UNIT 9 LINEAR REGRESSION

Structure

9.1 INTRODUCTION

In Block 2, you have studied the curve fitting, correlation, rank correlation and intra-class correlation. Correlation studies the linear relationship between two variables. Correlation coefficient measures the strength of linear relationship and direction of the correlation whether it is positive or negative. When one variable is considered as an independent variable and another as dependent variable, and if we are interested in estimating the value of dependent variable for a particular value of independent variable, we study regression analysis. For example we might be interested in estimation of production of a crop for particular amount of rainfall or in prediction of demand on the price or prediction of marks on the basis of study hours of students. In these types of cases, regression would be the choice of statisticians or researchers. In general sense, regression analysis means estimation or prediction of the unknown value of one variable from the other variable.

In this unit, you will study the concept of regression, linear regression, regression coefficient with its properties and angle between two linear regression lines. Since correlation coefficient plays very important role in understanding the regression coefficient and its properties, you are advised to see the correlation coefficient with its properties carefully. You also go through the principle of least squares which will be used in finding the regression lines. This unit will also clearly discriminate the correlation and regression.

Section 9.2 explains the concept of linear regression and Section 9.3 describes how to obtain the regression line of dependent variable y on independent variable x. Regression line considering the x as a dependent variable and y as an independent variable is also discussed. Regression coefficients of y on x and x on y are defined in Section 9.4 whereas Section 9.5 gives the properties of regression coefficients with their proofs.

Section 9.6 differentiate between correlation and regression. Angles between two linear regression lines are explained in Section 9.7.

Objectives

After reading this unit, you will be able to

- define independent variable and dependent variable;
- explain the concept of regression and linear regression;
- describe lines of regression of y on x and x on y;
- define the regression coefficients of y on x and x on y;
- explore the properties of regression coefficient;
- explain the distinction between correlation and regression; and
- define the acute angle and obtuse angle.

9.2 CONCEPT OF LINEAR REGRESSION

Prediction or estimation is one of the major problems in most of the human activities. Like prediction of future production of any crop, consumption, price of any good, sales, income, profit, etc. are very important in business world. Similarly, prediction of population, consumption of agricultural product, rainfall, revenue, etc. have great importance to the government of any country for effective planning.

If two variables are correlated significantly, then it is possible to predict or estimate the values of one variable from the other. This leads us to very important concept of regression analysis. In fact, regression analysis is a statistical technique which is used to investigate the relationship between variables. The effect of price increase on demand, the effect of change in the money supply on the increase rate, effect of change in expenditure on advertisement on sales and profit in business are such examples where investigators or researchers try to construct cause and affect relationship. To handle these type of situations, investigators collect data on variables of interest and apply regression method to estimate the quantitative effect of the causal variables upon the variable that they influence.

Regression analysis describes how the independent variable(s) is (are) related to the dependent variable i.e. regression analysis measures the average relationship between independent variables and dependent variable. The literal meaning of regression is “stepping back towards the average” which was used by British Biometrician Sir Francis Galton (1822-1911) regarding the height of parents and their offspring’s.

Regression analysis is a mathematical measure of the average relationship between two or more variables.

There are two types of variables in regression analysis:

1. Independent variable
2. Dependent variable

The variable which is used for prediction is called independent variable. It is also known as regressor or predictor or explanatory variable.

The variable whose value is predicted by the independent variable is called dependent variable. It is also known as regressed or explained variable.

If scatter diagram shows some relationship between independent variable X and dependent variable Y, then the scatter diagram will be more or less concentrated round a curve, which may be called the curve of regression.

When the curve is a straight line, it is known as line of regression and the regression is said to be linear regression.

If the relationship between dependent and independent variables is not a straight line but curve of any other type then regression is known as non-linear regression.

Regression can also be classified according to number of variables being used. If only two variables are being used this is considered as simple regression whereas the involvement of more than two variables in regression is categorized as multiple regression.

Let us solve some little exercises.

E1) What do you mean by independent and dependent variables?
E2) Define regression.

