Chapter 12
SIMPLE REGRESSION ANALYSIS

12.1 Introduction

Correlation coefficient measures the direction and strength of linear relationship between two variables. It does not measure the cause-and-effect relationship between two variables. By regression analysis we can measure the cause-and-effect relationship between two variables.

The concept of regression was first introduced by a British biometrician, Sir Francis Galton in 1877 while studying the relationship between the heights of fathers and sons. He found that tall fathers tend to have tall sons and short fathers' short sons; but the average height of the sons of a group of tall fathers is less than that of the tall fathers and the average height of the sons of a group of short fathers is greater than that of the short fathers. But now it is widely used in statistics.

Today regression analysis is a very powerful tool in the field of statistical analysis in predicting the value of one variable on the basis of the given value of another variable, when these two variables are related to each other.

There are many situations in business and other fields where we are interested to measure the relationship between two variables.

Some examples of these related variables are

i) Fertilizer used and yield of various plots of land.
ii) Ages of husbands and ages of wives of a group of couples.
iii) Incomes and expenditures of a class of people.
iv) The price of a commodity and amount demanded.
v) The advertising expenditures and the volume of sales of a product.
vi) The volume of sales and the experience of the salesman of a departmental store.
vii) The deposit in a bank and the number of clients.
viii) The performance of a student in a high school and the performance of the same student in the college.
ix) The heights and weights of students in a class etc.

In all the above examples, there are two variables involved, the value of one variable depending upon the value of the other variable. For example, within the limits, the yield of a plot depends upon the kind and amount of fertilizer used. Hence the yield variable y is known as the dependent variable and the fertilizer variable x is known as independent variable. The expenditure y of a person would depend upon the income x of the person. Similarly, expenditure of a family depends on his income, demand of a commodity depends on price, sale of a store depends on its experience salesman etc. Since the value of y is related to the value of x, knowing this relationship would help us predict the value of y for a given value of x. This general process of predicting the value of dependent variable y on the basis of known value of the independent variable x is known as the regression analysis.

12.1.1 Dependent and independent variables. In regression analysis there are two types of variables. They are known as (i) dependent variable and (ii) independent variable.

i) **Dependent variable :** The variable whose value is influenced or is to be predicted is called dependent variable. It is usually denoted by y. Dependent variable is also known

Scanned with CamScanner

as explained variable, predictand, regressed, response or endogenous variable. In example (i) yield per plot y is called dependent variable.

ii) **Independent variable :**. The variable, which influences the values or is to use for prediction, is called independent variable. It is usually denoted by x. In regression analysis independent variable is also known as explanatory variable, predictor, regressor, control or exogenous variable. In example (i) amount of fertilizer x per plot is known as independent variable.

12.1.2 Regression Analysis. Regression analysis is a statistical technique, which has developed to study and measure the statistical relationship among two or more variables with a vision to estimate or predict the value of dependent variable for some known value of the independent variable.

In this section, we shall study the relationship between two variables only. The process is called Simple **Regression.**

12.1.3 The purpose of regression analysis. The main purpose of simple regression analysis is to

i) Establish a functional or mathematical relationship between dependent and independent variables;

ii) Estimate or predict the values of the dependent variable for given values of the independent variable;

iii) Show the pattern or trend of the dependent variables for various values of the independent variable.

Broadly speaking, the mathematical or functional relationship between the dependent and independent variable may be (i) linear or (ii) non-linear.

Examples of linear relation. (i) $y = a + bx$; (ii) $y = a - bx$; (iii) $y = x$ etc.

Examples of non-linear relation. (i) $y = x^2$; (ii) $y = \sqrt{x}$; (iii) $y = a + bx + cx^2$ etc.

In simple regression analysis, we consider the linear or straight-line relationship between the variables. In simple linear regression, a mathematical regression equation is developed to describe the functional relationship that exists between the two variables x and y, and this association is exhibited by plotting the values of paired coordinated (XY) on a graph, with dependent variable y along the vertical Y-axis and the independent variable x along the horizontal X-axis.

12.2 Population Regression Line and Model

Suppose we have N individuals in a population. Suppose we want to study two characteristics say income and expenditure of the individuals by two variables X and Y respectively. If the expenditure variable Y functionally related with income variable X, then the population simple regression equation is a straight line of Y on X and is defined as

$$Y = \alpha + \beta X \qquad \text{... (12.1)}$$

where α is the Y- intercept- the value of Y when $X = 0$ and β is the slope of the line, defined as the change in Y for a one-unit change in X as shown in Fig. 12.1. In regression analysis β is called the regression coefficient of Y on X.

Scanned with CamScanner

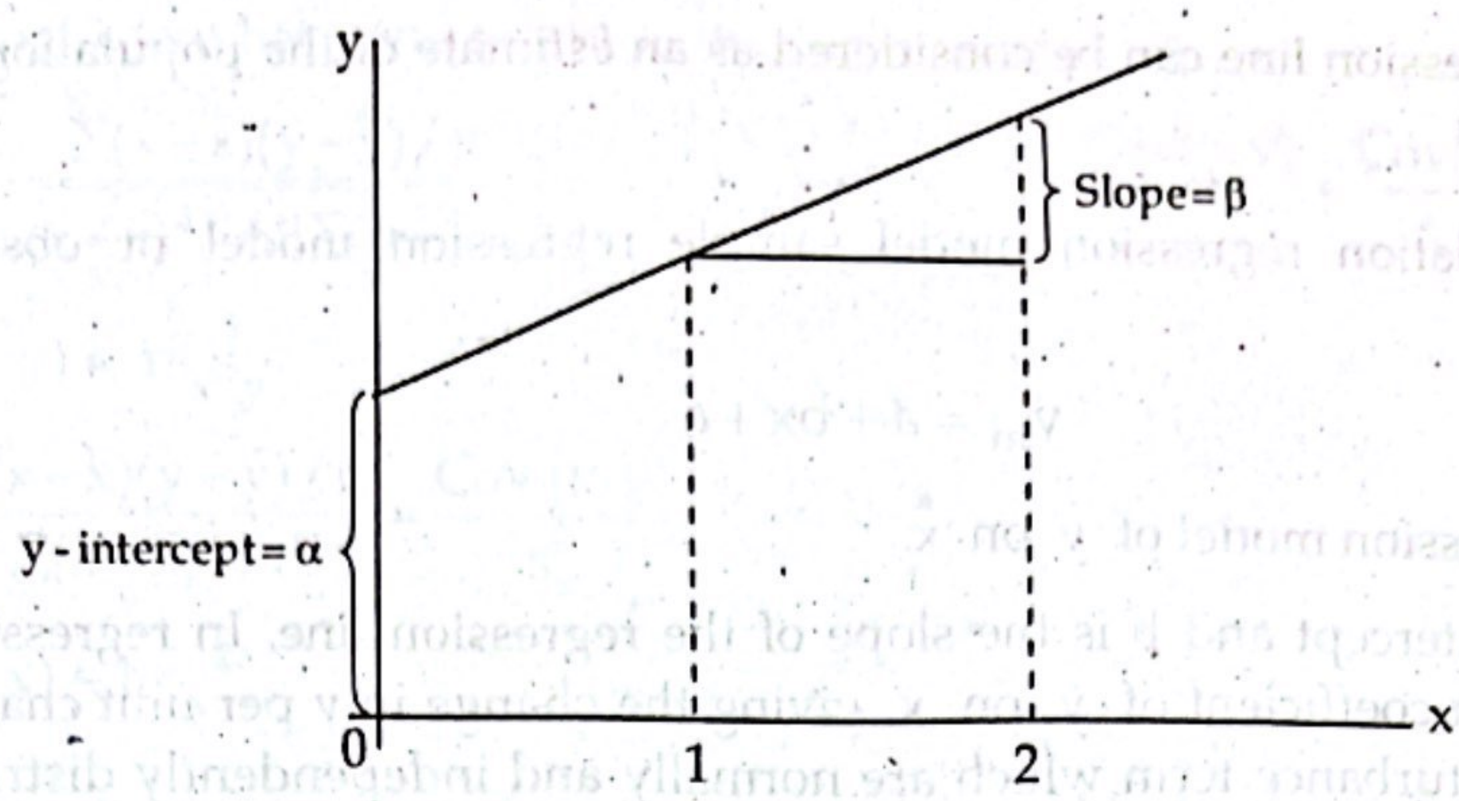

Fig. 12.1

In practice, expenditure of an individual does not depend only on income. There may be other factors such as number of family members and others. Hence in general we can assume a regression model of Y on X as

$$Y_m = \alpha + \beta X + \varepsilon \qquad \text{... (12.2)}$$

Here α and β have the same meaning as mentioned above. ε is the error or disturbance term. Actually, it is the distance between the observed value Y_m and the expected value of Y. The error term ε is normally and independently distributed with zero mean and constant variance σ^2. Under the assumption, the mean value of Y_m is

$$E[Y \mid X] = \alpha + \beta X.$$

To find the best fit of the population regression model or equation we need N pairs of observations of the two variables on the N individuals of the population. Suppose (X_1, Y_1), (X_2, Y_2), ..., (X_N, Y_N) are N pairs of observations. We can find the best fit of the population regression line by obtaining the values of α and β in such a way that the error sum of square i.e. $\sum \varepsilon^2$ is minimum. This is done by the method of least squares, which will be discussed, for sample data.

Remember that the model defined in (12.2) is created for a population of measurements that is generally unknown to us. However, we can use sample information to estimate the values of α and β, which are the coefficients of the line of means, $E[Y \mid X] = \alpha + \beta X$. These estimates are used to form the best-fitted line for a given set of data, called the least squares line or regression line. We review how to calculate the intercept and the slope of this line in the next section.

12.3 Sample Regression Equation and Model

In practice, it is not always possible to get population data. Suppose we have a sample of n pairs of observations say (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) from a bivariate population of interest. The sample regression equation is the best-fitted straight line of y on x is

$$\hat{y} = a + bx \qquad \text{... (12.3)}$$

Here a and b are the estimates of the intercept and slope parameters α and β, respectively.

Scanned with CamScanner

This sample regression line can be considered as an estimate of the population regression line, $Y = \alpha + \beta X$.

Like population regression model sample regression model or observation can be defined as

$$y_m = a + bx + e \qquad \text{... (12.4)}$$

This is the regression model of y on x.

Here a is the intercept and b is the slope of the regression line. In regression analysis b is called the regression coefficient of y on x giving the change in y per unit change of x. e's are random error or disturbance term which are normally and independently distributed with zero mean and constant variance s^2. The simple regression equation is the best fitted straight line in the least -squares sense with the sample data. It is defined as

$$\hat{y} = a + bx \qquad \text{... (12.5)}$$

But the observation y follows the model, $y = a + bx + e$ ... (12.6)

Then the error is $e = y - \hat{y} = (y - a - bx)$, and the error sum of squares is

$$L = \Sigma e^2 = \Sigma(y - a - bx)^2 \qquad \text{... (12.7)}$$

To find the best fitted regression, we have to find the values of a and b in such a way that error sum of squares is minimum. This is done by the method of least squares developed by Legendre. This can be done by taking partial derivatives of L with respect to both a and b, and equate to zero. That is

$$\frac{\partial L}{\partial a} = -2\Sigma(y - a - bx) = 0 \Rightarrow \Sigma y = na + \Sigma x \qquad \text{... (12.8)}$$

$$\frac{\partial L}{\partial b} = -2\Sigma(y - a - bx)x \Rightarrow \Sigma xy = a\Sigma x + b\Sigma x^2 \qquad \text{... (12.9)}$$

Equations (12.8) and (12.9) are called normal equations for finding a and b. By solving the equations (12.8) and (12.9) we get the values of a and b as

$$a = \bar{y} - b\bar{x} \quad \text{and} \quad b = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}.$$

Now by putting the calculated values of a and b in (12.3) we get the best fitted sample regression line of y on x as

$$\hat{y} = a + bx = \bar{y} - b\bar{x} + bx \quad \Rightarrow \hat{y} - \bar{y} = b(x - \bar{x}). \qquad \text{... (12.10)}$$

This regression line is used to estimate or predict the values of y for given values of x. The advantage of using this method is that we do not need the value of the intercept a.

The fitted regression line of y on x can also be obtained with the help of the correlation coefficient between x and y.

Scanned with CamScanner

The correlation coefficient between x and y is

$$r=\frac{\Sigma(x-\bar{x})(y-\bar{y})/n}{\sqrt{\{\Sigma(x-\bar{x})^2/n\}\{\Sigma(y-\bar{y})^2/n\}}}=\frac{Cov(x,y)}{\sqrt{var(x)\,var(y)}}=\frac{Cov(x,y)}{\sqrt{s_x^2\times s_y^2}}=\frac{Cov(x,y)}{s_x s_y}$$

$$\Rightarrow \quad Cov(x, y) = rs_x s_y \qquad \text{... (12.11)}$$

$$b=\frac{\Sigma(x-\bar{x})(y-\bar{y})/n}{\Sigma(x-\bar{x})^2/n}=\frac{Cov.(x,y)}{s_x^2}$$

$$\Rightarrow \quad Cov(x, y) = b\,s_x^{\,2}. \qquad \text{... (12.12)}$$

From (12.11) and (12.12), we have $bs_x^2 = rs_x s_y$. Hence, $b=\frac{rs_y}{s_x}$.

Hence, the best fitted regression line of y on x can be written as

$$\hat{y}=\bar{y}+r\frac{s_y}{s_x}(x-\bar{x}).$$

So the best-fitted regression line of y on x can be written in three different ways as

i) $\hat{y}=a+bx$; by finding the values of a and b by the least squares method.

ii) $\hat{y}=\bar{y}+b(x-\bar{x})$; when the values of b, $\bar{y}$ and $\bar{x}$ are available.

iii) $\hat{y}=\bar{y}+r\frac{s_y}{s_x}(x-\bar{x})$; when the values of r, s_x and s_y are available.

