\,X$ and $X = a' + b'Y$ respectively, we get,

$$b = 2.8 \text{ and } b' = -0.5$$

i.e.

$$b_{YX} = 2.8 \text{ and } b_{XY} = -0.5$$

Clearly equation (*i*) and (*ii*) can not be the estimated regression equations of Y on X and X on Y respectively. This is because of the fact that the signs of both the regression coefficients b_{YX} and b_{XY} should be same which is not so in this case.

Example 3. Given, variance of X = 9, regression equations are $8X - 10Y + 66 = 0$, $40X - 18Y - 214 = 0$.

What are

(*i*) the mean values of X and Y ? (*ii*) the correlation coefficient between X and Y ?

(*iii*) standard deviation of Y ?

(P.U. B.C.A. April 2004)

Sol. Given regression equations are

$$8X - 10Y + 66 = 0 \qquad ...(1)$$

and

$$40X - 18Y - 214 = 0 \qquad ...(2)$$

(*i*) We know that the regression equations pass through the point $(\overline{X}, \overline{Y})$ so in order to find $\overline{X}$ and $\overline{Y}$, we shall solve equations given by (1) and (2).

Now multiply equation (1) by 5 and subtract from equation (2), we have,

$$32Y - 544 = 0 \quad \text{or} \quad Y = 17.$$

$\therefore$ from equation (1), we have

$$X = \frac{-66 + 10\,(17)}{8} = 13$$

Hence the mean values of X and Y are 29.5 and 17 respectively.

(*ii*) Let equation (1) be the regression line of Y on X and equation (2) be the regression line of X on Y.

Rewriting equations (1) and (2) as

$$Y = 6.6 + 0.8X \qquad ...(3)$$

and

$$X = 5.35 + 0.45Y \qquad ...(4)$$

Comparing equations (3) and (4) with the equations $Y = a + bX$ and $X = a' + b'Y$, we have

$$b = 0.8 \text{ and } b' = 0.45 \quad i.e. \quad b_{YX} = 0.8 \text{ and } b_{XY} = 0.45$$

Further $b_{YX} \cdot b_{XY} = 0.8 \times 0.45 = 0.36 < 1$

Hence our assumption regarding the selection of regression equation is correct.

Now, coefficient of correlation r is given by

$$r = \pm \sqrt{b_{YX} \cdot b_{XY}}.$$

Since both b_{YX} and b_{XY} are positive so r should also be positive.

$\therefore$

$$r = \sqrt{b_{YX} \cdot b_{XY}} = \sqrt{0.8 \times 0.45} = 0.6$$

(*iii*) Since

$$b_{XY} = r \frac{\sigma_X}{\sigma_Y}$$

$\therefore$

$$0.45 = 0.6 \frac{(3)}{\sigma_Y} \qquad [\because \sigma_X^2 = 9]$$

So

$$\sigma_Y = 4.$$

Example 4. For 50 students of a class the regression equation of marks in Statistics (X) on the marks in Accountancy (Y) is $3Y - 5X + 108 = 0$. The mean marks of Accountancy is 44 and the variance of marks in Statistics is 9/16 of the variance of marks in Accountancy. Find the mean marks of Statistics and the coefficient of correlation between marks in two subjects.

Sol. Given regression equation of X on Y as

$$3Y - 5X + 108 = 0 \qquad \ldots(i)$$

i.e. $5X = 3Y + 108$

i.e. $X = \frac{3}{5}Y + 21.6$

Compare this equation with the equation $X = a' + b'Y$, we get

$$b' = \frac{3}{5} \quad i.e. \quad b_{XY} = \frac{3}{5}$$

It is given that variance of $X = \frac{9}{16}$ variance Y or $\sigma_X^2 = \frac{9}{16}\sigma_Y^2$

$$\frac{\sigma_X^2}{\sigma_Y^2} = \frac{9}{16}$$

$$\frac{\sigma_X}{\sigma_Y} = \frac{3}{4}$$

Now $b_{XY} = r\frac{\sigma_X}{\sigma_Y}$

$\therefore$ $\frac{3}{5} = r\frac{3}{4}$

$$r = \frac{3}{5} \times \frac{4}{3} = 0.8$$

Also given, mean marks of Accountancy = 44 *i.e.* $\overline{Y} = 44$

We know that every regression line passes through the point $(\overline{X}, \overline{Y})$,