9.3 LINES OF REGRESSION

Regression lines are the lines of best fit which express the average relationship between variables. Here, the concept of lines of best fit is based on principle of least squares.

Let X be the independent variable and Y be the dependent variable and we have observations (x_i, y_i); $i = 1, 2, ..., n$; Let the equation of line of regression of y on x be

$$y = a + bx \qquad \ldots (1)$$

In this case, the value of the dependent variable changes at a constant rate as a unit change in the independent variable or explanatory variable.

Let $Y_i = a + bx_i$ be the estimated value of y_i for the observed or given value of $x = x_i$. According to the principle of least squares, we have to determine a and b so that the sum of squares of deviations of observed values of y from expected values of y i.e.

$$U = \sum_{i=1}^{n} (y_i - Y_i)^2$$

or, $$U = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \qquad \ldots (2)$$

is minimum. From the principle of maxima and minima, we take partial derivatives of U with respect to a and b and equating to zero, i.e.

$$\frac{\partial U}{\partial a} = 0$$

$$\Rightarrow \frac{\partial}{\partial a}\sum_{i=1}^{n}(y_i - a - bx_i)^2 = 0$$

$$\Rightarrow 2\sum_{i=1}^{n}(y_i - a - bx_i)(-1) = 0$$

$$\Rightarrow \sum_{i=1}^{n}(y_i - a - bx_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \quad \ldots (3)$$

and $\frac{\partial U}{\partial b} = 0$

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

$$\Rightarrow 2\sum_{i=1}^{n}(y_i - a - bx_i)(-x_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n}(y_i x_i - ax_i - bx_i^2) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \quad \ldots (4)$$

Equations (3) and (4) are known as normal equations for straight line (1).
Dividing equation (3) by n, we get

$$\bar{y} = a + b\bar{x} \quad \ldots (5)$$

This indicates that regression line of y on x passes through the point $(\bar{x}, \bar{y})$.
By the definition of covariance, we know that

$$\mathrm{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$\mathrm{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \mathrm{Cov}(x, y) + \bar{x}\,\bar{y} \quad \ldots (6)$$

The variance of variable x can be expressed as

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \sigma_x^2 + \bar{x}^2 \qquad \ldots (7)$$

Dividing equation (4) by n we get

$$\frac{1}{n}\sum_{i=1}^{n} y_i x_i = a\frac{1}{n}\sum_{i=1}^{n} x_i + b\frac{1}{n}\sum_{i=1}^{n} x_i^2 \qquad \ldots (8)$$

Using equations (6) and (7) in equation (8) gives

$$Cov(x, y) + \bar{x}\,\bar{y} = a\bar{x} + b(\sigma_x^2 + \bar{x}^2) \qquad \ldots (9)$$

Multiplying equation (5) by $\bar{x}$, we get

$$\bar{y}\,\bar{x} = a\bar{x} + b\bar{x}^2 \qquad \ldots (10)$$

Subtracting equation (10) from equation (9), we get

$$Cov(x, y) = b\sigma_x^2$$

$$\Rightarrow b = \frac{Cov(x, y)}{\sigma_x^2} \qquad \ldots (11)$$

Since b is the slope of the line of regression of y on x and the line of regression passes through the point $(\bar{x}, \bar{y})$, so the equation of line of regression of y on x is

$$(y - \bar{y}) = b(x - \bar{x})$$

$$= \frac{Cov(x, y)}{\sigma_x^2}(x - \bar{x}) \qquad \left(\because b = \frac{Cov(x, y)}{\sigma_x^2}\right)$$

$$= \frac{r\sigma_x\sigma_y}{\sigma_x^2}(x - \bar{x}) \qquad (Cov(x, y) = r\sigma_x\sigma_y)$$

$$(y - \bar{y}) = \frac{r\sigma_y}{\sigma_x}(x - \bar{x}) \qquad \ldots (12)$$

This is known as regression line of y on x.