It is to be noted that the formulae for finding a and b are quite easy when the values of and y are measured from their respective means. In that case

$$a=0 \quad \text{and} \quad b=\frac{\Sigma uv}{\Sigma u^2}; \quad \text{where } u=x-\bar{x} \text{ and } v=y-\bar{y}.$$

Then the regression equation of v on u is

$$\hat{v}=bu \qquad \text{... (12.1}$$

By putting the values of u and v in (12.13) we get the required regression equation of on x as

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

In many cases the two variables x and y are inter-dependent. For example, ages of husban and ages of wives are inter-dependent. Here one can use any one as dependent variable. He two regression lines are meaningful. Let us assume that the age of wife x depends on the age husband y. Then the best fifed regression line of x on y in the least squares sense is

$$\hat{x}=c+dy \qquad \text{... (12.14}$$

Here c is the intercept and d is the slope of the line. d is also called the regression coefficien of x on y, giving the change in x per unit change of y. e's are random error which are normally

Scanned with CamScanner

and independently distributed with zero mean and constant variance s^2. The values of c and d are obtained by the following formulae :

$$c=\bar{x}-d\bar{y} \quad \text{and} \quad d=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2}=\frac{\Sigma xy-\frac{(\Sigma x)(\Sigma y)}{n}}{\Sigma y^2-\frac{(\Sigma y)^2}{n}}=\frac{n\Sigma xy-(\Sigma x)(\Sigma y)}{n\Sigma y-(\Sigma y)^2}.$$

By putting the value of c and d in (12.14), we get the best-fitted regression line of x on y as

$$\hat{x}-\bar{x}=d(y-\bar{y}) \qquad \text{... (12.15)}$$

From this regression line, we can estimate or predict the values of x for different given values of y.

Similarly, the best-fitted regression line of x on y can be written in three different ways as

i) $\hat{x}=c+dy$; by finding the values of c and d by the least squares method.

ii) $\hat{x}=\bar{x}+d(y-\bar{y})$; when the values of d, $\bar{y}$ and $\bar{x}$ are available.

iii) $\hat{x}=\bar{x}+r\frac{s_x}{s_y}(y-\bar{y})$; when the values of r, s_x and s_y are available.

Remarks. It is to be noted that the numerators of r, b and d are same. They have the same sign when they are calculated from the same set of data. That is if b is negative, then d and r are also negative. If r is positive, then b and d are also positive.

12.4 Relationship between Correlation Coefficient and Regression Coefficients

There is a mathematical relationship between the correlation coefficient and the regression coefficients. Actually, correlation coefficient is the geometric mean of the regression coefficients.

Theorem 12.4.1 Correlation coefficient is the geometric mean of the regression coefficients.

Proof. Suppose r is the correlation coefficient between x and y, b is the regression coefficient of y on x and d is the regression coefficient of x on y, then we have to prove, $r=\sqrt{c\times d}$.

The coefficient of correlation between x and y is

$$r=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2}}.$$

The regression coefficients of y on x and x on y are

$$b=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} \quad \text{and} \quad d=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2}.$$

Then, $$b\times d=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}\times\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2}=\left[\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2}}\right]^2=r^2.$$

Hence, $r=\sqrt{b\times d}$.

Scanned with CamScanner

Theorem 12.4.2 Regression coefficient is independent of the shift of origin but depends on the change of scale.

Proof. Suppose $b_{y/x}$ is the regression coefficient of y on x. That is

$$b_{y/x} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}.$$

Let us define two new variables u and v by shifting the origin and changing the scale as

$$u = \frac{x-A}{h} \Rightarrow x = A + hu \; ; \quad v = \frac{y-B}{k} \Rightarrow y = B + kv.$$

The regression coefficient of v on u is

$$b_{v/u} = \frac{\Sigma(u-\bar{u})(v-\bar{v})}{\Sigma(u-\bar{u})^2}.$$

Now, $b_{y/x} = \dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$

$$= \frac{\Sigma(A+hu-A-h\bar{u})(B+kv-B-k\bar{v})}{\Sigma(A+hu-A-h\bar{u})^2} = \frac{kh\,\Sigma(u-\bar{u})(v-\bar{v})}{h^2\,\Sigma(u-\bar{u})^2} = \frac{k}{h} b_{v/u}.$$

This shows that regression coefficient depends on h and k but independent of A and B. This means regression coefficient is independent of the shift of origin but depends on change of scale. That is, the value of regression coefficient will remain same if we subtract a constant quantity from all the values of x and another constant quantity from the entire values of y. But regression coefficient will be different if we divide all the values of x and y by two constant quantities.

12.5 Some Important Properties of Regression Coefficient

i) The regression co-efficient measures the average change in dependent variable for a unit change in independent variable.

ii) Regression coefficients are not symmetrical function of x and y. Suppose $b_{y/x}$ and $b_{x/y}$ are regression coefficients of y on x and x on y respectively, then $b_{y/x} \neq b_{x/y}$.

iii) Both the regression coefficients have the same sign.

iv) The correlation coefficient is the geometric mean of two regression coefficients, that is $r_{xy} = \sqrt{b_{x/y}.b_{y/x}}$.

v) The arithmetic mean of the regression coefficients is equal to or greater than the correlation coefficient, that is $\dfrac{b_{x/y}+b_{y/x}}{2} \geq r_{xy}$.

vi) If one of the regression coefficients is greater than one, then the other regression coefficient must be less than one since $r_{xy}^2 = b_{x/y}.b_{y/x} \leq 1$.

vii) The sign of correlation coefficient and the sign of regression coefficients are same, since all the measures depend on the sign of the covariance appearing in the numerator.

Scanned with CamScanner

12.6 Difference between Simple Correlation and Simple Regression

Although correlation and regression both are used to analyze the relationship between two variables and there is an algebraic relationship between two coefficients, but there are differences between these two types of statistical tools with respect to their analysis, characteristics and interpretation. Some of the differences are listed below.

	Simple Correlation	Simple Regression
1	Simple correlation measures the direction and strength of linear relationship between two variables.	Regression measures the effect of independent variable on dependent variable.
2	Correlation does not measure cause and effect relationship between the variables under study.	However, regression analysis measures the cause and effects relationship between the variables. Here, the variable corresponding to cause is taken as independent variable and the variable corresponding to effect is taken as dependent variable.
3	Question of dependent and independent variables do not arise in correlation analysis.	Dependent variable is regressed on the independent variable in regression analysis.
4	Correlation analysis is confined only to the study of linear relationship between the variables, and therefore has limited applications.	Regression analysis has much wider application as it studies linear as well as non-linear relationship between the variables.
5	Correlation coefficient is symmetrical about the variables, that means, $r_{xy} = r_{yx}$.	Regression coefficients are not symmetrical, that is $b_{y/x} \neq b_{x/y}$.
6	The value of the correlation coefficient lies between –1 and + 1.	Regression coefficient can take any real value between $-\alpha$ to $+\alpha$.
7	Correlation coefficient is a pure number. It is a relative measurement.	Regression co-efficient is an absolute measurement. It depends on the units of measurement of the variables.
8	Correlation coefficient is independent of the shift of origin and change of scale.	Regression co-efficient is independent of shift of origin, but depends on change of scale.

12.7 The Coefficient of Determination r^2

The coefficient of determination r^2 is really the square of the correlation coefficient r. But it is a much more precise measure of the strength of the relationship between the two variables and is more precise interpretation because it can be presented as a proportion or as a percentage.

The coefficient of determination r^2 can be defined as the proportion of the variation in the dependent variable y that is explained by the independent variable x, in the regression model. In other words, coefficient of determination is

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\Sigma(\hat{y}-\bar{y})^2}{\Sigma(y-\bar{y})^2}.$$

$\hat{y}$: Best fitted regression line; $\bar{y}$: mean of $\hat{y}$; y: observation and $\bar{y}$: mean of y.

Scanned with CamScanner

Coefficient of determination can also be calculated from the following formula

$$r^2 = \frac{[\Sigma(x-\bar{x})(y-\bar{y})]^2}{\Sigma(x-\bar{x})^2\,\Sigma(y-\bar{y})^2}.$$

It is the square of the simple correlation coefficient. It follows from the definition that the value of r^2 varies from 0 to 1.

A short-cut method for finding sample coefficient of determinate is

$$r^2 = \frac{a\Sigma y + b\Sigma xy - n\bar{y}^2}{\Sigma y^2 - n\bar{y}^2}.$$

- **Interpretation of r^2.** The coefficient of determination is a summary measure that tells us how well the sample regression line fits the observed data. It is a measure of the goodness of fit of a regression.

Usually, the closer the value of r^2 to 1, the better the model, that means, a regression line is better if its r^2 value is closed to 1. On the other hand, if it is closed to zero, it indicates that the regression line does not fit the data well and it fails to predict the future value of y for given value of x. Actually, r^2 measures the proportion or percentage of the total variation in the dependent variable explained by the regression model.

For example, if for a model $y = a + bx + e$, $r^2 = 0.88$, it means 88% variation of the dependent variable is explained by the independent variable and only 12% variation of y remains still unexplained. It means 88% observations i.e.,88% data points should fall within the regression line. In regression analysis r^2 is more meaningful than r.

12.8 Some Examples

Example 12.8.1. Let b be the regression coefficient of Y on X and d be the regression coefficient of X on Y, and r be the correlation coefficient of X and Y. Comment on the following sets of the values of the correlation coefficient and regression coefficients :

i) $b = 1.2$ and $d = 1.4$,
ii) $b = 0.6$ and $d = -0.8$,
iii) $b = -0.2$, $d = -0.8$ and $r = -0.4$,
iv) $b = 0.8$ and $d = 0.4$,
v) $b = 1.36$ and $d = 0.613$.

Ans. (i) We know $r = \sqrt{b \times d}$. Here $r = \sqrt{1.2 \times 1.4} = 1.29$ this is not possible. The value of r must be less than or equal to 1. That is if one of the regression coefficients is greater than one the other must be less than one. Hence the values $b = 1.2$ and $d = 1.4$ are not accepted.

(ii) The values of $b = 0.6$ and $d = -0.8$ are not possible since both the regression coefficients must have the same sign. In other words, it is not possible that one of the regression coefficients is having minus sign and the other plus sign.

(iii) Here $b = -0.2$, $d = -0.8$ and $r = -0.4$. We know $r = \sqrt{(-.2)(-0.8)} = -0.4$. Here the value of r should be minus, since both the regression coefficients are minus. In other words, correlation coefficient and regression coefficients must have the same sign. Hence the values $b = -0.2$, $d = -0.8$, and $r = -0.4$ are accepted.

Scanned with CamScanner

(iv) Here $r = \sqrt{0.8 \times 0.4} = 0.566$ is less than one. Moreover, $(b+d)/2 = (0.8+0.4)/2 = 0.6$ which is greater than the value of $r = 0.566$ In other words, the average of regression coefficients must be greater than or equal to the correlation coefficient. Hence $b = 0.8$, $d = 0.4$ are possible.

(v) $r = \sqrt{1.36 \times 0.613} = 0.91$, which is less than one and have the same sign. One of the regression coefficients is greater than one the other is e less than one. The average of the two regression coefficients is $(1.36 + 0.613)/2 = 0.9865$ which is greater than the value of $r = 0.91$. Hence the values of $b = 1.36$ and $d = 0.613$ are possible.

Example 12.8.2 The following data give the test scores and sales made by nine salesmen during the last year of a big departmental store :

Test Scores	: y	14	19	24	21	26	22	15	20	19
Sales (in lakh Taka)	: x	31	36	48	37	50	45	33	41	39

i) Find the regression equation of test scores on sales.
ii) Find the test score when the sale is Tk. 40 lakh.
iii) Find the regression line of sales on test scores.
iv) Predict the value of sale if the test score is 30.
v) Compute the value of correlation coefficient with the help of regression coefficients and also with the original formula.
vi) Find coefficient of determination and comment.

Solution. (i) The best fitted regression equation of test scores y on sales x is

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

Table for calculation of regression equation

Sales (x)	Test Scores (y)	Xy	x^2	y^2
31	14	434	961	196
36	19	684	1296	361
48	24	1152	2304	576
37	21	777	1369	441
50	26	1300	2500	676
45	2	990	2025	484
33	15	495	1089	225
41	20	820	1681	400
39	19	741	1521	361
$\Sigma x = 360$	$\Sigma y = 180$	$\Sigma xy = 7393$	$\Sigma x^2 = 14746$	$\Sigma y^2 = 3720$

Here $\bar{y} = \frac{180}{9} = 20$, $\bar{x} = \frac{360}{9} = 40$

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{7393 - \frac{180 \times 360}{9}}{14746 - \frac{(360)^2}{9}} = \frac{7393 - 7200}{14746 - 14400} = \frac{193}{346} = 0.56.$$

$$a = \bar{y} - b\bar{x} = 20 - 0.56 \times 40 = 20 - 22.4 = -2.4.$$

Scanned with CamScanner

Hence the required regression equation of test scores on sales is

$$\hat{y} = a + bx = -2.4 + 0.56x.$$

(ii) When $x = 40$, the value of y is $\hat{y} = -2.4 + 0.56 \times 40 = -2.4 + 22.4 = 20.$

The test score is 20 when the sale is Tk. 40 lakh.

(iii) The regression equation of sales x on test scores y is $x = c + dy$.

Here, $$d = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma y^2 - \frac{(\Sigma y)^2}{n}} = \frac{7393 - \frac{180 \times 360}{9}}{3720 - \frac{(180)^2}{9}} = \frac{7393 - 7200}{3720 - 3600} = \frac{193}{120} = 1.61.$$

$$c = \bar{x} - d\bar{y} = 40 - 1.61 \times 20 = 40 - 32.2 = 7.8.$$

Hence the required regression equation of sales on test scores is

$$\hat{x} = c + dy = 7.8 + 1.61y.$$

(iv) When $y = 30$, then x would be Tk. 56.1 lakh.

The predicted sales would be Tk. 56.1 lakh if the test score is 30.

$x = 7.8 + 1.61 \times 30 = 7.8 + 48.3 =$

(v) The formula of correlation coefficient with the help of regression coefficients is

$$r = \sqrt{c \times d} = \sqrt{0.56 \times 1.61} = \sqrt{0.9016} = 0.95.$$

The value of correlation coefficient would be positive since the regression coefficients are positive.

The correlation coefficient from the original formula is

$$r = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\sqrt{\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right\}\left\{\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right\}}} = \frac{193}{\sqrt{346 \times 120}} = \frac{193}{20396} = 0.95.$$

Here the coefficient of determination r^2 is 0.9016.

Comment. 90.16% variations of the sales of the salesmen are explained by their test scores. Hence both the regression lines fit well. Any prediction or estimate made by the regression lines may be accepted.

When the means of x and y are whole numbers. The calculation of regression equation discussed above is quite difficult when the values of x and y are large. The work can be simplified if the all values of x and y are subtracted from their respective means $\bar{x}$ and $\bar{y}$ when they are whole number. Now we shall cite an example.

