LINEAR REGRESSION

Introduction

we have established the fact that if two variables are closely related we may be interested in estimating the value of one variable given the value of another. For example, if we know that in milk, the content of total solids and fat levels are correlated we want to find out expected total solids in milk for a given fat level. Similarly, if we know that spoilage of milk (in %) and the temperature (°C) of storage of milk in a dairy plant are closely related we may find out the level of temperature at which spoilage of milk starts. It is often of interest to determine how change of values of some variables influences the change of values of other variables. Regression analysis reveals average relationship between two variables and this makes possible estimation or prediction. The literal or dictionary meaning of the word Regression is stepping back or returning to the average value. The term was first used by British biometrician Sir Francis Galton in the later part of the 19^{th} century in connection with some studies made on estimating the extent to which the stature of the population. Actually regression means to regress i.e., to step back or to fall back or to return back to a former state. So regression means returning of retrogression. Falconer (1936) conducted an experiment in which he took two groups of parents, one group having more height while others having shorter than the normal height. It was found that the children of the first group of parents try to go back to normal height while the children of second group of parents try to reach the normal height. Regression analysis in the general sense means the estimation or prediction of the unknown value of one variable from the known value of the other variable. It is one of the very important statistical tools which are extensively used in almost all branches of science-natural, social and physical. It is specially used in business and economics to study the relationship between two or more variables that are related casually and for estimation of demand and supply curves, cost functions, production and consumption functions, etc.

25.2 Definition of Regression

Regression analysis is one of the very scientific techniques for making predictions. In the words of M.M. Blair "Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data".

According to Morris Hamburg the term "regression analysis refers to the methods by which estimates are made of the values of a variable from a knowledge of the values of one or more values of other variables and to the measurement of the errors involved in this estimation process".

Ya-Lun Chou defined "Regression analysis attempts to establish the nature of the relationship between variables-that is, to study the functional relationship between the variables and thereby provide a mechanism for prediction or forecasting".

It is clear from the above definitions that regression analysis is a statistical device with the help of which estimate (predict) the unknown values of one variable from known values of another variable. In regression analysis there are two types of variables. The variables whose value is influenced or is to be predicted is called dependent variable and the variable which influence the values or is used for prediction, is called independent variable. In regression analysis independent variable is also known as regressor or predictor or explanatory while the dependent variable is also known as regressed or explained variable.

25.3 Types of Regression Analysis

The main types of regression analysis are as follows:

a) Simple and Multiple
b) Linear and Non- Linear

25.3.1 Simple and multiple

The regression analysis confined to the study of only two variables at a time is termed as simple regression. In simple regression analysis one variable is dependent and another is independent. The functional relationship between total solids and fat content in milk samples is an example of simple regression. But quite often the values of a particular phenomenon may be affected by multiplicity of factors. The regression analysis where we study more than two variables at a time is known as multiple regression for example, the study of effect of fat and SNF contents, on total solids in milk of samples; the study of Total Quality as affected by Methyl Blue Reduction Time (MBR) and Standard Plate Counts (SPC) etc.

25.3.2 Linear and non linear

If the given bivariate data are plotted on a graph, the points so obtained on the scatter diagram will more or less concentrate round a curve, called the" ***curve of regression"***. Often such a curve is not distinct and is quite confusing and sometimes complicated too. The mathematical equation of the regression curve, usually called the regression equation, enables us to study the average change in the value of the dependent variable for any given value of the independent variable. If the regression curve is a straight line, we say that there is linear regression between the variables under study. The equation of such a curve is the equation of a straight line, *i.e.*, a first degree equation in the variable X and Y. In case of linear regression, the values of the dependent variable increase by a constant absolute amount for a unit change in the value of the independent variable. However, if the curve of regression is not a straight line, the regression is termed as

curved or non-linear regression. In that case the regression equation is a functional relation between X and Y involving transformed values of X and Y, i.e., involving terms of the type X^2, Y^2, XY, log X, log Y etc. However, in this chapter we shall confine our discussion to linear regression between two variables only.

Simple Linear Regression

In practice, simple linear regression is often used and under this, regression lines, regression equations and regression coefficients are very important to be studied, which are discussed in the subsequent sections.