If we consider the straight line $x = c + dy$ and proceeding similarly as in case of equation (1), we get the line of regression of x on y as

$$(x - \bar{x}) = \frac{Cov(x, y)}{\sigma_y^2}(y - \bar{y})$$

$$(x - \bar{x}) = \frac{r\sigma_x}{\sigma_y}(y - \bar{y}) \qquad \ldots (13)$$

Therefore, we have two lines of regression, one of y on x and other of x on y. In case of perfect correlation $(r = \pm 1)$, either perfect positive or perfect negative, both lines of regression coincide and we have single line.

Lines of regression provide the best estimate to the value of dependent variable for specific value of the independent variable.

Equation (12) shows that the regression line of y on x passes through $(\bar{x}, \bar{y})$ and equation (13) also implies that regression line of x on y passes through $(\bar{x}, \bar{y})$.

Passing of both lines through the points $(\bar{x}, \bar{y})$ indicates that the point $(\bar{x}, \bar{y})$ is the intersection point of both lines. We can also conclude that solution of both lines as simultaneous equations provide mean of both variables, i.e. $\bar{x}$ and $\bar{y}$.

Regression line of y on x is used to estimate or predict the value of dependent variable y for the given value of independent variable x. Estimate of y obtained by this line will be best because this line minimizes the sum of squares of the errors of the estimates in y. If x is considered as dependent variable and y as independent variable then regression line of x on y is used to estimate or predict the value of variable x for the given value of y. Estimate of x obtained by regression line of x on y will be best because it minimizes the sum of squares of the errors of the estimates in x.

It is important to know that these regression lines y on x and x on y are different. These lines can't be interchanged. Regression line of y on x cannot be used for the prediction of independent variable x. Similarly regression line of x on y cannot be used for the prediction of independent variable y.

When two regression lines cut each other at right angle (at the angle of 90 degree), it shows no correlation between y and x.

Let us solve some exercises.

E3) Which line is used for the prediction of dependent variable x and why?

E4) How many regression lines exist when two variables are considered in regression analysis and why?

9.4 REGRESSION COEFFICIENTS

If regression line of y on x is

$$(y - \bar{y}) = \frac{r\sigma_y}{\sigma_x}(x - \bar{x})$$

Then $\frac{r\sigma_y}{\sigma_x}$ is called the regression coefficient of y on x and it is denoted by b_{yx}. Thus, Regression coefficient of y on x ,

$$b_{yx} = \frac{r\sigma_y}{\sigma_x}$$

It gives the change in the dependent variable y as a unit change in the independent variable x. If we are considering the production of a crop by y and amount of rain fall by x, then regression coefficient of y on x represents change in production of crop as a unit change in rainfall i.e. if rainfall increases or decreases by one cm or by one inch how much production of a

crop increases or decreases, is given by regression coefficient b_{yx}. In another example of investment in advertisement (x) and sales revenue (y), regression coefficient b_{yx} gives the change in sales revenue when the investment in advertisement is changed by a unit (a unit may be one thousand or one lakh or any other convenient figure).

Similarly, the regression coefficient of x on y gives the change in the value of dependent variable x as a unit change in the value of independent variable y and it is defined as

Regression coefficient of x on y, $b_{xy} = \frac{r\sigma_x}{\sigma_y}$.

For calculation purpose we can use the formula of regression coefficient of x on y as

$$b_{xy} = \frac{r\sigma_x}{\sigma_y}$$

$$= \frac{\left(\sum (x-\bar{x})(y-\bar{y}) \Big/ \sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}\right)\sqrt{\frac{1}{n}\sum (x-\bar{x})^2}}{\sqrt{\frac{1}{n}\sum (y-\bar{y})^2}}$$

$$b_{xy} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (y-\bar{y})^2}$$

Similarly, $b_{yx} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$

9.5 PROPERTIES OF REGRESSION COEFFICIENTS

Property 1: Geometric mean of the regression coefficients is correlation coefficient.

Description: If regression coefficient of y on x is b_{yx} and regression coefficient of x on y is b_{xy} then geometric mean of b_{yx} and b_{xy} is correlation coefficient i.e. $\sqrt{b_{yx} \times b_{xy}} = r$.