Example 12.8.3 The following data relate to advertising expenditure (in lakhs of taka and sales (in crores of taka) of a firm :

Advertising Expenditure (in lakh Tk.)	: x	10	12	13	17	18
Sales (in crores of Tk.)	: y	5	6	7	9	13

i) Find the equation of the regression line of sales y on expenditure x.

ii) Predict the sales target for an advertising expenditure of Tk. 20 lakhs.

iii) Find the equation of the regression line of expenditure x on sales y.

Scanned with CamScanner

iv) Predict the advertising expenditure for a sales target Tk. 20 crores.
v) Find the correlation coefficient with the help of regression coefficients.
vi) Compute coefficient of determination and comment.
vii) Compute coefficient of determination and comment.

Solution. Here $\bar{x}=\frac{\Sigma x}{n}=\frac{70}{5}=14$; $\bar{y}=\frac{\Sigma y}{n}=\frac{40}{5}=8$. Since $\bar{x}$ and $\bar{y}$ are whole numbers, we can subtract all the values of x and y from their respective means to simply the calculations. Now let us make a table for calculation of regression equation.

X	u = (x – 14)	u^2	Y	v = (y – 8)	v^2	uv
10	– 4	16	5	– 3	9	12
12	– 2	4	6	– 2	4	4
13	– 1	1	7	– 1	1	1
17	3	9	9	1	1	3
18	4	16	13	5	25	20
$\Sigma x = 70$	$\Sigma u = 0$	$\Sigma u^2 = 46$	$\Sigma y = 40$	$\Sigma v = 0$	$\Sigma v^2 = 40$	$\Sigma uv = 40$

(i) The regression of y on x is: $\hat{y}-\bar{y}=b(x-\bar{x})$. Here, $b=\frac{\Sigma uv}{\Sigma u^2}=\frac{40}{46}=0.87$.

Hence the regression equation of y on x is

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

$$\Rightarrow \hat{y}=\bar{y}+b(x-\bar{x})=8+0.87(x-14)=8+0.87x-12.18=-4.18+0.87x.$$

(ii) When $x=20$, the value of y is: $\hat{y}=-4.18+0.87\times 20=$ Tk. 13.22 crores.

(iii) The best fitted regression line of expenditure x on sales y is

$$\hat{x}-\bar{x}=d(y-\bar{y}).$$

Here d is the regression coefficient of x on y. The value of d is

$$d=\frac{\Sigma uv}{\Sigma v^2}=\frac{40}{40}=1.0.$$

Hence the required regression line of x on y is

$$\hat{x}-\bar{x}=d(y-\bar{y})$$

$$\Rightarrow \hat{x}=\bar{x}+d(y-\bar{y})=14+y-8=6+y.$$

(iv) When $y=20$, then the advertising expenditure of the firm is

$$\hat{x}=6+20=Tk.\ 26\ lakhs.$$

(v) The correlation coefficient between sales and expenditure is

$$r=\sqrt{b\times d}=\sqrt{0.87\times 1}=0.933.$$

The value of correlation coefficient is positive, since b and d are positive. The value of $r=0.933$ means there exists a strong relationship between the advertising expenditures and sales of the firm.

vi) The coefficient of determination is $r^2=0.87$. This shows that 87% variations of the sales of the firm are explained by the advertising expenditures.

Scanned with CamScanner

When the means of x and y are not whole numbers. When the actual means of x and y are not whole numbers, the calculation can also be simplified by subtracting two suitable whole numbers from each value of x and y since the regression coefficients are independent of the shift of origin.

Example 12.8.4 The following data give the ages and blood pressure of 10 women.

Age (in years) : x	56	42	36	47	49	42	60	72	63	55
Blood pressure : y	147	125	118	128	145	140	155	160	149	150

i) Obtain the regression line of y on x.

ii) Estimate the blood pressure of a woman whose age is 50 years.

Solution. Here, $\bar{x}=\frac{\Sigma x}{n}=\frac{522}{10}=52.2$ and $\bar{y}=\frac{\Sigma y}{n}=\frac{1417}{10}=141.7$.

Here the means of x and y are not whole number. So, we can subtract two suitable whole numbers from the values of x and y to make the calculation easy. Here we can take A = 49 and B = 145 as they are the two values of x and y situated in the middle of the two series of x and y. Then the new variables are $u=x-49$ and $v=y-145$.

Then the table of calculation will be as follows :

Calculation of regression equation

Age : x	Blood Pressure : y	u = x – 49	v = y – 145	u^2	v^2	uv
56	147	7	2	49	4	14
42	125	–7	–20	49	400	140
36	118	–13	–27	169	729	351
47	128	–2	–17	4	289	34
49	145	0	0	0	0	0
42	140	–7	–5	49	25	35
60	155	11	10	121	100	110
72	160	23	15	529	225	345
63	149	14	4	196	16	56
55	150	6	5	36	25	30
522	1417	32	–33	1202	1813	1115

(i) The equation of the best-fitted regression line is: $\hat{y}=a+bx$;

where, $a=\bar{y}-b\bar{x}$ and $b=\dfrac{\Sigma uv-\frac{(\Sigma u)(\Sigma v)}{n}}{\Sigma u^2-\frac{(\Sigma u)^2}{n}}$.

Here we have to find b first,

$$b=\frac{1115-\frac{32\times(-33)}{10}}{1202-\frac{(32)^2}{10}}=\frac{1115+105.6}{1202-1024}=\frac{1220.6}{109.6}=1.11,$$

$$a=141.7-1.11\times 52.2=83.758.$$

Hence the best-fitted regression line is $\hat{y}=83.76+1.11x$.

(ii) The blood pressure of the woman whose age 50 is $\hat{y}=83.76+1.11(50)=139.26$.

Scanned with CamScanner

Example 12.8.9. The following summary data relate to the advertising expenditures and sales of a company.

	Advertising Expenditures (x) (lakh Tk.)	Sales (y) (lakh Tk.)
Mean	10	90
Standard deviation	3	12
Correlation coefficient	0.8	

i) Find the regression lines.
ii) Find the likely sales of the company when the advertising expenditure is Tk. 15 lakhs
iii) What should be advertising expenditure if the company wants to obtain a sales target of Tk. 120 lakhs.

Solution. (i) The regression line of y on x is

$$\hat{y}-\bar{y}=r\frac{\sigma_y}{\sigma_x}(x-\bar{x}) \Rightarrow y-90=0.8\times\frac{12}{3}(x-10) \Rightarrow y=90+3.2x-32 \Rightarrow y=58+3.2x.$$

The regression line of x on y is

$$\hat{x}-\bar{x}=r\frac{\sigma_x}{\sigma_y}(y-\bar{y}) \Rightarrow \hat{x}-10=0.8\times\frac{3}{12}(y-90) \Rightarrow \hat{x}=10+0.2y-18 \Rightarrow \hat{x}=-8+0.2y.$$

(ii) The regression equation of sales y on advertising expenditure x is

$$\hat{y}=58+3.2x.$$

When the advertising expenditure x = Tk. 15 lakh, the sales y of the company is

$$\hat{y}=58+3.2\times15=58+48=\text{Tk.}106\text{ lakh.}$$

(iii) The regression equation of advertising expenditure x on sales y is

$$\hat{x}=-8+0.2y.$$

When the sales is Tk. 120 lakhs, the advertising expenditure will be

$$\hat{x}=-8+0.2\times120=-8+24=\text{Tk.}16\text{ lakh.}$$

Example 12.8.10 To study the relationship between expenditure on accommodation x and the expenditure on food y, an inquiry of 50 families gave the following results.

$$\Sigma x=8500 \text{ and } \Sigma y=9600$$
$$\sigma_x=60,\ \sigma_y=20 \text{ and } r=0.6.$$

Estimate the expenditure on food when the expenditure on accommodation is Tk. 200.

Solution. Here first we have to find the regression of food y on accommodation x. The regression of y on x is $y-\bar{y}=r\frac{\sigma_y}{\sigma_x}(x-\bar{x})$.

From the given data, $\bar{x}$ and $\bar{y}$ are $\bar{x}=\frac{8500}{50}=170$ and $\bar{y}=\frac{9600}{50}192.$

The regression equation of y on x is

$$y-\bar{y}=r\frac{\sigma_y}{\sigma_x}(x-\bar{x}) \Rightarrow y-192=0.6\times\frac{20}{60}(x-170) \Rightarrow y=192+0.2x-34$$

Hence, $y=158+0.2x.$

When the expenditure on accommodation x is Tk. 200, the expenditure on food y is

$$y=158+(0.2)(200)=158+40=\text{Tk. }198.$$

Scanned with CamScanner

Example 12.8.11 The following calculations have been made for prices of 12 stocks (X) on the Chittagong Stock Exchange on a certain day along with the volume of sales in thousands of shares (Y). $\Sigma X = 580$, $\Sigma Y = 370$, $\Sigma XY = 11494$, $\Sigma X^2 = 41658$, $\Sigma Y^2 = 17206$.

From these calculations find the regression equation of prices of stocks on the volume of sales of shares.

Solution. The regression equation of stocks X on volume of sales Y is

$$X - \bar{X} = d(Y - \bar{Y})$$

The values of $\bar{X}$, $\bar{Y}$ and d are calculated from the given data as follows.

$$\bar{X} = \frac{\Sigma X}{n} = \frac{580}{12} = 48.33 \quad \text{and} \quad \bar{Y} = \frac{\Sigma Y}{n} = \frac{370}{12} = 30.83.$$

$$d = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{n}}{\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}} = \frac{11494 - \frac{580 \times 370}{12}}{17206 - \frac{(370)^2}{12}} = \frac{11494 - 1788333}{17206 - 1140833} = \frac{-638933}{579767} = -1.102.$$

Therefore, the regression equation is

$$X - 48.33 = (-1.102)(Y - 30.83)$$

$$\Rightarrow \quad X = 48.33 - 1.102Y + 33.98 = 82.31 - 1.102Y.$$

Example 12.8.12 The equations of two regression lines obtained in a correlation analysis are the following.

$$2x + 3y - 8 = 0 \quad \text{and} \quad x + 2y - 5 = 0.$$

Obtain the correlation coefficient and the variance of y given that the variance of x is 12.

Solution. Let us assume that the regression of x on y is

$$2x + 3y - 8 = 0 \qquad \text{... (i)}$$

The regression of y on x is

$$x + 2y - 5 = 0. \qquad \text{... (ii)}$$

From (i) $\quad 2x = 8 - 3y \quad \Rightarrow \quad x = 4 - 1.5y$

Therefore, $d = -1.5$.

From (ii) $\quad 2y = 5 - x \quad \Rightarrow \quad y = 2.5 - 0.5x$

Therefore, $b = -0.5$.

Hence both the values of b and d are reasonable and the assumed regression equations are accepted. We know the correlation coefficient is the geometric mean of the regression coefficients. Therefore, the required correlation coefficient r is

$$r = \sqrt{(-1.5)(-0.5)} = \sqrt{0.75} = -0.867.$$

It is here mentioned that the sign of the correlation coefficient must be negative, since the regression coefficients are negative.

The value of variance of y can be determined from any regression coefficient. The regression coefficient of y on x is

$$b_{y/x} = r\frac{\sigma_y}{\sigma_x}; \quad r = -0.867, \quad \sigma_x = \sqrt{12}, \quad b_{y/x} = -0.5.$$

$$(b_{y/x})^2 = r^2\frac{\sigma_y^2}{\sigma_x^2} \quad \Rightarrow \quad \sigma_y^2 = \frac{(b_{y/x})^2\sigma_x^2}{r^2} = \frac{(-0.5)^2 \times 12}{(0.867)^2} = \frac{0.25 \times 12}{0.75} = 4.$$

Scanned with CamScanner

Remarks. Regression coefficient coefficients and correlation coefficient must have the same sign.

Example 12.8.13 Determine the regression equation of production (in thousand units) as function of capacity utilization (in percentage) from the summarized data given below :

	Average	Standard Deviation
Capacity Utilization (in percentage) : X	84.2	12.5
Production (in thousand Units) : Y	35.6	6.8

Correlation Coefficient between X and Y is 0.68. Estimate the production when the capacity utilization is 75%.

Solution. First we have to find the regression equation of production Y on the capacity utilization X. The regression equation of Y on X is

$$Y-\bar{Y}=r\frac{\sigma_y}{\sigma_x}(X-\bar{X})$$

$$\Rightarrow \quad Y-35.6=0.68\frac{6.8}{12.5}(X-84.2)$$

$$\Rightarrow \quad Y=35.6+0.37X-31.15=4.5+0.37X.$$

The estimated production when the capacity utilization 75% is

$$Y = 4.5 + (0.37)(75) = 4.5 + 27.75 = 32.25 \text{ thousand unit.}$$

Example 12.8.14 In a crop production experiment conducted to study the relationship between yield per acre of a crop (Y) and the dose of NPK fertilizer (X), following results were obtained.

Sample size, $n=20$, $\bar{X}=12.8, \bar{Y}=130.7, \sigma_x^2=3.53, \sigma_y^2=4.93, \text{cov}(X,Y)=3.42.$

Assuming linear relationship, (i) Develop the least squares regression line, (ii) Compute the standard error of the estimator, and (iii) Estimate yield corresponding to $X=12$.

Solution. (i) The regression line of Y on X is $Y=a+bx$.

The least squares estimates a and b are

$$a=\bar{Y}-b\bar{X} \quad \text{and} \quad b=\frac{\text{Cov.}(X,Y)}{\text{Var.}(X)}.$$

Here, $b=\dfrac{Cov.X,Y)}{Var(X)}=\dfrac{3.42}{3.53}=0.97$ and $a=130.7-0.97\times12.8=130.7-12.40=118.28.$

Hence, the least squares regression line of Y on X is: $Y=118.28+0.97X.$

(ii) The standard error of the estimate is: $S.E.=\sigma_y\sqrt{1-r^2}$.

Now, $r=\dfrac{Cov.(X,Y)}{\sqrt{\sigma_x^2\sigma_y^2}}=\dfrac{3.42}{\sqrt{3.53\times4.93}}=\dfrac{3.42}{\sqrt{17.39}}=\dfrac{3.42}{4.17}=0.82.$

Hence, $S.E.=\sigma_y\sqrt{1-r^2}=2.22\sqrt{1-(0.82)^2}=2.22\sqrt{.33}=1.27;$ where $\sigma_y=2.22.$

(iii) Estimated yield corresponding to $X=12$ is

$$Y=118.28+0.97(12)=118.28+11.64=129.92$$

Example 12.8.15 The following results relate to the regression analysis of an experiment. The two regression lines are $5Y=4X+33$ and $20X=9Y+107$. The variance of Y is $\sigma_y^2=16$.