$\therefore$ from (*i*), we have,

$$3\overline{Y} - 5\overline{X} + 108 = 0$$

i.e. $3(44) - 5\overline{X} + 108 = 0$ *i.e.* $5\overline{X} = 132 + 108$

i.e. $\overline{X} = 48$

Hence, mean marks in Statistics = 48

EXERCISE 10.3

1. If the regression coefficient of X on Y is $\frac{-1}{6}$ and that of Y on X is $\frac{-3}{2}$ then what is the value of correlation coefficient between X and Y ?

(P.U. B.C.A. Sept. 2005)

2. Given $b_{xy} = 0.85$, $b_{yx} = 0.89$, $\sigma_x = 6$, find the value of 'r' and σ_y.

3. A student obtained the two regression lines as

$$2x - 5y - 7 = 0$$

and $$3x + 2y - 8 = 0$$

Do you agree with him ?

4. The equations of two lines of regression are $3x + 12y = 19$ and $9x + 3y = 46$. Find

(*i*) mean of x and mean of y.

(*ii*) the regression coefficients b_{yx} and b_{xy}.

(*iii*) coefficient of correlation between x and y. (P.U. B.C.A. April 2003)

5. Equations of two regression lines are : $3x + 2y = 26$ and $6x + y = 31$, find

(*i*) mean values of x and y (*ii*) correlation coefficient between x and y.

(P.U. B.C.A. Sept. 2003)

6. The equations of two lines of regression are $4x + 3y + 7 = 0$ and $3x + 4y + 8 = 0$. Find :

(*a*) the mean values of x and y.

(*b*) the regression coefficients b_{yx} and b_{xy}.

(*c*) the correlation coefficient between x and y.

(G.N.D.U. B.Sc. C.Sc. Sept. 2006)

7. The two regression lines are given by $y = \frac{40}{18}x - \frac{214}{18}$ and $x = \frac{10}{8}y - \frac{66}{8}$.

Find

(*i*) correlation coefficient between x and y. (*ii*) y when $x = 10$, (*iii*) x when $y = 10$

(*iv*) σ_y if $\sigma_x^2 = 9$.

8. If $y = -1.5x$ and $x = -0.3y$, determine the most likely value of coefficient of correlation.

ANSWERS

1. -0.5

2. 0.87, 6.14

3. No

4. (*i*) $\bar{x} = 5$, $\bar{y} = \frac{1}{3}$ (*ii*) $b_{yx} = -\frac{1}{4}$, $b_{xy} = -\frac{1}{3}$ (*iii*) -0.289

5. (*i*) $\bar{x} = 4$, $\bar{y} = 7$ (*ii*) -0.5

6. (*a*) $\bar{x} = -\frac{4}{7}$, $\bar{y} = -\frac{11}{7}$ (*b*) $b_{yx} = -\frac{3}{4}$, $b_{xy} = -\frac{3}{4}$ (*c*) -0.75

7. (*i*) 0.6 (*ii*) 14.6 (*iii*) 9.85 (*iv*) 4

8. -0.67

10.6.4 REGRESSION ANALYSIS IN CASE OF GROUPED SERIES

In case of bivariate frequency series or grouped series, the regression coefficients can be determined by any of the following methods :

(*i*) Direct method

The regression coefficient of Y on X is given by

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

and regression coefficient of X on Y is given by

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

(*ii*) Short-cut Method

The regression coefficient of Y on X is given by

$$b_{YX} = \frac{N \Sigma f\, dx\, dy - \Sigma f\, dx \Sigma f\, dy}{N \Sigma f\, dx^2 - (\Sigma f\, dx)^2}$$

and the regression coefficient of X on Y is given by

$$b_{XY} = \frac{N \Sigma f\, dx\, dy - \Sigma f\, dx \Sigma f\, dy}{N \Sigma f\, dy^2 - (\Sigma f\, dy)^2}$$

where $dx = X - A_x$ and $dy = Y - A_y$; A_x and A_y being the assumed means for X and Y-series respectively.