Proof: If regression coefficient of y on x is b_{yx} and regression coefficient of x on y is b_{xy}, then

$$b_{yx} \times b_{xy} = r\frac{\sigma_y}{\sigma_x} \times r\frac{\sigma_x}{\sigma_y}$$

$$\Rightarrow b_{yx} \times b_{xy} = r^2$$

$$\Rightarrow \pm\sqrt{b_{yx} \times b_{xy}} = r$$

It shows that geometric mean of regression coefficients is correlation coefficient.

Property 2: If one of the regression coefficients is greater than one, then other must be less than one.

Description: If b_{yx} is greater than one then b_{xy} must be less than one.

Proof: Let b_{yx}, the regression coefficient of y on x is greater than one i.e.

$$b_{yx} > 1$$

or

$$\frac{1}{b_{yx}} < 1$$

We know that

$$r^2 \le 1 \Rightarrow b_{yx} \times b_{xy} \le 1 \qquad \text{(From Property 1)}$$

$$\Rightarrow b_{xy} \le \frac{1}{b_{yx}} < 1.$$

Thus if b_{yx} is greater than one then b_{xy} is less than one.

Property 3: Arithmetic mean of the regression coefficients is greater than the correlation coefficient i.e. $\frac{1}{2}(b_{yx} + b_{xy}) \ge r$, subject to the condition r > 0.

Proof: Suppose that arithmetic mean of regression coefficients is greater than correlation coefficient thus,

$$\frac{1}{2}(b_{yx} + b_{xy}) \ge r$$

$$\Rightarrow (b_{yx} + b_{xy}) \ge 2r$$

$$\Rightarrow (b_{yx} + b_{xy}) \ge 2(\pm\sqrt{(b_{xy} \times b_{yx})}$$

(From Property 1 $r = \pm\sqrt{b_{yx} \times b_{xy}}$)

Therefore,

$$\Rightarrow (b_{yx} + b_{xy}) \mp 2\sqrt{b_{yx}} \times \sqrt{b_{xy}} \ge 0$$

$$\Rightarrow (\sqrt{b_{yx}} \mp \sqrt{b_{xy}})^2 \ge 0$$

which is always true since the square of a real quantity is always positive.

Thus, $\frac{1}{2}(b_{yx} + b_{xy}) \ge r$, i.e. arithmetic mean of regression coefficients is greater than correlation coefficient.

Let us do some problems related to regression coefficients.

Example 1: Height of fathers and sons in inches are given below:

Height of Father	65	66	67	67	68	69	70	71
Height of Son	66	68	65	69	74	73	72	70

Find two lines of regression and calculate the estimated average height of son when the height of father is 68.5 inches.

Solution: Let us denote the father's height by x and son's height by y then two lines of regression can be expressed as

$$(y-\bar{y}) = \frac{r\sigma_y}{\sigma_x}(x-\bar{x}) \text{ and}$$

$$(x-\bar{x}) = \frac{r\sigma_x}{\sigma_y}(y-\bar{y})$$

To find the regression lines we need $\bar{x}$, $\bar{y}$, σ_x, σ_y and r which will be obtained from the following table:

x	y	x^2	y	xy
65	66	4225	4356	4290
66	68	4356	4624	4488
67	65	4489	4225	4355
67	69	4489	4761	4623
68	74	4624	5476	5032
69	73	4761	5329	5037
70	72	4900	5184	5040
71	70	5041	4900	4970
$\sum x = 543$	$\sum y = 557$	$\sum x^2 = 36885$	$\sum y^2 = 38855$	$\sum xy = 37835$