Scanned with CamScanner

Calculate: i) The mean vales of the two series X and Y,

ii) The correlation coefficient of X and Y, and

iii) The standard deviation of X.

Solution. (i) The mean values of X and Y are given by the point of intersection of two regression equations. Therefore, in order to calculate the mean values of X and Y, we solve the following simultaneous equations

$$5Y = 4X + 33 \quad \ldots (1)$$

$$20X = 9Y + 107 \quad \ldots (2)$$

Multiply equation (1) by 5 and subtract it from (2), we have

$$\begin{array}{r} 20X - 25Y = -165 \\ 20X - 9Y = 107 \\ \hline -16Y = -272 \end{array}$$

$$\Rightarrow \quad Y = 272/16 = 17.$$

Substituting the value of Y in equation (2), we get

$$20X = 9(17) + 107 = 260 \quad \Rightarrow X = 260/20 = 13.$$

Therefore, the mean value of X is 13 and the mean value of Y is 17.

(ii) From the equation (1), we get $Y = 6.6 + 0.8x$,

Then the regression coefficient of Y on X is: $b = 0.8$.

Now from the equation (2) we get $X = 5.35 + 0.45y$

The regression coefficient of X on Y is $d = 0.45$ Both the regression coefficients are accepted. We know that the correlation coefficient is the geometric mean of the regression coefficients. Therefore the correlation coefficient is

$$r = \sqrt{b \times d} = \sqrt{0.8 \times 0.45} = \sqrt{0.36} = 0.6.$$

(iii) The standard deviation of X is obtained from the relation,

$$b = r\frac{\sigma_y}{\sigma_x}; \qquad \text{where } r = 0.6,\ \sigma_y = 4,\ b = 0.8$$

$$\Rightarrow \quad \sigma_x = \frac{0.6 \times 4}{0.8} = 3.$$

Hence the standard deviation of X is 3.

Example 12.8.16 The Personal Manager of a large industrial unit is interested to find a measure that can be used to fix the yearly wage of skilled workers. On an experimental basis, he compiled data on the length of service and their yearly wages in thousand taka from a group of 10 randomly selected workers.

Length of service (in years) : X	4	3	12	10	6	8	5	9	7	11
Yearly Wages : Y	7	6	16	14	10	13	9	10	11	14

Scanned with CamScanner

i) Develop a best fitted regression equation of wage (Y) on the length of service (X),

ii) On the basis of (i), what would be the starting salary to a skill worker who has already served 13 years?

iii) Is it necessary to consider other factors (in addition, to length of service) for fixing the wages of a skilled workers?

Solution. (i) The required regression equation is the regression of Y on X which is

$$Y = a + bx.$$

The values of a and b obtained by the least squares method give best fitted regression equation of wage on the length of service. The least squares values of a and b are obtained from the following formulae.

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

Calculation of regression equation

X	Y	X^2	Y^2	XY
4	7	16	49	28
3	6	9	36	18
12	16	144	256	192
10	14	100	196	140
6	10	36	100	60
8	13	64	169	104
5	9	25	81	45
9	10	81	100	90
7	11	49	121	77
11	14	121	196	154
75	110	645	1304	908

$$\overline{X} = \frac{\Sigma X}{10} = \frac{75}{10} = 7.5 \qquad \overline{Y} = \frac{\Sigma Y}{10} = \frac{110}{10} = 11.$$

The value of b is

$$b = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{n}}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}} = \frac{908 - \frac{75 \times 110}{10}}{645 - \frac{(75)^2}{10}} = \frac{908 - 825}{645 - 5625} = \frac{83}{82.5} = 1.006$$

$$\therefore \quad a = \overline{Y} - b\overline{X} = 11 - (1.006)(75) = 11 - 7.545 = 3.455$$

Thus the best fitted regression line equation of wage (Y) on length of service (X) is

$$Y = 3.455 + 1.006X.$$

(ii) For a skill worker with 13 years of service, the starting wages would be

$$Y = 3.455 + 1.006(13) = 3.455 + 13.078 = \text{Tk. } 16.533 \text{ thousands.}$$

(iii) We have to find the value of coefficient of determination. r^2 is called the coefficient of determination. The value of r is

Scanned with CamScanner

$$r=\frac{\Sigma XY-\frac{\Sigma X\Sigma Y}{n}}{\sqrt{\left\{\Sigma X^2-\frac{(\Sigma X)^2}{n}\right\}\left\{\Sigma Y^2-\frac{(\Sigma Y)^2}{n}\right\}}}=\frac{908-\frac{75\times 110}{10}}{\sqrt{\left(645-\frac{(75)^2}{10}\right)\left(1304-\frac{(110)^2}{10}\right)}}$$

$$=\frac{908-825}{\sqrt{(645-562.5)(1304-1210)}}=\frac{83}{\sqrt{82.5\times 94}}=\frac{83}{88.06}=0.94.$$

The coefficient of determination is $r^2=0.88$.

The value of $r^2=0.88$ indicates that 88% variation of the wage is explained by the length of the service. One can include some more factors such as education or efficiency of the workers to increase the value of r^2 to some more extend.

Example 12.8.17 You are given the following information.

	x-Series	y-Series
No. of items	15	15
Average	25	18
Sum of squares of deviation from mean	136	138

Sum of the product of deviation of x and y-series from their respective means is 122.

Requirements. (i) Compute coefficient of correlation between x and y, (ii) Find two regression coefficients, and (iii) Calculate the geometric mean of these two regression coefficients and comment on the result. *C.U. Acct. 2012*

Solution. Here we have the following summary data.

$n=15,\ \bar{x}=25,\ \bar{y}=18,$ sum of squares of $x=\Sigma(x-\bar{x})^2=136,$

sum of squares of $y=\Sigma(y-\bar{y})^2=138$ and $\Sigma(x-\bar{x})(y-\bar{y})=122.$

Then the correlation coefficient between x and y is

$$r=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\,\Sigma(y-\bar{y})^2}}=\frac{122}{\sqrt{136\times 138}}=\frac{122}{136996}=0.89.$$

The regression line of y on x is: $y-\bar{y}=b(x-\bar{x})$

The value of b is obtained by the formula, $b=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}=\frac{122}{136}=0.897.$

Hence the regression line of y on x is

$$y-18=(0.897)(x-25)$$

$$\Rightarrow\quad y=18+0.897x-22.43=-4.43+0.897x.$$

The regression line of x on y is: $x-\bar{x}=d(y-\bar{y})$.

The value of d is obtained by the formula: $d=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2}=\frac{122}{138}=0.884$

Scanned with CamScanner

Hence the regression line of x on y is

$$x-25=(0.884)(y-18)$$

$$\Rightarrow \quad x=25+0.884y-15.91=9.087+0.884y.$$

We know correlation coefficient is the geometric mean of the regression coefficients.

Hence, $r=\sqrt{b\times d}=\sqrt{(0.897)(0.884)}=0.89.$

The value of r is same for both the methods which is supposed to be.

Example 12.8.18 A survey firm studying the relation between kilowatt-hours (thousands) use and the number of rooms in a private family residence. A random sample of 10 homes yield the following :

Number of rooms	4	5	6	5	7	4	3	6	5	4
Kilowatt-hours(thousands)	8	7	9	5	8	6	8	10	7	4

C.U. Acct. 2011

i) Determine two regression equations.

ii) Determine the expected number of kilowatt-hours for an eight-room house.

Solution. (i) Let x be the number of rooms and y be the kilowatt hours used. Then the requir regression equations are

$$y=a+bx \quad \text{and} \quad x=c+dy.$$

Calculation of regression equations

x	y	x^2	y^2	xy
4	8	16	64	32
5	7	25	49	35
6	9	36	81	54
5	5	25	25	25
7	8	49	64	56
4	6	16	36	24
3	8	9	64	24
6	10	36	100	60
5	7	25	49	35
4	4	16	16	16
49	72	253	548	361

$$\bar{x}=\frac{\Sigma x}{n}=\frac{49}{10}=4.9; \qquad \bar{y}=\frac{72}{10}=7.2.$$

The value of b is

$$b=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}=\frac{\Sigma xy-\frac{\Sigma x\Sigma y}{n}}{\Sigma x^2-\frac{(\Sigma x)^2}{n}}=\frac{361-\frac{49\times 72}{10}}{253-\frac{(49)^2}{10}}=\frac{361-3528}{253-2401}=\frac{8.2}{12.9}=0.636$$

Scanned with CamScanner

$$a = \bar{y} - b\bar{x} = 7.2 - (0.636)(4.9) = 7.2 - 3.12 = 4.08.$$

Hence the regression equation of y on x is: $y = 4.08 + 0.636x$.

The value of the regression coefficient of x on y is

$$d = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2} = \frac{122}{138} = 0.884.$$

The value of c is: $c = \bar{x} - d\bar{y} = 4.9 - (0.884)(7.2) = 4.9 - 6.36 = -1.46$

Hence the regression line of x on y is: $x = c + dy = -1.46 + (0.884)y$.

(ii) For an eight room house, the expected number of kilowatt hours used is

$$Y = 4.08 + (0.636)(8) = 4.08 + 5.04 = 9.12.$$

Example 12.8.19. The following data refer to advertising expenditure and sales in lakh Tk. of a company for last five years.

Advertising Exp.	10	12	15	23	20
Sales	14	17	23	25	21

Requirements

i) Predict the sales when the amount of advertisement exp. is Tk. 25 lakh.

ii) Predict the advertisement expenses to make sale of Tk. 30 lakh. *C.U. Acct. 2011*

Solution. Let x be the advertising expenditures and y be the sales of the company for last 5 years. To predict the sales and the advertisement expenses of the company we have to find both the regression equations. Then the required regression equations are

$$y = a + bx \quad \text{and} \quad x = c + dy.$$

Calculation of regression equations

x	y	x^2	y^2	xy
10	14	100	196	140
12	17	144	289	204
15	23	225	529	345
23	25	529	625	575
20	21	400	441	420
80	100	1398	2080	1684

$$\bar{x} = \frac{\Sigma x}{n} = \frac{80}{5} = 16; \quad \bar{y} = \frac{100}{5} = 20.$$

The value of b is

$$b = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}} = \frac{1684 - \frac{80\times 100}{5}}{1398 - \frac{(80)^2}{5}} = \frac{1684-1600}{1398-1280} = \frac{84}{118} = 0.71,$$

$$a = \bar{y} - b\bar{x} = 20 - (16)(0.71) = 20 - 11.36 = 8.64.$$

Scanned with CamScanner

Hence the regression equation of y on x is: $\hat{y} = 8.64 + 0.71x$.

The value of the regression coefficient of x on y is

$$d = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2} = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\Sigma y^2 - \frac{(\Sigma y)^2}{n}} = \frac{84}{2080 - \frac{(100)^2}{5}} = \frac{84}{80} = 1.05.$$

The value of c is: $c = \bar{x} - d\bar{y} = 16 - (105)(20) = 16 - 21 = -5.$

Hence the regression line of x on y is: $\hat{x} = c + dy = -5 + (1.05)y$.

(i) When the advertisement expenditure Tk.25lakh, the predicted sales of the company is
$\hat{y} = 8.64 + (0.71)(25) = 8.64 + 17.75 = Tk.\ 26.39\ lakh.$

(ii) When the sales of the company Tk.30 lakh, the predicted advertisement expenditure is
$\hat{x} = -5.0 + (1.05)(30) = -5.0 + 31.5 = Tk.\ 26.5\ lakh.$

Example 12.8.20 The following data relates to the ages of husbands and wives in year of 10 couples.

Ages of husbands	25	28	30	32	35	36	38	39	42	45
Ages of wives	20	26	29	30	25	10	26	34	35	45

i) Obtain two regression equations.

ii) Determine the most likely age of husbands for age of wife 25 years.

iii) Determine the most likely age of wife for age of husband 0 years. *C.U. Acct. 2009*

Solution. (i) Let x be the age of husband and y be the age of wife. Then the two regression equations are

$$y = a + bx \quad \text{and} \quad x = c + dy.$$

Calculation of regression lines

X	y	u = x – 35	v = y – 30	u^2	v^2	uv
25	20	–10	–10	100	100	100
28	26	–7	–4	49	16	28
30	29	–5	–1	25	1	5
32	30	–3	0	9	0	0
35	25	0	–5	0	25	0
36	10	1	–20	1	400	–20
38	26	3	–4	9	16	–12
39	34	4	4	16	16	16
42	35	7	5	49	25	35
45	45	10	15	100	225	150
		0	20	358	824	302

$\bar{u} = 0$ or $\bar{x} = 35$, $\bar{v} = \frac{-20}{10} = -2$ or $\bar{y} = 30 - 2 = 28$.

Scanned with CamScanner

The value of b is

$$b=\frac{\Sigma uv-\frac{\Sigma u\Sigma v}{n}}{\Sigma u^2-\frac{(\Sigma u)^2}{n}}=\frac{302-\frac{0\times 20}{10}}{358-0}=\frac{302}{258}=1.17.$$

$$a=\bar{y}-b\bar{x}=28-(1.17)(35)=28-41.78=-1297.$$

Hence the regression equation of the age of wife y on the age of husband x is

$$y=-1297+1.17x.$$

The value of the regression coefficient of x on y is

$$d=\frac{\Sigma uv-\frac{\Sigma u\Sigma v}{n}}{\Sigma v^2-\frac{(\Sigma v)^2}{n}}=\frac{302}{824-\frac{400}{10}}=\frac{302}{824-40}=\frac{302}{784}=0.39.$$

The value of c is: $c=\bar{x}-d\bar{y}=35-(0.39)(28)=35-10.79=24.21.$

Hence the regression line of the age of husband x on the age of wife y is

$$x=c+dy=-24.21+(0.39)y.$$

(ii) When the age of wife 25 years, then the most likely age of husband is

$$x=24.21+(0.39)(25)=24.21+9.75=33.96 \text{ years.}$$

(iii) When the age of husband 30 years, the most likely age of wife is

$$y=-1297+(1.17)(30)=-1297+35.1=22.13 \text{ years.}$$

Example 12.8.21 The National Highway Association is studying the relationship between the number of bidders on a highway project and the winning (lowest) bid for the project. Of particular interest, is whether the number of bidders increases or decreases the amount of the winning bid, a sample of 10 projects revealed the following information.