(*iii*) Step Deviation Method

The regression coefficient of Y on X is given by

$$b_{YX} = \frac{N \Sigma f\, d'x\, d'y - \Sigma f\, d'x \Sigma f\, d'y}{N \Sigma f\, d'x^2 - (\Sigma f\, d'x)^2} \times \frac{c_y}{c_x}$$

and the regression coefficient of X on Y is given by

$$b_{XY} = \frac{N \Sigma f\, d'x\, d'y - \Sigma f\, d'x \Sigma f\, d'y}{N \Sigma f\, d'y^2 - (\Sigma f\, d'y)^2} \times \frac{c_x}{c_y}$$

where $d'x = \frac{dx}{c_x}$ and $d'y = \frac{dy}{c_y}$; c_x and c_y being the common factors in dx and dy respectively.

10.6.5 ASSUMPTIONS OF REGRESSION ANALYSIS

Regression analysis is based on the following assumptions :

1. Regression analysis assumes linear relationship between related variables. It means the relationship can be expressed by a straight line $Y = a + bX$ or $X = a' + b'Y$.
2. There is no error of measurement and aggregation error in the variable.
3. The relationship under study is exactly identified.
4. The residual errors are normally distributed. They have zero mean and constant variances.

10.6.6 LIMITATIONS OF REGRESSION ANALSYSIS

The following are the limitations of regression analysis :

1. The regression analysis assumes linear relationship between variables but this assumption does not hold true in case of social sciences. In these cases, we mostly find non-linear or curvilinear relationships.
2. Regression analysis assumes static relationship between variables. It reduces its applicability in social fields.
3. The linear relationship between the variables can be ascertained within the limits. When these limits are crossed the results become incorrect or inconsistent. For example, Increase in agricultural production may be associated with an increase in use of fertilizers. But if use of fertilizer increases to a great extent, then production may fall.

Despite all these limitations, the regression analysis is considered as one of the most useful statistical technique.

CHECKPOINTS

1. What are the underlying assumptions in regression analysis ?
2. What are the limitations of regression analysis ?

ILLUSTRATIVE EXAMPLES

Example 1. By calculating the two regression coefficients obtain two regression lines from the following data :

Y → / ↓X	0 – 5	5 – 10	10 – 15
0 – 10	2	5	7
10 – 20	1	3	2
20 – 30	8	4	0

(G.N.D.U. B.C.A. April 2002)

Sol. First we construct the following table :

$c_y = 5$ $d'y = dy/c_y$	-1	0	1
$A_y = 7.5$ $dy = Y - A_y$	-5	0	5
M.V. (Y)	2.5	7.5	12.5

$c_x = 10$ $d'x = dx/c_x$	$A_x = 15$ $dx =$ $X - A_x$	M.V. (X)	Y→ ↓X	0–5	5–10	10–15	f(X)	$fd'x$	$fd'x^2$	$fd'x\,d'y$
-1	-10	5	0–10	2 | 2	5 | 0	7 | -7	14	-14	14	-5
0	0	15	10–20	1 | 0	3 | 0	2 | 0	6	0	0	0
1	10	25	20–30	8 | -8	4 | 0	0 | 0	12	12	12	-8
				f(Y) = 11	12	9	N = 32	$\Sigma fdx' = -2$	$\Sigma fd'x^2 = 26$	$\Sigma fd'x\,d'y = -13$
				$fd'y$	-11	0	9	$\Sigma fd'y = -2$		
				$fd'y^2$	11	0	9	$\Sigma fd'y^2 = 20$		
				$fdx'\,dx'$	-6	0	-7	$\Sigma fd'x\,d'y = -13$		

Now $$\overline{X} = A_x + \frac{\Sigma f\,d'x}{N}\,c_x = 15 + \frac{-2}{32} \times 10 = 14.375$$

and $$\overline{Y} = A_y + \frac{\Sigma f\,d'y}{N}\,c_y = 7.5 + \frac{-2}{32} \times 5 = 7.1875$$