Now,

Mean of x = $\bar{x} = \frac{1}{n}\sum x = \frac{1}{8}543 = 67.88$

Mean of y = $\bar{y} = \frac{1}{n}\sum y = \frac{1}{8}557 = 69.62$

Standard Deviation of x = $\sigma_x = \sqrt{\frac{1}{n}\sum (x-\bar{x})^2} = \sqrt{\frac{1}{n}\sum x^2 - \bar{x}^2}$

$$= \sqrt{\frac{1}{8}\times 36885 - (67.88)^2}$$

$$= \sqrt{4610.62 - 4607.69}$$

$$= \sqrt{2.93} = 1.71$$

Similarly,

Standard deviation of y = $\sigma_y = \sqrt{\frac{1}{n}\sum (y_i-\bar{y})^2} = \sqrt{\frac{1}{n}\sum y^2 - \bar{y}^2}$

$$= \sqrt{\frac{1}{8}\times 38855 - (69.62)^2}$$

$$= \sqrt{4856.88 - 4846.94}$$

$$= \sqrt{9.94} = 3.15$$

Now, correlation coefficient

$$r = \text{Corr}(x, y) = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left\{n\sum x^2 - \left(\sum x\right)^2\right\}\left\{n\sum y^2 - \left(\sum y\right)^2\right\}}}$$

$$= \frac{8\times 37835 - (543)\times(557)}{\sqrt{\left\{(8\times 36885 - (543)^2\right\}\left\{(8\times 38855) - (557)^2\right\}}}$$

$$= \frac{302680 - 302451}{\sqrt{\{(295080 - 294849\}\{310840 - 310249\}}}$$

$$= \frac{229}{\sqrt{231\times 591}}$$

$$= \frac{229}{\sqrt{136521}} = \frac{229}{369.49}$$

$$r = 0.62$$

Substituting the value of $\bar{x}$, $\bar{y}$, σ_x , σ_y and r in regression equations, we get regression line of y on x as

$$(y - 69.62) = 0.62 \times \frac{3.15}{1.71}(x - 67.88)$$

$$y = 1.14x - 77.38 + 69.62$$

$$y = 1.14x - 7.76$$

and regression line of x on y

$$(x - 67.88) = 0.62 \times \frac{1.71}{3.15}(y - 69.62)$$

$$x = 0.34\,y - 23.67 + 67.88$$

$$x = 0.34\,y + 44.21$$

Estimate of height of son for the height of father = 68.5 inch is obtained by the regression line of y on x

$$y = 1.14x - 7.76$$

Putting x = 68.5 in above regression line

$$y = 1.14 \times 68.50 - 7.76 = 78.09 - 7.76 = 70.33$$

Thus, the estimate of son's height for the father's height 68.5 inch is 70.33 inch.

Note: For the estimation of x for the given value of y, we use regression line of x on y whereas for the estimation of y for the given value of x we use regression line of y on x.

Example 2: Regression line of y on x and x on y respectively are

$$2x - 3y = -8$$

$$5x - y = 6$$

Then, find

(i) the mean values of x and y,

(ii) coefficient of correlation between x and y, and

(iii) the standard deviation of y for given variance of x = 5.

Solution:

(i) Since regression lines of y on x and x on y passes through $\bar{x}$ and $\bar{y}$ so $\bar{x}$ and $\bar{y}$ are the intersection points. Thus to get the mean values of variable x and y, we solve given simultaneous equations

$$2x - 3y = -8$$

$$5x - y = 6$$

By solving these equations as simultaneous equations we get x =2 and y = 4 which are means of x and y respectively.

Note: You have already solved simultaneous equations in Unit 1 of the Block-2 during the fitting of various curves for given data. If you face some problems in solving these equations go through problems given in Unit 5 of the Block 2.

(ii) To find the correlation coefficient, given equations are expressed as

$$y = 2.67 + 0.67x \quad \ldots (14)$$

$$x = 1.20 + 0.20y \quad \ldots (15)$$

Note: To find the regression coefficient of y on x, regression line of y on x is expressed in the form of $y = a + bx$, where b is the regression coefficient of y on x. Similarly, to find the regression coefficient of x on y, regression line of x on y is expressed in the form of $x = c + dy$, where d is the regression coefficient of x on y. In our problem, by dividing the first line by 3, i.e. by the coefficient of y gives equation in the form $y = a + bx$ which is y = 2.67 +0.67 x. Similarly, dividing the second regression equation by 5 gives equation (15).