Project	1	2	3	4	5	6	7	8	9	10
Number of Bidders	10	12	16	20	9	11	8	13	15	18
Winning Bid(Tk. Crore)	5.1	8	9.7	7.8	8.3	7.8	6.9	7.3	9.1	8.2

i) Determine the regression equations. Interpret the equation. Do more bidders tend to increase or decrease the amount of winning bid?

ii) Estimate the amount of winning bid if there were 8 bidders. *C.U. Acct. 2012*

Solution. Let x be the number of bidders and y be the amount of winning bid. The possible regression equations are

Amount of winning bid on the number bidders x i.e.

$$y=a+bx.$$

(i) The number of winning bidders x on the amount of winning bid y i.e.

Scanned with CamScanner

$$x = c + dy$$

Calculation of regression equations

x	y	x^2	y^2	xy
10	5.1	100	26.01	51
12	8	144	64	96
16	9.7	256	94.09	155.2
20	7.8	400	60.84	156
9	8.3	81	68.89	74.7
11	7.8	121	60.84	85.8
8	6.9	64	47.61	55.2
13	7.3	169	53.29	94.9
15	9.1	225	82.81	136.5
18	8.2	324	67.24	147.6
132	78.2	1884	625.62	1052.9

$$\bar{x} = \frac{\Sigma x}{n} = \frac{132}{10} = 13.2; \qquad \bar{y} = \frac{78.2}{10} = 7.82$$

The value of b is

$$b = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma 3x^2 - \frac{(\Sigma x)^2}{n}} = \frac{10529 - \frac{132 \times 78.2}{10}}{1884 - \frac{(132)^2}{10}} = \frac{10529 - 103224}{1884 - 17424} = \frac{20.66}{141.6} = 0.146.$$

$$a = \bar{y} - b\bar{x} = 7.82 - (0.146)(13.2) = 7.82 - 1.926 = 5.89.$$

Hence the regression equation of y on x is: $y = 5.89 + 0.146x$.

More bidders will slightly increase the amount of winning bid since the regression coefficient of winning bid y on the number of bidders x is positive (0.147) but small.

The value of the regression coefficient of x on y is

$$d = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma y^2 - \frac{(\Sigma y)^2}{n}} = \frac{20.66}{625.62 - \frac{(78.2)^2}{10}} = \frac{20.66}{625.62 - 611.524} = \frac{20.66}{14.096} = 1.47.$$

The value of c is: $c = \bar{x} - d\bar{y} = 13.2 - (1.47)(7.82) = 13.2 - 11.46 = 1.74$.

Hence the regression line of x on y is: $x = c + dy = -1.74 + (1.47)y$.

(ii) The estimated value of the amount of winning when the number of bidders 8 is

$$Y = 5.89 + (0.146)(8) = 5.89 + 1.168 = \text{Tk. } 7.1 \text{ crore.}$$

Example 12.8.22 Compute the appropriate regression equation for the following data.

X (independent variable)	2	4	5	6	8	11
Y (Dependent variable)	18	12	10	8	7	5

N.U. BB. 2011

Scanned with CamScanner

Solution. The appropriate regression equation of y on x is: $y = a + bx$.

The formula for finding the values of a and b are

$$a = \bar{y} - b\bar{x} \quad \text{and} \quad b = \frac{\Sigma xy - (\Sigma x \Sigma y)/n}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}.$$

Computation table for finding the values of a and b

y	y	x^2	xy
2	18	4	36
4	12	16	48
5	10	25	50
6	8	36	48
8	7	64	56
11	5	121	55
36	60	266	293

$$\bar{x} = \frac{\Sigma x}{n} = \frac{36}{6} = 6; \quad \bar{y} = \frac{\Sigma y}{n} = \frac{60}{6} = 10$$

$$b = \frac{\Sigma xy - (\Sigma x \Sigma y)/n}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}} = \frac{293 - \frac{36 \times 60}{6}}{266 - \frac{(36)^2}{6}} = \frac{293 - 360}{266 - 216} = \frac{-67}{50} = -1.34.$$

$$a = \bar{y} - b\bar{x} = 10 - (-1.34)(6) = 10 + 8.04 = 18.04.$$

The appropriate regression is: $y = 18.04 - 1.34x$.

Example 12.8.23 Given the advertising expenses and sales related data for a company:

	Advertising expenses (Lakh Tk.) : x	Sales (Lakh Tk.) : y
Mean	20	120
Standard deviation	5	25
Correlation Coefficient	0.8	

i) Calculate the both regression equations.

ii) If the advertise expenses Tk. 25 lakh what will be the estimated sales. *N.U. BBS 2010*

Solution. (i) The regression equation of y on x is

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

We have, $\bar{x} = 20$, $\bar{y} = 120$, $\sigma_x = 5$, $\sigma_y = 25$ and $r = 0.8$. Then

$$y = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) = 120 + (0.8)\frac{25}{5}(x - 20) = 120 + 4x - 80 = 40 + 4x.$$

The regression equation of x on y is

$$x = \bar{x} + r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) = 20 + (0.8)\frac{5}{25}(y - 120) = 20 + 0.16y - 19.2 = 0.8 + 0.16y.$$

(ii) When the advertising expenses Tk. 25 lakh, then the sales of the company is

$$Y = 40 + 4(25) = \text{Tk. } 140 \text{ lakh.}$$

Scanned with CamScanner

Example 12.8.24 Calculate two regression equations and the coefficient of correlation from the data given below.

x	1	2	3	4	5	6	7	8	9
y	9	8	10	12	11	13	14	16	15

N.U. BBS.-2009

Solution. The regression equation of y on x is: $y = a + bx$.

The values of a and b are calculated by the following formulae.

$$a = \bar{y} - b\bar{x} \quad \text{and} \quad b = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}.$$

The regression equation of x on y is: $x = c + dx$.

The values of c and d are calculated by the following formulae.

$$c = \bar{x} - d\bar{y} \quad \text{and} \quad d = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma y^2 - \frac{(\Sigma y)^2}{n}}.$$

Calculation for finding regression equations

x	Y	x^2	y^2	Xy
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	11	25	121	55
6	13	36	169	78
7	14	49	196	98
8	16	64	256	128
9	15	81	225	135
45	108	285	1356	597

$$\bar{x} = \frac{45}{9} = 5, \quad \bar{y} = \frac{108}{9} = 12.$$

$$b = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}} = \frac{597 - \frac{45 \times 108}{9}}{285 - \frac{(45)^2}{9}} = \frac{597 - 540}{285 - 225} = \frac{57}{60} = 0.95.$$

$$a = \bar{y} - b\bar{x} = 12 - (0.95)(5) = 12 - 4.75 = 7.25.$$

Hence the regression equation of y on x is: $y = 7.25 + 0.95x$.

$$d = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma y^2 - \frac{(\Sigma y)^2}{n}} = \frac{597 - 540}{1356 - \frac{(108)^2}{9}} = \frac{57}{1356 - 1296} = \frac{57}{60} = 0.95.$$

$$c = \bar{x} - d\bar{y} = 5 - (0.95)(12) = 5 - 11.4 = -6.4.$$

Scanned with CamScanner

Hence regression of x on y is: $x = -6.4 + 0.95y$.

We know the correlation coefficient between x and y is the geometric mean of the regression coefficients. Hence the correlation coefficient between x and y is

$$r = \sqrt{b \times d} = \sqrt{(0.95 \times 0.95} = 0.95.$$

Example 12.8.25. From the following bi-variate data.

x	1	5	3	2	1	1
y	6	1	0	0	1	2

i) Fit a regression line of y on x and hence predict y if $x = 10$

ii) Fit a regression line of x on y and hence estimate x when $x = 2.5$. *N.U. BBS. 2006*

Solution. (i) The regression equation of y on x is: $y = a + bx$.

The values of a and b are calculated by the following formulae.

$$a = \bar{y} - b\bar{x} \quad \text{and} \quad b = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}.$$

Calculation for finding regression equations

x	y	x^2	y^2	Xy
1	6	1	36	6
5	1	25	1	5
3	0	9	0	0
2	0	4	0	0
1	1	1	1	1
1	2	1	4	2
13	10	41	42	14

$$\bar{x} = \frac{13}{6} = 2.17, \quad \bar{y} = \frac{10}{6} = 1.67.$$

$$b = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}} = \frac{14 - \frac{13 \times 10}{6}}{41 - \frac{(13)^2}{6}} = \frac{14 - 21.67}{41 - 28.17} = \frac{-7.67}{12.83} = -0.60.$$

$$a = \bar{y} - b\bar{x} = 1.67 - (-0.60)(2.17) = 1.67 + 1.30 = 2.97.$$

Hence the regression equation of y on x is: $y = 2.97 - 0.60x$.

The predicted value of y when $x = 10$ is

$$y = 2.97 - (0.6)(10) = 2.97 - 6.0 = -3.03.$$

(ii) The regression equation of x on y is: $x = c + dx$.

The values of c and d are calculated by the following formulae.

$$c = \bar{x} - d\bar{y} \quad \text{and} \quad d = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma y^2 - \frac{(\Sigma y)^2}{n}}.$$

Scanned with CamScanner

Now, $d = \dfrac{\Sigma xy - \dfrac{\Sigma x \Sigma y}{n}}{\Sigma y^2 - \dfrac{(\Sigma y)^2}{n}} = \dfrac{-7.67}{42 - \dfrac{(10)^2}{6}} = \dfrac{-7.67}{42 - 16.67} = \dfrac{-7.67}{25.33} = -0.303.$

$c = \bar{x} - d\bar{y} = 2.17 - (-0.303)(1.6) = 2.17 + 0.506 = 2.68.$

Hence regression of x on y is: $x = -2.68 - (0.303)y$

The estimated value of x when $y = 2.5$ is

$$x = 2.68 - (0.303)(2.5) = 2.68 - 0.76 = 1.92.$$

Example 12.8.26 In a partially destroyed laboratory of an analysis of correlation data, the following results are legible; the variance of x equals to 9. The regression equations are

$$8x - 10y + 66 = 0 \quad \text{and} \quad 40x - 18y = 214.$$

Find on the basis of the above information (i) The average values of x and y; (ii) Correlation coefficient between x and y; and (iii) Standard deviation of y. *N.U. BBS-2005*

Solution. (i) Since the two regression lines intersect at the point $(\bar{x}, \bar{y})$, the solution of the two regression equations will give the values of $\bar{x}$ and $\bar{y}$.

We have, $8x - 10y + 66 = 0$ ... (i) and $40x - 18y = 214$ ... (ii)

Now Multiplying (i) by 5 and then subtracting it from (ii); we have

$$\begin{array}{r} 40x - 50y = -330 \\ 40x - 18y = 214 \\ \hline -32y = -544 \end{array}$$

$y = 17$. Hence the value of $\bar{y} = 17$

Now by putting the value of y in (i) we have

$$8x - 10(17) = -66 \quad \Rightarrow 8x = -66 + 170 = 104 \quad \Rightarrow x = 13.$$

Hence the value $\bar{x} = 13$.

(ii) For finding the value of r, we have to determine the regression coefficients. Since we don't know which equation is regression of y on x and which is of x on y. We assume that the first equation is the regression line of y on x.

Then, $8x - 10y + 66 = 0 \Rightarrow y = 0.8x + 6.6.$

Here the value of the regression coefficient b is 0.8. Suppose the second equation is the regression line of x on y.

Then, $40x - 18y = 214 \Rightarrow x = 5.35 + 0.45y.$

Hence the value of the regression coefficient d is 0.45. Both the regression coefficients are less than one, so the values are accepted. The correlation coefficient between x and y is the geometric mean of the regression coefficient.

Hence the correlation coefficient is $r^2 = 0.8 \times 0.45 = 0.36 \Rightarrow r = 0.6$.

(iii) The value of standard deviation of Y can be determined from any regression coefficient. The regression coefficient of Y on X is

$$b = r\frac{\sigma_y}{\sigma_x} \quad \Rightarrow \sigma_y = \frac{b \times \sigma_x}{r} = \frac{0.8 \times 3}{0.6} = \frac{2.4}{0.6} = 4.$$

Scanned with CamScanner

12.9 Testing the Usefulness of the Linear Regression Model

12.9.1 The Standard error of estimate of the simple regression equation. The standard error of estimate measures the reliability of the estimating equation. It measures the variability or scatter of the observed dependent variable y around the regression line $\hat{y}$. We can compare this measure with the standard deviation. The formula for standard error of estimate is

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

where y = values of dependent variable; $\hat{y}$ = estimated value of the dependent variable y from the estimating equation; and n = number of data points used to fit the regression line. Here it is noted that sum of the squared deviations is divided by $n-2$ not by n since 2 degrees of freedom have been lost as a and b are obtained from a sample data points in estimating the regression line. To calculate the value of the standard error we must compute $\hat{y}$ for every value of y which needs lot of calculation. Fortunately, we have an alternative short cut formula for finding the standard error of estimate which is given below :

$$S_e = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n-2}}$$

where x = values of the independent variable; y = values of the dependent variable.; a = Y-intercept; b = slope of the estimating equation; and n = number of data points. This equation for standard error is short cut because all the calculations except Σy^2 are needed for finding a and b in estimating equation.

The standard error of estimate can also be calculated by the following formula.

$$S_e = S_x\sqrt{\Sigma(x-\bar{x})^2/n-1}.$$

- **Interpreting the standard error of estimate.** Like standard deviation, the larger the standard error of estimate, the greater the scattering of points around the regression line. If $S_e = 0$, we expect the estimating equation to be perfect estimator of the dependent variable. In that case, all the data points would lie directly on the regression, and no points would be scattered around it.

F test is used to test the goodness of fit of the regression equation and t tests are used to test significance of the regression parameters α and β. At this moment we need the following assumptions.

1) The observed values for y are normally distributed around each estimated value of $\hat{y}$.

2) The variance of the distributions around each possible value of y is the same.

12.9.2 Approximate Prediction Intervals. Considering normality assumption for large samples, that is $n > 30$, 68% observations of the dependent variable lie in the interval $\hat{y} \pm 1S_e$, 95.5% and 99.7% observations lie in the intervals $\hat{y} \pm 1S_e$ and $\hat{y} \pm 1S_e$ respectively. For small samples t distribution is used to find the prediction intervals.

12.9.3 Inference about the regression as a whole (Using F test). Given any simple regression, it is natural to ask whether the value of r^2 really indicates that the independent

Scanned with CamScanner

variable x explain dependent variable y, or might have happened just by chance. Or is the regression as a whole significant? Our null hypothesis in this case is

$H_0 : \beta = 0$ ← Null Hypothesis: y does not depend on x.