Now regression coefficient of Y on X is given by

$$b_{YX} = \frac{N\Sigma f\,d'x\,d'y - \Sigma f\,d'x\,\Sigma f\,d'y}{N\Sigma f\,d'x^2 - (\Sigma f\,d'x)^2} \times \frac{c_y}{c_x}$$

$$= \frac{32(-13) - (-2)(-2)}{32 \times 26 - (-2)^2} \times \frac{5}{10} = \frac{-416 - 4}{832 - 4} \times \frac{1}{2}$$

$$= -0.25$$

... is given by

$$b_{XY} = \frac{N \Sigma f\, d'x\, d'y - \Sigma f\, d'x \, \Sigma f\, d'y}{N \Sigma f\, d'y^2 - (\Sigma f\, d'y)^2} \times \frac{c_x}{c_y}$$

$$= \frac{32(-13) - (-2)(-2)}{32(20) - (-2)^2} \times \frac{10}{5}$$

$$= \frac{-416 - 4}{640 - 4} \times 2 = -1.32$$

Also, regression equation of Y on X is given by

$$Y - \overline{Y} = b_{YX}(X - \overline{X})$$

i.e. $$Y - 7.1875 = -0.25\,(X - 14.375)$$

i.e. $$Y = -0.25\,X + 10.78$$

and regression equation of X on Y is given by

$$X - \overline{X} = b_{XY}(Y - \overline{Y})$$

i.e. $$X - 14.375 = -1.32\,(Y - 7.1875)$$

i.e. $$X = -1.32\,Y + 29.86;$$

which are the required equations of regression lines.

EXERCISE 10.4

1. Find the two lines of regression from the marks in statistics and Economics :

Marks in Economics	Marks in Statistics						Total
	40–50	50–60	60–70	70–80	80–90	90–100	
40 – 50	3	5	4	–	–	–	12
50 – 60	3	6	6	2	–	–	17
60 – 70	1	4	9	5	2	–	21
70 – 80	–	–	5	10	8	1	24
80 – 90	–	–	1	4	6	5	16
90 – 100	–	–	–	2	4	4	10
Total	7	15	25	23	20	10	100

2. Determine the regression equation of Saving on income. Estimate the amount of saving when the income is Rs. 500, from the following data.

Income (in Rs.)	Saving (in Rs.)			
	50	100	150	200
400	10	4	–	–
600	8	12	24	6
800	–	9	7	2
1000	–	–	10	5
1200	–	–	9	4

3. Following table gives the ages of husbands and Wives for 50 newly married couples. Find the two regression lines. Also estimate

(*a*) the age of husband when wife is 20 and

(*b*) the age of wife when husband is 30.

Age of Wife	Age of Husband			
	20 – 25	25 – 30	30– 35	Total
16 – 20	9	14	–	23
20 – 24	6	11	3	20
24 – 28	–	–	7	7
Total	15	25	10	50

ANSWERS

1. $Y = 0.82\,X + 10.952$, $X = 0.72\,Y + 21.36$ 2. $X = 0.106\,Y + 52.33$, Rs. 105.33

3. $X = 0.470Y + 8.03$, $Y = 0.723X + 12.02$, (*a*) 26.48 (*b*) 22.13

MISCELLANEOUS EXERCISE

1. Compute the least squares regression of Y on X from the following data :

X :	89	86	74	65	64	63	66	67	72	79
Y :	92	91	84	75	73	72	71	75	78	84

2. Given the following pair of values of X and Y :

X :	2	4	6	8	8	9
Y :	3	6	8	10	11	12

Calculate coefficient of correlation and find the regression coefficients of Y on X and X on Y.

(P.U. B.C.A. Sept. 2001)

3. Obtain two regression equations and find correlation coefficient between X and Y from the following data :

X :	10	9	7	8	11
Y :	6	3	2	4	5

(G.N.D.U. B.C.A. April 2003)

4. A small hospital is planning for expansion of its maternity wing. The past eight years data is given below :

Year :	1	2	3	4	5	6	7	8
Births :	565	590	583	597	615	611	610	623

(*a*) Use simple linear regression to forecast the annual number of births for next three years.