From equations (14) and (15) we, find the regression coefficient of y on x and x on y respectively as

$$b_{yx} = \frac{r\sigma_y}{\sigma_x} = 0.67 \text{ and}$$

$$b_{xy} = \frac{r\sigma_x}{\sigma_y} = 0.20$$

By the property of regression coefficients

$$\pm\sqrt{b_{yx} \times b_{xy}} = r \Rightarrow r = \sqrt{0.67 \times 0.20} = 0.37$$

Thus, correlation coefficient r = 0.37

Note: We are taking $(+)$ sign because correlation coefficient and regression coefficients have same sign.

(iii) By the definition of regression coefficient of y on x i.e.

$$b_{yx} = \frac{r\sigma_y}{\sigma_x} = 0.67$$

Variance of x i.e., $\sigma_x^2 = 5 \Rightarrow \sigma_x = 2.24$

Now,

$$b_{yx} = \frac{r\sigma_y}{\sigma_x} \Rightarrow 0.67 = \frac{0.37\sigma_y}{2.24} \Rightarrow \sigma_y = 4.05,$$

$$\Rightarrow \sigma_y^2 = (4.05)^2 = 16.45$$

Thus, the variance of y is 16.45.

E5) What is the geometric mean of regression coefficients?

E6) Both regression coefficients can have same sign. Do you agree?

E7) Marks of 6 students of a class in paper I and paper II of Statistics are given below:

Paper I	45	55	66	75	85	100
Paper II	56	55	45	65	62	71

Find

(i) both regression coefficients,

(ii) both regression lines, and

(iii) correlation coefficient.

E8) We have data on variables x and y as

x	5	4	3	2	1
y	9	8	10	11	12

Calculate

(i) both regression coefficients,

(ii) correlation coefficient,

(iii) regression lines of y on x and x on y, and

(iv) estimate y for x = 4.5.

E9) If two regression lines are

$$6x + 15y = 27$$

$$6x + 3y = 15,$$

Then, calculate

(i) correlation coefficient, and

(ii) mean values of x and y.

9.6 DISTINCTION BETWEEN CORRELATION AND REGRESSION

Both correlation and regression have important role in relationship study but there are some distinctions between them which can be described as follow:

(i) Correlation studies the linear relationship between two variables while regression analysis is a mathematical measure of the average relationship between two or more variables.

(ii) Correlation has limited application because it gives the strength of linear relationship while the purpose of regression is to "predict" the value of the dependent variable for the given values of one or more independent variables.

(iii) Correlation makes no distinction between independent and dependent variables while linear regression does it, i.e. correlation does not consider the concept of dependent and independent variables while in regression analysis one variable is considered as dependent variable and other(s) is/are as independent variable(s).

9.7 ANGLE BETWEEN TWO LINES OF REGRESSION

When we consider y as dependent variable and x as independent variable than regression line of y on x is

$$(y-\bar{y}) = r\frac{\sigma_y}{\sigma_x}(x-\bar{x})$$

Similarly, when x is considered as dependent variable and y as an independent variable then regression line of x on y is

$$(x-\bar{x}) = r\frac{\sigma_x}{\sigma_y}(y-\bar{y})$$

If m_1 and m_2 are the slopes of two lines and θ be the angle between them therefore

$$\tan\theta = \left|\frac{m_1 - m_2}{1+m_1 m_2}\right|$$

For these regression lines it is observed that the slope of regression line of y on x is $\frac{r\sigma_y}{\sigma_x}$ and slope of regression line of x on y is $\frac{\sigma_y}{r\sigma_x}$. If the angle between the two lines of regression is denoted by θ, then

$$\theta = \tan^{-1}\left\{\frac{(1-r^2)}{r}\left(\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}\right)\right\} \quad \ldots (16)$$

If $r = 0$ i.e. variables are uncorrelated then

$$\tan\theta = \infty \Rightarrow \theta = \frac{\pi}{2}$$

In this case lines of regression are perpendicular to each other.

If $r = \pm 1$ i.e. variables are perfect positive or negative correlated then

$$\tan\theta = 0 \Rightarrow \theta = 0 \text{ or } \pi$$

In this case, regression lines either coincide or parallel to each other.