$H_0 : \beta \neq 0$ ← Alternative Hypothesis: y depends on x.

It says that the total variation in y can be broken down into two parts, the explained part and the unexplained part. Suppose y is an observed value of the dependent variable as shown in Fig. 12.9.1, $\hat{y}$ is an estimated value of y and $\bar{y}$ is the mean of y. Then the total deviation of y from its mean $\bar{y}$ would be $(y - \bar{y})$. The explained part of the total deviation is $(\hat{y} - \bar{y})$ and the unexplained part is $(y - \hat{y})$. Now consider a set of observations of the dependent variable y instead of only one value. The total variation in y is the sum of the squared total deviation $\Sigma(y - \bar{y})^2$ is equal to the sum of the squared explained deviations $\Sigma(\hat{y} - \bar{y})^2$ and the sum of the squared unexplained deviations $\Sigma(y - \hat{y})^2$. That is

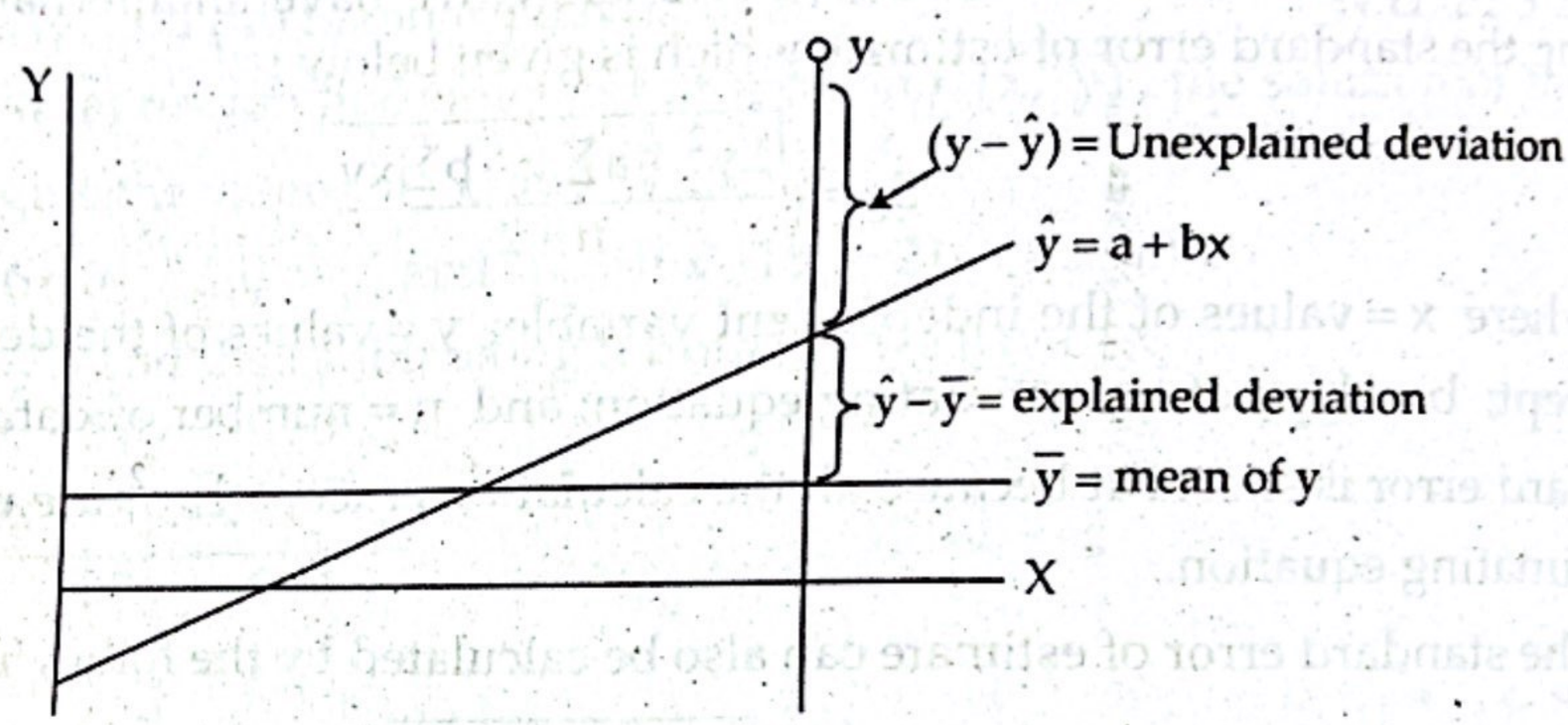

Fig. 12.9.1.

Total sum of Squares (SST) = Regression sum of squares (SSR) + Error sum of squares (SSE). Symbolically, we can write as

$$\Sigma(y - \bar{y})^2 = \Sigma(\hat{y} - \bar{y})^2 + \Sigma(y - \hat{y})^2.$$

Each of this sum of squares has an associated number of degrees of freedom. SST has $n-1$ degrees of freedom. One degree of freedom is less since $n-1$ observations are independent as the sample mean is fixed. SSR has one degree of freedom since one independent variable is being used to explain Y. Finally, SSE has $n-2$ degrees of freedom since we used our n observations to estimate two constants a and b. If the null hypothesis is true, the ratio below has F distribution with 1 numerator degree of freedom and $n-2$ denominator degrees of freedom. That is F ratio is defined by

$$F = \frac{SSR/1}{SSE/(n-2)}$$

If the F ratio is too high (as determined by the significance level of the test statistic), we reject H_0 and conclude that the regression as a whole is significant. This part of output by SPSS includes the computed F ratio for the regression, and is sometimes called the analysis of variance (ANOVA) for the regression.

Scanned with CamScanner

- **Test of population regression Coefficient β (Slope of the population regression line)**

Suppose the null hypothesis is $H_0 : \beta = \beta_0$ and the alternative hypothesis is $H_1 : \beta = \beta_1$. To find the test statistic for β, it is necessary first to find the standard error of the regression coefficient. Here the regression is b, so the standard error of this coefficient denoted by S_b is defined as

$$S_b = \frac{S_e}{\sqrt{\sum x^2 - n\bar{x}}}$$

where S_b = standard error of the regression coefficient; S_e = standard error of the estimate $\hat{y}$; x = values of the independent variable; $\bar{x}$ = mean of the values of the independent variable; and n = number of data points. The value of the test statistic under the null hypothesis is

$$t = \frac{b - \beta_0}{s_b}$$

where b = slope of fitted regression line; β_0 = value of β under the null hypothesis and s_b value of S_b.

Remarks: To test the null hypothesis $H_0 : \beta = 0$ and the alternative hypothesis is $H_1 : \beta \neq 0$ can also be tested with the test statistic $t = \frac{b-0}{s_b} = \frac{b}{s_b}$ which follows t distribution $n-2$ degrees of freedom. This is an equivalent test statistic that can also be used for testing regression equation as a whole by using an F test mentioned above. This is due to the fact that F is equal to t^2. The square of a t statistic with degrees freedom (df) has the same distribution as an F statistic with 1 numerator df and denominator degrees of freedom. The F test is a more general test of the usefulness of the model and can be used when the model has more than one independent variable.

Example 12.9.1 The authority of the municipality of Dhaka city is interested to know whether there is any relationship between the age of a truck in year and the amount of annual repair expense during the last in thousands of takas. The records of the last year gave the following information.

Annual Track Repair Expenses

Truck Number	Age of truck In Years (x)	Repair Expense During Last Year in Thousands of Takas (y)
1001	1	4
1002	3	6
1003	3	7
1004	5	7

a) Develop the estimating equation that best describes these data.
b) Find the expected annual repair cost of a truck of 6 years old.
c) Find the standard error of the estimate.
d) Find 90% prediction interval for your estimated value by (b).
e) Determine whether there is a significant linear relationship between the annual repair cost and the age of the truck. Test at the 5% level of significance.

Scanned with CamScanner

Solution. (a) The best fitted regression equation of annual repair expense y on the age of truck x is

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

The values of a and b are obtained from the formulae

$$a = \bar{y} - b\bar{x} \quad \text{and} \quad b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

Table for calculation of regression equation

	Age (x)	Repair Expense(y)	xy	x^2	y^2
	1	4	4	1	16
	3	6	18	9	36
	3	7	21	9	49
	5	7	35	25	49
Total	12	24	78	44	150

$$\bar{x} = \frac{\sum X}{n} = \frac{12}{4} = 3; \quad \bar{y} = \frac{\sum y}{n} = \frac{24}{4} = 6.$$

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} = \frac{78-(4)(3)(6)}{44-(4)(9)} = \frac{78-72}{44-36} = \frac{6}{8} = 0.75;$$

$$a = \bar{y} - b\bar{x} = 6 - (0.75)(3) = 6 - 2.25 = 3.75.$$

The best fitted regression line of y on x is: $\hat{y} = a + bx = 3.75 + 0.75x$.

(b) The expected annual repair cost of the truck of age 6 years is

$$\hat{y} = 3.75 + 0.75(6) = 3.75 + 4.50 = 8.25 \text{ thousand takas} = \text{Tk. } 8250.$$

Thus, the authority might expect to spend about Tk. 8250 annually in repairs on a 6-year-old truck.

(c) The standard error of estimate is

$$s_e = \sqrt{\frac{\sum y^2 - a\sum y - b\sum xy}{n-2}} = \sqrt{\frac{150-(3.75)(24)-(0.75)(78)}{4-2}} = \sqrt{\frac{150-90-58.5}{2}} = \sqrt{0.75} = 0.866$$

(d) Here the estimated value $\hat{y}$ follows t distribution with 2 degrees of freedom, since $n = 4$ is very small for normal approximation. Here the value of t with 2 degrees of freedom at 10% level of significance is 2.92. We use this value to find the prediction intervals for our estimated value will lie 90% confidence. Our 90% prediction interval limits for the estimated value Tk. 8250 are

$$\hat{y} + t(s_e) = 8250 + (2.92)(86.6)) = 8250 + 252.872 = \text{Tk. } 8502.872 \text{ and}$$

$$\hat{y} + t(s_e) = 8250 - (2.92)(86.6)) = 8250 - 252.872 = \text{TK. } 7997.124.$$

So the authority of the municipality can be 90% certain that the annual repair expense on a 6-year old truck will be between Tk. 850.872 and Tk. 7997.124.

(e) The hypothesis to be tested are

$$H_0 : \beta = 0 \quad \text{versus} \quad H_1 : \beta \neq 0. \quad t_{\alpha/2(2)} = 4.303 \text{ (for two tailed test)}$$

Scanned with CamScanner

The observed value of the test statistic is calculated as

$$t = \frac{b-0}{s_b} = \frac{b}{s_b} = \frac{0.75}{0.306} = 2.45$$

$$[s_b = \frac{s_e}{\sqrt{\Sigma x^2 - n\bar{x}^2}} = \frac{0.866}{\sqrt{44-4\times 9}} = \frac{0.866}{\sqrt{8}} = \frac{0.866}{2.83} = 0.306]$$

with 2 degrees of freedom.

Conclusion. The observed value of t falls in the acceptance region at 5% level of significance. We can conclude that there is no significant linear relationship between the annual repair cost and the age of the truck.

Example 12.9.2 A chemical firm wants to see whether there is any relationship between the amount of money spent on research and the annual profit of the firm. The related data on the last six years are given below :

Year	Amount of Money spent on research in million Takas(x)	Annual Profit in Million (y)
2006	4	30
2006	3	25
2007	5	31
2008	5	34
2009	2	20
2010	11	40

i) Find the least-squares regression line that could be used to predict annual profit from the amount of money annually spent on research.

ii) If the firm spends Tk.9 million for research in year 2012 what will be the profits of firm during that year.

iii) Test the hypothesis at 10% level of significance that the population regression coefficient $\beta = 2.1$.

iv) Also find the 90% confidence interval for β.

olution. (i) The best fitted regression equation of annual profit y on research expense x is

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

The values of a and b are obtained from the formulae

$$a = \bar{y} - b\bar{x} \quad \text{and} \quad b = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n\bar{x}^2}.$$

Table for calculation of regression equation

	Expenditures Research (x)	Annual Profits (y)	xy	x^2	y^2
	4	30	120	16	900
	3	25	75	9	625
	5	31	155	25	961
	5	34	170	25	1156
	2	20	40	4	400
	11	40	440	121	1600
Total	30	180	1000	200	5642

Scanned with CamScanner

$$\bar{x} = \frac{\Sigma X}{n} = \frac{30}{6} = 5; \qquad \bar{y} = \frac{\Sigma y}{n} = \frac{180}{6} = 30$$

$$b = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n\bar{x}^2} = \frac{1000-(6)(5)(30)}{200-(6)(25)} = \frac{1000-900}{200-150} = \frac{100}{50} = 2;$$

$$a = \bar{y} - b\bar{x} = 30 - (2)(5) = 30 - 10 = 20.$$

The best fitted regression line of y on x is

$$\hat{y} = a + bx = 20 + 2x.$$

(ii) The expected annual profits of the firm in 2012 when the expenditures on research TK. 9 million is

$$\hat{y} = 20 + 2(9) = 20 + 18 = 38 \text{ million takas.}$$

(iii) The standard error of estimate by the short-cut method is

$$s_e = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n-2}} = \sqrt{\frac{5642-(20)(180)-(2)(1000)}{6-2}} = \sqrt{\frac{5642-3600-2000}{4}}$$

$$= \sqrt{\frac{5642-5600}{4}} = \sqrt{\frac{42}{4}} = \sqrt{10.5} = 3.24.$$

(iv) Here we could define hypothesis as

$$H_0 : \beta_0 = 2.1 \quad \leftarrow \text{Null Hypothesis}$$

$$H_1 : \beta_1 \neq 2.1 \quad \leftarrow \text{Alternative hypothesis}$$

To test the null hypothesis, we use the test statistic: $t = \frac{b-\beta}{S_b}$

which follows student's t with $n-2$ degrees freedom. S_b is the standard error of the regression coefficient b calculated by the formula.

$$S_b = \frac{S_e}{\sqrt{\Sigma x^2 - n\bar{x}^2}}.$$

The value of S_b is: $S_b = \frac{S_e}{\sqrt{\Sigma x^2 - n\bar{x}^2}} = \frac{3.24}{\sqrt{200-(6)(25)}} = \frac{3.24}{\sqrt{50}} = \frac{3.24}{7.07} = 046.$

Now the calculated value of the test statistic under the null hypothesis is

$$t = \frac{b-\beta_0}{s_b} = \frac{2.0-2.1}{0.46} = \frac{-0.1}{0.46} = -0.217.$$

Conclusion. The tabulated value of t at 10% level of significance with 4 degrees of freedom is 2.132 which is greater than the calculated value of t with the same degrees of freedom. Hence we have no reason to reject the null hypothesis. That is, we accept the null hypothesis.