(*b*) Determine the correlation for the data and interpret its meaning.

(G.N.D.U. B.Sc. C.Sc. April 2005)

5. From the following data, obtain the two lines of regression :

X :	43	44	46	40	44	42	45	42	38	40	42	57
Y :	29	31	19	18	19	27	27	29	41	30	26	10

Hence obtain the value of correlation coefficient between X and Y.

(G.N.D.U. B.Sc. I.T. April 2009)

6. The regression equations of 60 observations are $5x = 6y + 24$ and $1000y = 768x - 3608$. What is the probable error of the coefficient of correlation ?

Also show that the ratio of the coefficient of variation of x to that of y is 5/24. What is the ratio of variances of x and y ?

7. If $ax + by + c = 0$ is a line of regression of y on x and $a_1x + b_1y + c_1 = 0$ that of x on y. Prove that $ab_1 \leq a_1b$.

8. The regression lines of y on x and x on y are respectively $y = ax + b$ and $x = cy + d$. Show that the means are $\bar{x} = \frac{bc+d}{1-ac}$ and $\bar{y} = \frac{ad+b}{1-ac}$ and the Correlation Coefficient between x and y is $\sqrt{ac}$. Also show that the ratio of Standard Deviation of y and x is $\sqrt{\frac{a}{c}}$ (a and c both > 0).

9. Given $x = 4y + 5$ and $y = kx + 4$, the lines of regression as $x = f(y)$ and $y = f(x)$. Show that $0 \leq 4k \leq 1$. If $k = 1/16$ determine coefficient of correlation.

10. As a furniture retailer in a certain locality you are interested in studying whether some relationship does exist between the number of building permits issued in that locality in the past years, and the volume of your sales in those years. You accordingly collect data for your sales (Y), in thousands of rupees and the number of building permits issued (X in hundreds) in the past 10 years. The results worked out are as under :

$$\Sigma X = 200,\ \Sigma Y = 2200,\ \Sigma XY = 45800,\ \Sigma X^2 = 4600 \text{ and } \Sigma Y^2 = 490400$$

Using the appropriate regression equation, find the level of sales you can expect next year when 2,000 building permits are to be issued.

11. Find the likely increased percentage of production corresponding to rainfall 90'' for the following data :

	Rainfall	Production
Average :	65''	5000 Kg
Standard Deviation :	12''	210 Kg

Coefficient of correlation = 0.7.

12. The regression equation of profits (X) on sales (Y) of a certain firm is $3Y - 5X + 108 = 0$. The average sales of a firm were Rs. 44,000 and variance of profits is 9/16 of the variance of sales. Find out average profits and coefficient of correlation between sales and profits.

13. The regression equation of production (x) on capacity utilization (y) of a certain firm is $5x - 3.830y + 146.80 = 0$. The average capacity utilization of the firm was 70% and the variance of capacity utilization is (9/16)th of the variance of production. Find the average production and coefficient of correlation between production and capacity utilization.

14. Establish the angle between the two lines of regression. Also discuss when the lines are

(*i*) Perpendicular to each other

(*ii*) Coincide with each other

ANSWERS

1. $Y = 0.808\,X + 20.92$
2. $b_{YX} = 1.240,\ b_{XY} = 0.796,\ r = 0.993$
3. $Y = 0.8\,X - 3.2,\ X = 0.8\,Y + 5.8,\ r = 0.8$
4. (*a*) 632, 639, 646 (*b*) 0.93
5. $Y = -1.38\,X + 85.64,\ X = -0.44\,Y + 54.80,\ -0.78$
6. Probable error = 0.0068, C.V. of x : C.V. of $y = 5 : 24$
9. 0.5
10. 220 thousand
11. 6.125%
12. Average Profit = Rs. 26421.6, $r = 0.8$
13. Average production = 24.26 units, $r = 0.5745$
14. $\theta = \tan^{-1}\left[\dfrac{1-r^2}{r}\cdot\dfrac{\sigma_X\sigma_Y}{\sigma_X^{\,2}+\sigma_Y^{\,2}}\right]$