There are two angles between regression lines whenever two lines intersect each other, one is acute angle and another is obtuse angle. The tan θ would be greater than zero if θ lies between 0 and $\frac{\pi}{2}$ then θ is called acute angle, which is obtained by

$$\theta_1 = \text{acute angle} = \tan^{-1}\left\{\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}.\frac{1-r^2}{r}\right\}, \quad r>0, \quad \text{if } 0<\theta<\frac{\pi}{2}$$

The tan θ would be less than zero if θ lies between $\frac{\pi}{2}$ and π then θ is called obtuse angle, which is obtained by

$$\theta_2 = \text{obtuse angle} = \tan^{-1}\left\{\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}.\frac{r^2-1}{r}\right\}, \quad r>0, \quad \text{if } \frac{\pi}{2}<\theta<\pi$$

9.8 SUMMARY

In this unit, we have discussed:

1. The concept of regression,
2. How to obtain lines of regression and regression coefficient,
3. Properties of regression coefficients,
4. How to use regression line for the prediction of dependent variable on the basis of value of independent variable,
5. The difference between correlation and regression, and
6. The angle between the two regression lines.

9.9 SOLUTIONS /ANSWERS

E1) The variable which is used for prediction is called independent variable. It is also known as regressor or predictor or explanatory variable.

The variable whose value is predicted by the independent variable is called dependent variable. It is also known as regressed or explained variable.

E2) If two variables are correlated significantly, then it is possible to predict or estimate the values of one variable from the other. This leads us to very important concept of regression analysis. In fact, regression

analysis is a statistical technique which is used to investigate the relationship between variables. The effect of price increase on demand, the effect of change in the money supply on the inflation rate, effect of change in expenditure on advertisement on sales and profit in business are such examples where investigators or researchers try to construct cause and affect relationship. To handle these type of situations investigators collect data on variables of interest and apply regression method to estimate the quantitative effect of the causal variables upon the variable that they influence.

The literal meaning of regression is "stepping back towards the average". It was first used by British Biometrician Sir Francis Galton (1822-1911) in connection with the height of parents and their offspring's.

Regression analysis is a mathematical measure of the average relationship between two or more variables.

E3) If x is considered as dependent variable and y as independent variable then regression line of x on y is used to estimate or predict the value of variable x for the given value of y. Estimate of x obtained by regression line of x on y will be best because it minimizes the sum of squares of errors of estimates in x.

E4) When two variables are considered in regression analysis, There are two regression lines

(i) Regression line of y on x and

(ii) Regression line of x on y.

Regression line of y on x is used to estimate or predict the value of dependent variable y for the given value of independent variable x. Estimate of y obtained by this line will be best because this line minimizes the sum of squares of the errors of the estimates in y. If x is considered as dependent variable and y as independent variable then regression line of x on y is used to estimate or predict the value of variable x for the given value of y. Estimate of x obtained by regression line of x on y will be best because it minimizes the sum of squares of the errors of the estimates in x.

E5) Correlation coefficient is the geometric mean of the regression coefficients.

Description: If regression coefficient of y on x is b_{yx} in and regression coefficient of x on y is b_{xy}, then $\sqrt{b_{yx} \times b_{xy}} = r$.

Proof:

If regression coefficient of y on x is b_{yx} and regression coefficient of x on y is b_{xy}, then

$$b_{yx} \times b_{xy} = r\frac{\sigma_y}{\sigma_x} \times r\frac{\sigma_x}{\sigma_y}$$

$$\Rightarrow b_{yx} \times b_{xy} = r^2$$

$$\Rightarrow \pm\sqrt{b_{yx} \times b_{xy}} = r$$

It shows that geometric mean of regression coefficients is correlation coefficient.

E6) Yes, both regression coefficients have same sign (positive or negative).