(b) The value $b = 2$ is a point estimate of β. The 90% confidence interval for β is

$$b \pm t_{0.05(4)}(s_b)$$

Here $b = 2$, $s_b = 0.46$ and $t_{0.05(4)} = 2.132$. Hence the upper limit of the confidence interval is

$$b + t_{0.05(4)}(s_b) = 2 + (2.132)(0.46) = 2 + 0.981 = 2.981$$

Scanned with CamScanner

The lower limit of the confidence interval is

$$b - t_{0.05(4)}(s_b) = 2 - (2.132)(0.46) = 2 - 0.981 = 1.019.$$

That is, we are 90 percent confident that the true value of β lies between 1,019 and 2.981. It means, each additional million taka spent on research increases annual profits by some amount between Tk. 1.02 million and Tk. 2.09 million.

Example 12.9.3 Suppose in a regression problem with a sample of size 17, the slope was found to be 3.73 and the standard error of estimate is 28.654. The quantity

$$(\Sigma x^2 - n\bar{x}^2) = 87.56.$$

a) Find the standard error of the regression slope coefficient.

b) Construct a 96% percent confidence interval for the population slope.

c) Interpret the confidence interval of part (b).

Solution. (a) The standard error of the regression coefficient b is

$$s_b = \frac{s_e}{\sqrt{\Sigma x^2 - n\bar{x}^2}} = \frac{28.654}{\sqrt{471.56}} = 0.9706$$

(b) The 96% confidence interval is

$$b \pm t(s_b) = 3.73 \pm 2.602(0.9706) = 3.73 \pm 2.53 = (1.20,\ 6.26)$$

(Here $\alpha = .04$ and the degrees of freedom of t is 15. The tabulated value of t with 15 df at 2% level of significance is ± 2.602).

(c) In repeated sampling, 96% out of 100 intervals constructed as above would contain the true, unknown population slope β. For our single sample, we can say that we are 96% confident that our computed interval contains β.

Example 12.9.4 In finance, it is interest to look at the relationship between y, a stock average return, and x, the overall market return. The regression coefficient computed by linear regression is called the stock's beta by investment analysts. A beta greater than 1 indicates that the stock is relatively sensitive to change in the market; a beta less than 1 indicates that the stock is relatively insensitive. For the following information, compute (i) the regression line of stock return y on the overall market return x; (ii) the beta and test to see whether it is significantly less than 1. Use $\alpha = .05$. $\Sigma x = 113$, $\Sigma y = 107$, $\Sigma xy = 1{,}301$, $\Sigma x^2 = 1{,}501$, $\Sigma y^2 = 1{,}189$

Solution. Simple regression analysis can be easily done by using the statistical software SPSS in a few seconds.

Group-A : Short questions and answers

1. **What is a dependent variable?**
   Ans. The variable to be predicted or explained is called dependent variable.
2. **What is independent variable?**
   Ans. The variable included in a regression model to explant the variation of the dependent variable is called independent variable.

Scanned with CamScanner

3. **What is least square method?**
Ans. The method used to fit a regression line by a set of pair observations in such a way that the error sum of squares is minimum.
4. **What is least square regression line or best fitted regression line. ?**
Ans. A regression line obtained by the least square method is called least squares regression line or best fitted regression line.
5. **What is random error?**
Ans. The difference between the actual observed value and the predicted value of the dependent y is called error e.
6. **What is error sum of squares (SSE)?**
Ans. The sum of squares differences between the actual and the predicted values of y is called the error sum of squares. It is the proportion of the total sum of squares that is not explained by the regression model.
7. **What is regression sum of squares (SSR)?**
Ans. The portion of SST (total sum of squares) that is explained by the regression model is regression sum of squares.
8. **What is total sum of squares (TSS)?**
Ans. The sum of squares differences between the actual values of y and its mean $\bar{y}$.
9. **What is slope or regression coefficient?**
Ans. The coefficient of the independent variable x in a regression model that gives the change in y for unit change in x.
10. **What is y-intercept a ?**
Ans. The points at which the regression line intersects the vertical axis on which the dependent variable is marked is called y-intercept a. Actual, it is the value of y when x is zero.
11. **What is the equation of a sample regression line of y on x?**
Ans. The equation of a sample regression line of y on x is $y = a + bx$.
12. **What is the equation of a sample regression model of y on x?**
Ans. The equation of a sample regression model of y on x is $y = a + bx + e$.
13. **What is the equation of a population regression line of Y on X?**
Ans. The equation of a population regression line of Y on X is $Y = \alpha + \beta X$
14. **What is the equation of a population regression model of Y on X?**
Ans. The equation of a sample regression line of Y on X is $Y = \alpha + \beta X + \varepsilon$.
15. **What is the equation of a sample regression line of x on y?**
Ans. The equation of a sample regression line of x on y is $x = c + dy$.
16. **What is the equation of a sample regression model of x on y?**
Ans. The equation of a sample regression model of x on y is $x = c + dx + e$.
17. **What is the equation of a population regression line of X on Y?**
Ans. The equation of a population regression line of X on Y is $X = \alpha' + \beta' Y$
18. **What is the equation of a population regression model of Y on X=?**
Ans. The equation of a sample regression line of X on Y is $X = \alpha' + \beta' Y + \varepsilon'$.
19. **What is the relationship between the regression coefficients and the correlation coefficient?**
Ans. The correlation coefficient is the geometric mean of the regression coefficients.
20. **What are the signs of the regression coefficients if the correlation coefficient is positive?**
Ans. The signs of both the regression coefficients must be positive, since correlation coefficient and the regression coefficients must have the same sign.
21. **What will be the range of the regression coefficient of x on y if the regression coefficient of y on x is greater than 1.?**
Ans. The range of the regression coefficient x on y must be less than 1.

Scanned with CamScanner

Group-B & C: Broad questions and problems

1. Distinguish between correlation and regression. State some important properties of regression coefficient.
2. What are regression lines? Define regression coefficients. Show that correlation coefficient is the geometric mean of the regression coefficients.
3. Distinguish between regression model and regression line. Write the equation of the best fitted regression line of y on x. Discuss its role in business.
4. Define dependent and independent variables. Is it reasonable to write two regression lines (i) y on x and (ii) x on y? Justify in favour of your answer.
5. Distinguish between correlation and regression. State some uses of regression in business.

Exercises

6. Consider the following data set on two variables x and y.

x	1	2	3	4	5	6
y	6	4	3	5	4	2

a) Find the equation of the regression line y on x.
b) Graph the line on a scatter diagram.
c) Estimate the value of y when $x = 4.5$.
d) Predict the value of y when $x = 8$.

Ans. (a) $\hat{y} = 5.799 - 0.514x$; (c) $\hat{y} = 3.486$; (d) $\hat{y} = 1.687$

7. Calculate the regression equations of
   i) y on x and
   ii) x on y from the following pairs of data set

x	1	2	3	4	5
y	2	5	3	8	7

   iii) Estimate the value y when $x = 3.5$;
   iv) Predict the value x when $y = 10$.

Ans. (a) $\hat{y} = 1.10 + 1.30x$; (b) $\hat{x} = 0.5 + 0.5x$; (c) $\hat{y} = 5.65$; (d) $\hat{x} = 5.5$.

8. Calculate the regression lines of
   i) y on x
   ii) x on y
   iii) Estimate the value of y when $x = 16$.
   iv) Find the probable value of x when $y = 16$
   v) Find the correlation coefficient with the help of the regression coefficients.

Ans. $\hat{y} = 8.8 + 1.015x$, (ii) $\hat{y}_{16} = 5.04$, (iii) $\hat{x} = -7.05 + .91y$, (iv) $\hat{x}_{16} = 7.51$, $r = 0.96$.

Applications

9. The following summary data were obtained for closing prices of twelve stocks x on Chittagong Stock Exchange on a certain day, along with the volume of sales in thousands of shares y

$$\Sigma x = 580, \ \Sigma y = 370, \ \Sigma xy = 11494, \ \Sigma x^2 = 41{,}658, \ \Sigma y^2 = 17{,}206$$

Find i) The regression line of y on x.

Scanned with CamScanner

ii) The regression line of x on y.

iii) Correlation coefficient between x and y.

Ans. (i) $\hat{y} = 53.55 - 0.47x$, (ii) $\hat{x} = 79.16 - 1.1y$, (iii) $r = -0.517$

10. A study was made by a retail merchant to determine the relation between weekly advertising expenditure and sales. The following data were recorded.

Expenditure ($)	40	20	25	20	30	50	40	20	50	40	25	50
Sales ($)	385	400	395	365	475	440	490	420	560	525	480	510

i) Plot a Scatter diagram.

ii) Find the equation of the regression line to predict weekly sales from advertising expenditures.

iii) Estimate the weekly sales when advertising costs are 350.

Ans. (ii) $\hat{y} = 343.699 + 3.221x$; (iii) $\hat{y} = 456$.

11. A study was made on the amount of converted sugar in a certain process at various temperatures. The data were coded and recorded as follows.

Temperature (x)	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
Converted sugar (y)	8.1	7.8	8.5	9.8	9.5	8.9	8.6	10.2	9.3	9.2	10.5

a) Estimate the linear regression line of y on x.

b) Estimate the amount of converted sugar produced when the coded temperature is 1.75.

Ans. (a) $\hat{y} = 6.414 + 1.809x$; (b) $\hat{y} = 9.58$

12. In a study between the amount of rainfall and the quantity of air pollution removed, the following data were collected.

Daily Rainfall (centimeter)	4.3	4.5	5.9	5.6	6.1	5.2	3.8	2.1	7.5
Pollution Removed (Micrograms per cubic meter)	126	121	116	118	114	118	132	141	108

i) Find the equation of the regression line to predict the pollution removed from the amount of daily rainfall.

ii) Estimate the amount of pollution removed when the daily rainfall is $x = 4.8$ centimeter.

13. The following data relate to advertising expenditure (in lakh taka) and their corresponding sales in crores taka of a firm in last five years.

Advertising Expenditure : x	10	12	15	23	20
Sales : y	14	17	23	25	21

a) Find the regression line of sales on expenditure.

b) Estimate the sales of the firm when the advertising expenditure is Tk.30 lakhs.

c) Find the equation of the regression line of expenditure on sales

d) Estimate the expenditure of the firm when the sales is 35 crores.

e) Calculate the correlation coefficient and the coefficient of determination of advertising expenditure and sales.

Ans. (a) $\hat{y} = 8.608 + 0.712x$; (b) Tk.29.968 crores; (c) $\hat{x} = -5 + 1.05y$;

(d) Tk 31.75 Lakhs; $r = 0.864$ and (e) $r^2 = 0.747$.

14. A researcher wants to find out if there is any relationship between the ages of the husbands and the ages of the wives. In other words, do old husbands have old wives and young husbands have young wives? He took a random sample of 6 couples whose respective ages are given below :

Age of Husband : x	39	25	35	32	27	37
Age of Wife : y	37	18	25	25	20	30

Scanned with CamScanner

a) Find the regression line of wife on husband and estimate the probable age of wife if the age of husband is 30.

b) Find the regression line of husband on wife and predict the probable age of husband if the age of wife is 40.

c) Compute the Karl Pearson's coefficient of correlation with the help of the regression coefficients and from the original formula.

15. 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 taka)	10	20	25	30

Estimate the maintenance cost for a 7 year old car.

16. The following data give price and supply of a commodity for last 9 years.

Supply	80	82	86	91	83	85	89	96	93
Price	145	140	130	124	133	127	120	110	116

(a) Find a regression line of price on supply.

(b) Compute the correlation coefficient between price and supply.

Ans. (a) $y = 301.75 + 2x$; (b) $r = -0.96$.

17. i) Find the regression line of production on rainfall from the following data.

	Rainfall in inches : x	Production in kg : y
Average	30	500
Standard deviation	5	100
Correlation coefficient $= 0.8$		

ii) Also find the most likely production corresponding to a rainfall 40 inches.

Ans. (i) $\hat{y} = 20 + 16x$; (ii) 660kg.

18. The following data about the sales and advertisement expenditure of a firm is given below.

	Sales (in crores Tk.)	Advertisement Expenditure (in crores Tk.)
Mean	40	6
Standard deviation	10	1.5
Correlation coefficient: $r = 0.9$		

i) Find the regression line of sales x on expenditure y.

ii) Estimate the sales of the firm when the expenditure on advertisement is Tk. 10 crores.

iii) Find the regression line of advertisement expenditure on sales.

iv) What should be the advertisement expenditure if the firm proposes a sale target of Tk. 60 crores.

Ans. (i) $\hat{y} = 4 + 6x$, (ii) Tk 64crores, (iii) $\hat{x} = 0.6 + 0.135y$, (iv) Tk 8.7 crores

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 taka)	10	20	25	30

Estimate the maintenance cost for a 7-year-old car.

Scanned with CamScanner

20. The following table gives the per unit cost (in hundred Taka) and selling price (in hundred Taka) of 8 products of a company.

Product	A	B	C	D	E	F	G	H
Cost price	10	15	14	20	31	34	57	65
Selling price	11.5	18.0	18.5	20.9	33.2	39.0	64.2	74

a) Fit a suitable regression line and comment.

b) Estimate the profit incurred by the company for a product whose cost price is Taka 50 (hundred).

21. The following are the advertisement expenditure and sales of a certain firm:

Advertisement Expenditure (Tk. Lakh):	60	62	65	70	73	75	71
Sales (Tk. Crore):	10	11	13	15	16	19	14

(i) Find out the likely sales when advertisement budget is Tk.80 lakhs

(ii) Estimate the advertisement expenditure when the sales target is Tk. 25 Crores.

D.U. Acct. 2002 **Ans. (i) Tk. 20.12 Crores (ii) Tk. 87.69 lakhs.**

22. Calculate a regression line of y on x from the following information:

Production in mound (x):	20	25	28	32	36
Price in taka (y):	48	60	68	70	75

Ans. y=17.73 + 1.65 x

23. You are given the following information about advertisement expenditure and sales:

	Adv. Exp.(x) (Tk. Lakh)	Sales (y) (Tk. Lakh)
Mean	10	90
Standard Deviation	3	12
Correlation coefficient	r = 0.80	

(i) Calculate the two regression lines.