E7) (i) Let us denote the marks in paper I by x and marks in paper II by y then, here we will use the direct formula of b_{yx} and b_{xy} which are

$$b_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(y-\bar{y})^2} \text{ and}$$

$$b_{yx} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

Now, $\bar{x} = \frac{\sum x}{n} = \frac{426}{6} = 71$

and $\bar{y} = \frac{\sum y}{n} = \frac{354}{6} = 59$

S. No.	x	y	$(x-\bar{x})$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
1	45	56	-26	676	-3	9	78
2	55	55	-16	256	-4	16	64
3	66	45	-5	25	-14	196	70
4	75	65	4	16	6	36	24
5	85	62	14	196	3	9	42
6	100	71	29	841	12	144	348
Total	426	354	0	2010	0	410	626

Thus $b_{yx} = \frac{626}{2010} = 0.31$

$b_{xy} = \frac{626}{410} = 1.53.$

(ii) Regression line of y on x is $(y-\bar{y}) = b_{yx}(x-\bar{x})$ and

the regression line of x on y is $(x-\bar{x}) = b_{xy}(y-\bar{y})$

So we need $\bar{y}$, $\bar{x}$, b_{yx} and b_{xy}. In (i) we have calculated b_{yx} and b_{xy}.

Thus, the Regression line of y on x is $(y-59) = 0.31(x-71)$

$\Rightarrow y = 0.31x - 22.01 + 51$

$\Rightarrow y = 0.31x + 36.99$

Regression line of x on y is $(x-71) = 1.53(y-59)$

$\Rightarrow x = 1.53y - 19.27$

(iii) By the first property of regression coefficients, we know that

$r = \pm\sqrt{b_{yx} \times b_{xy}} \Rightarrow \sqrt{0.31 \times 1.53} = 0.68$

E8) Here, for the calculation in table

$\bar{x} = \frac{\sum x}{n} = \frac{15}{3} = 3$ and $\bar{y} = \frac{\sum y}{n} = \frac{50}{5} = 10$

S. No.	x	y	$(x-\bar{x})$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
1	5	9	2	4	-1	1	-2
2	4	8	1	1	-2	4	-2
3	3	10	0	0	0	0	0
4	2	11	-1	1	1	1	-1
5	1	12	-2	4	2	4	-4
Total	15	50	0	10	0	10	-9

(i) Regression coefficient of x on y $b_{xy} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (y-\bar{y})^2}$

$b_{xy} = \frac{-9}{10} = -0.9$

Regression coefficient of y on x $b_{yx} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$

$b_{yx} = \frac{-9}{10} = -0.9$

(ii) $r = \pm\sqrt{b_{yx} \times b_{xy}} \Rightarrow \pm\sqrt{(-0.9)\times(-0.9)} = -0.90$

Note: We are taking $(-)$ sign because correlation coefficient and regression coefficients have same sign.

(iii) To find regression lines we need $\bar{y}$, $\bar{x}$, b_{yx} and b_{xy}. From the calculation in table

$$\bar{x} = \frac{\sum x}{n} = \frac{15}{5} = 3 \text{ and}$$

$$\bar{y} = \frac{\sum y}{n} = \frac{50}{5} = 10$$

Thus, regression line of y on x $(y-10) = -0.90(x-3)$

Regression line of x on y $(x-3) = -0.90\,(y-10)$

(iv) To estimate y we use regression line of y on x which is

$$(y-10) = -0.9\,(x-3)$$

putting x = 4.5 we get

$$(y-10) = -0.9\,(4.5-3)$$

$$y = 8.65$$

E9) (i) To find correlation coefficient, we need regression coefficients which can be obtained from the given regression lines by presenting them in the following form

$y = 1.80 - 0.40x$ which gives $b_{yx} = -0.40$

and

$x = 2.50 - 0.50y$ which gives $b_{xy} = -0.50$

$$r = \pm\sqrt{b_{yx} \times b_{xy}} = \pm\sqrt{(-0.4)\times(-0.5)} = -0.44$$

(ii) Since both regression lines passes through means of y and x so the solution of these equation as a simultaneous equation gives mean values. Subtracting second equation from first equation we have

$6x + 15y = 27$

$6x + 3y = 15$

$12\,y = 12 \Rightarrow y = 1$

Substituting the value of y in first equation, we get x = 2.

Thus x = 2 and y = 1 are mean values of variables x and y respectively i.e. $\bar{x} = 2, \bar{y} = 1$