(ii) Find the likely sales when advertisement budget is Tk. 15 lakh . *D.U. Acct.2005, 1986*

Ans. (i) y=58 +3 .2x ; x=-8+0.2 (ii) Tk.106 lakh.

Concepts test

Write T or true and F for failure of the following.

a. The variable to be predicted is called independe4nt variable.

b. The variable in a regression model to explain the variation in the dependent is calle independent variable.

c. The best fitted regression line is obtained by the method of least squares.

d. The sample error term e in the regression model is the difference between the actu observation y and the predicted value of y.

e. The value of the correlation coefficient r is called the coefficient of determination.

f. The regression coefficient measures the change of dependent for a unit change of independen variable.

Scanned with CamScanner

g. The simple sample coefficient of determination r^2 measured proportion of variation in y that is explained by the independent variable x in the regression line

h. The regression analysis is a statistical technique to predict the value of a dependent for a known value of an independent variable.

i. The value of r = 0.75 means that 75% variation of y is explained the independent variable x.

j. The value of r^2 =0.81 means 81% variation of the predicted variable of y is explained by the independent variable x.

k. An r^2 value close to zero indicates a very strong relationship between x and y.

l. The regression line y = 3 - 4x indicates a positive and perfect relationship between x and y.

m. The regression line y = 2+3x indicates a perfect and positive relationship between x and y.

n. The correlation coefficient is the geometric mean of the regression coefficients. T

o. The equation of the regression line of x on y is x = c+dy and the regression line of y on x is y=a+bx.

p. The correlation coefficient and the regression coefficients have the same sign.

q. If one of the regression coefficient is greater than 1, then the other must be less than 1.

r. If the value of the regression coefficient y on x is b equals to -2 then the value of the regression coefficient of x on y must be a positive quantity.

s. If the value of correlation coefficient is positive then at least one of the regression coefficient must be negative.

Ans. a. F, b. T, c. T, d. T, e. F, f. T, g. T, h. T, i. F, j. T, k. F, l. F, m. T, n. T, o. T, p. T, q. T, r. F, s. F

Multiple Choice

Choose the correct answer of the following.

i. The range of the regression coefficient is

(a) -1 to +1 (b) 0 to 1

(c) $-\infty$ *to* $+\infty$ (d) 0 *to* $+\infty$

ii. If one of the regression coefficient is greater than 1, then the other must be

(a) greater than 1 (b) less than 1

(c) a and b (d) a or b

iii. A simple regression is a regression model that contains

(a) only one independent variable (b) only one dependent variable

(c) more than one independent variables (d) both a and b

iv. The relationship between an independent and a dependent variable in simple regres-sion is

(a) a straight line (b) a quadratic

(c) a curve (d) both a and b

Scanned with CamScanner

v. A regression model

(a) contains an error term (b) does not contain an error term

(c) gives a nonlinear relationship (d) both a and b

vi. The least squares regression line minimizes

(a) errors (b) absolute errors

(c) squared errors (d) sum of squared errors

vii. The degrees of freedom in simple regression line is

(a) n – 1 (b) n – 2

(c) n – 3 (d) n – 4

viii. The value of the coefficient of determination is always in the range

(a) -1 to +1 (b) 0 to 1

(c) $-\infty$ *to* $+\infty$ (d) 0 *to* $+\infty$

ix. The value of the coefficient of determination for a particular situation is 0.64. The value of the coefficient of correlation is

(a) 0.64 (b) 0.8

(c) 8 (d) None

x. Suppose the two regression coefficients are b and d. Then the correlation coefficient r is

(a) b/d (b) d/b

(c) bd (d) $\pm\sqrt{b \times d}$

xi. When the correlation coefficient r = 1, then the two regression lines

(a) are perpendicular to each other (b) are coincide each other

(c) are parallel to each other (d) do not exist

xii. The two regression lines are x+2y-5=0 and 2x+3y=8. Then the mean values of x and y are

(a) (2, 1) (b) (1, 2)

(c) (2, 2) (d) (1. 3)

xiii. The regression line of y on x minimizes the total sum of squares of

(a) horizontal deviations (b) vertical deviations

(c) both horizontal and vertical deviations (d) none of these

Answers

1.	c	2.	b	3.	d	4.	a	5.	a	6.	d	7.	b	8.	b	9.	b	10.	d
11.	b	12.	b	13.	b														

Scanned with CamScanner

Example 12.8.5 A researcher wants to find out if there is any relationship between the heights of the sons and the heights of the fathers. He took a random sample of seven fathers and their seven sons. Their heights in inches are given below :

Height of father (In inches)	68	63	66	67	65	67	66
Height of son (In inches)	70	66	65	69	68	67	64

i) Fit a regression line of the height of father y on the height of son x.

ii) Predict the height of son if father's height is 70 inches.

iii) Fit a regression line of the height of son on the height of father.

iv) Find the height of father if son's height is 65 inches.

v) Calculate the correlation coefficient with the help of regression coefficients and from the original formula.

Solution. Here, $\bar{x} = \frac{\Sigma x}{n} = \frac{462}{7} = 66$ *and* $\bar{y} = \frac{\Sigma y}{n} = \frac{469}{7} = 67$.

Here both the means are whole numbers. We define the new variables as

$$u = x - 66 \quad \text{and} \quad v = y - 67.$$

Calculation table

x	u = x – 66	u^2	y		v = y – 67	v^2	uv
68	2	4	70		3	9	6
63	–3	9	66		–1	1	3
66	0	0	65		–2	4	0
67	1	1	69		2	4	2
65	–1	1	68		1	1	–1
67	1	1	67		0	0	0
66	0	0	64		–3	9	0
$\Sigma x = 462$	$\Sigma u = 0$	$\Sigma u^2 = 16$	$\Sigma y = 469$		$\Sigma v = 0$	$\Sigma v^2 = 28$	$\Sigma uv = 10$

(i) Hence the required regression line of y on x is

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

$$\hat{y} = \bar{y} + b(x - \bar{x}) = 67 + 0.625x - (0.625)(66) = 67 + 0.625x - 41.25 = 25.75 + 0.625x.$$

(ii) Hence, if the height of the father is 69 inches or $x = 69$, the height of the son is

$$\hat{y} = 25.75 + (0.625)(69) = 25.75 + 43.125 = 68.875 \text{ inches.}$$

(iii) The regression line of son x on father y is: $\hat{x} - \bar{x} = d(y - \bar{y})$.

Here d is the regression coefficient of x on y: $d = \frac{\Sigma uv}{\Sigma v^2} = \frac{10}{28} = 0.357$.

Hence the required regression line of x on y is

$$\hat{x} - \bar{x} = d(y - \bar{y})$$

$$\hat{x} = \bar{x} + d(y - \bar{y}) = 66 + 0.357y - (0.357)(67) = 66 + 0.357y - 23.919 = 42.081 + 0357y.$$

(iv) Hence, if the height of son is 71 inches, the height of father would be

$$\hat{x} = 42.081 + (0.357)(71) = 42.081 + 25.357 = 67.438 \text{ inches.}$$

Scanned with CamScanner

(v) Correlation coefficient by using regression coefficients is

$$r=\sqrt{b\times d}=\sqrt{(0.625)(0357}=\sqrt{0.223}=0.472.$$

Correlation coefficient by the original formula

$$r=\frac{\Sigma uv}{\sqrt{(\Sigma u^2)(\Sigma v^2)}}=\frac{10}{\sqrt{16\times 28}}=\frac{10}{21.166}=0.472.$$

Example 12.8.6 A researcher wants to find out if there is any relationship between the ages of the husbands and the ages of the wives. In other words, do old husbands have old wives and young husbands have young wives? He took a random sample of 7 couples whose respective ages are given below :

Age of Husband (in years) : x	39	25	29	35	32	27	37
Age of Wife (in years) : y	37	18	20	25	25	20	30

i) Compute the regression line of y on x by direct method and the short cut method.
ii) Predict the age of wife whose husband's age is 45 years.
iii) Find the regression line of x on y and estimate the age of husband if the age of his wife is 28 years.

Solution. (i) The equation of the best-fitted regression line of y on x is: $\hat{y}=a+bx$;

Here, $b=\dfrac{\Sigma xy-\dfrac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2-\dfrac{(\Sigma x)^2}{n}}$; $a=\bar{y}-b\bar{x}$

Computation table

X	Y	x^2	y^2	xy
39	37	1521	1369	1443
25	18	625	324	450
29	20	841	400	580
35	25	1225	625	875
32	25	1024	625	800
27	20	729	400	540
37	30	1369	900	1110
$\Sigma x = 224$	$\Sigma y = 175$	$\Sigma x^2 = 7334$	$\Sigma y^2 = 4643$	$\Sigma xy = 5798$

$$b=\frac{\Sigma xy-\dfrac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2-\dfrac{(\Sigma x)^2}{n}}=\frac{5798-(224)(175)/7}{7334-\dfrac{(224)^2}{7}}=\frac{5798-5600}{7334-7168}=\frac{198}{166}=1.193$$

$$a=\bar{y}-b\bar{x}=25-(1.193)(32)=25-38.176=-13.176.$$

Hence the required regression line is: $\hat{y}=-13.176+1.193x$.

(ii) Hence, if the age of husband is 45, the probable age of wife would be

$$\hat{y}=-13.176+1.193\times 45=40.51 \text{ years.}$$

Scanned with CamScanner

Short cut method for finding the regression lines

Here, $\bar{x} = \frac{\Sigma x}{n} = \frac{224}{7} = 32$ and $\bar{y} = \frac{\Sigma y}{n} = \frac{175}{7} = 25$. Here both the means are whole numbers.

We define the new variables u and v as, $u = x - 32$ and $v = y - 25$.

(i) The equation of the best fitted regression of y on x is: $\hat{y} - \bar{y} = b(x - \bar{x})$.

Here b is the regression coefficient of y on x is: $b = \frac{\Sigma uv}{\Sigma u^2}$.

(ii) The equation of the best fitted regression of x on y is: $\hat{x} - \bar{x} = d(y - \bar{y})$.

Here d is the regression coefficient of x on y, is: $d = \frac{\Sigma uv}{\Sigma v^2}$.

Computation table

x	u = x–32	u^2	y	v = y–25	v^2	uv
39	7	49	37	12	144	84
25	–7	49	18	–7	49	49
29	–3	9	20	–5	25	15
35	3	9	25	0	0	0
32	0	0	25	0	0	0
27	–5	25	20	–5	25	25
37	5	25	30	5	25	25
$\Sigma x = 224$	$\Sigma u = 0$	$\Sigma u^2 = 166$	$\Sigma y = 175$	$\Sigma v = 0$	$\Sigma v^2 = 268$	$\Sigma uv = 198$

The equation of the best fitted regression of y on x is

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

$$\Rightarrow \hat{y} = \bar{y} + b(x - \bar{x}).$$

The formula for b is: $b = \frac{\Sigma uv}{\Sigma u^2} = \frac{198}{166} = 1.193$.

Hence the required regression line of y on x is

$$\hat{y} = \bar{y} + b(x - \bar{x}) = 25 + 1.193(x - 32) = -13.176 + 1.193x.$$

The equation of the best fitted regression of x on y is:

$$\hat{x} - \bar{x} = d(y - \bar{y})$$

$$\Rightarrow \hat{x} = \bar{x} + d(y - \bar{y}).$$

The regression coefficient of x on y is: $d = \frac{\Sigma uv}{\Sigma v^2} = \frac{198}{268} = 0.739$.

Hence the required regression line of x on y is:

$$\hat{x} = \bar{x} + d(y - \bar{y}) = 32 + 0.739y - (0.739)(25) = 32 + 0.739y - 18.475 = 13.525 + 0.739y.$$

Hence, if the age of wife is 28 years, the estimate age of husband is:

$$\hat{x} = 13.525 + (.739)(28) = 34.22 \text{ years.}$$

Scanned with CamScanner

Example12.8.7 A researcher got the following summary data form a community: Correlation coefficient between the ages of husbands and wives is 0.9.

- The average age of husbands = 30 years,
- The average age of wives = 25 years,
- The standard deviation of the ages of husbands = 5,
- The standard deviation of the ages of wives = 6.

On the basis of the above information, compute the regression lines of the ages of

i) Husband on the ages of wives and estimates the age of husband if the age of wife is 18 years;

ii) Wives on the ages of husbands and predict the age of wife if the age of husband is 40 years.

Solution. Suppose the age of wife is y and the age husband is x.

Here, we have $\bar{x}=30,\ \bar{y}=25,\ \sigma_x=5,\ \sigma_y=6$ and $r=0.9$.

(i) The required regression line will be x on y. On the basis of the above information, the regression equation of x on y is

$$\hat{x}=\bar{x}+r\frac{\sigma_x}{\sigma_y}(y-\bar{y})=30+\frac{(0.9)(5)}{6}(y-25) = 30+0.75y-18.75 = 11.25+0.75y.$$

When $y=18$, then the probable age of husband is: $\hat{x}_{18}=11.25+0.75\times 18=24.75$ years.

(ii) The regression line of the age of wife y on age of husband x is

$$\hat{y}=\bar{y}+r\frac{\sigma_y}{\sigma_x}(x-\bar{x})=25+\frac{(0.9)(6)}{5}(x-30) = 25+1.08x-32.4 = -7.4+1.08x.$$

When $x=40$, the probable age of wife is $\hat{y}_{40}=-7.4+1.08\times 40=38.8$ years.

Example 12.8.8 The following summary data refers to the rainfall and production of a Rabi crop in a particular region.

	Rainfall (In inches)	Production (In quintals)
Mean	29	40
Standard deviation	3	6

Coefficient of correlation between rainfall and production is 0.8.

i) Find the regression line of production on rainfall.

ii) Find the most probable production corresponding to a rainfall of 40 inches.

Solution. (i) Let rainfall be denoted by x and production by y. The equation of the best fitted regression line of production on rainfall is

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

Here, $\bar{x}=29,\ \bar{y}=40,\ \sigma_x=3,\ \sigma_y=6$ *and* $r=0.8$. Then

$$\hat{y}=\bar{y}+r\frac{\sigma_y}{\sigma_x}(x-\bar{x})=40+(0.4)\frac{6}{3}(x-29)=40+0.8x-23.2=16.8+0.8x.$$

(ii) When $x=40$ inches, the possible production the Rabi crop is

$$\hat{y}_{40}=16.8+(0.8)(40)=16.8+32=48.8 \text{ quintals}$$

Scanned with CamScanner