Regression Lines

The regression line shows the average relationship between two variables. It is the line which gives the best estimate of one variable for given value of other variable. The term best fit is interpreted in accordance with the Principle of Least Squares which consists in minimizing the sum of the squares of the residuals or the errors of estimates, i.e., the deviations between the given observed values of the variable and their corresponding estimated values as given by the line of best fit. In case of two variables X and Y, we shall have two lines regression one for Y on X and the other for X on Y.

Regression line of Y on X

Line of regression of Yon X is the line which gives the best estimate for the value of Y for any specified value of X and is obtained by minimising the sum of squares of the errors parallel to Y-axis.

Regression line of X on Y

Line of regression of X on Y is the line which gives the best estimate for the value of X for any specified value of Y and is obtained by minimising the sum of squares of the errors parallel to X-axis.

25.6 Derivation of Line of Regression of Y on X

Let (X_1, Y_1), (X_2, Y_2),...., (X_n, Y_n) be n pairs of observations on two variables understand (Eq. 25.1)

be the line of regression (best fit of Y on X). For a given point $P_i (X_i, Y_i)$ in the scatter diagram, the error estimate or residual as given by the line of best fit (Eq. 25.1) is P_iH_i as shown in figure 25.1. Now the X- coordinate of H_i is same as that of P_i, so X_i lies on the same line (25.1) the Y-coordinates of H_i i.e., H_i M is given by $(a + b X_i)$. Hence, the error of estimate for P_i is given by

$P_iH_i = P_iM - H_i M = Y_i (a + bX_i)$ (Eq. 25.2)

This error is parallel to Y-axis for the i^{th} point and we compute such error for all points of scatter diagram. The P_iH_i which lie above the line be positive and below the line, the error will be negative. There will be several lines passing throw these scatter of points and we have to find that particular line of best fit for which deviation or residual is minimum.

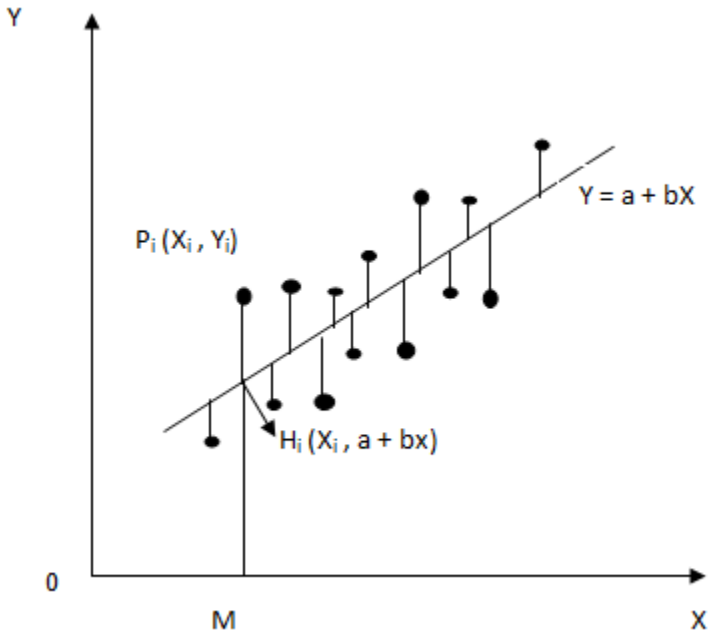

Fig. 25.1 Scatter diagram with an estimating line

According to principle of least squares, we have to determine the constants a and b in equation (25.1) such that the residual or deviation sum of squares of the errors is minimum. In other words we have to minimise the residual sum of squares due to error ❑E❑

$$E = \sum_{i=1}^{n}(P_iH_i)^2 = \sum_{i=1}^{n}\left(Y_i - (\hat{a} + \hat{b}X_i)\right)^2$$

Differentiating E partially with respect to $\hat{a}$ and $\hat{b}$ we get

$$\frac{\partial E}{\partial \hat{a}} = \sum_{i=1}^{n}(Y_i - \hat{a} - \hat{b}X_i)(-1) \Rightarrow \sum(Y_i - \hat{a} - \hat{b}X_i) = 0$$

$$\sum Y_i - n\hat{a} - \hat{b}\sum X_i = 0 \Rightarrow \sum Y_i = n\hat{a} + \hat{b}\sum X_i$$

$$\frac{\partial E}{\partial \hat{b}} = \sum_{i=1}^{n}(Y_i - \hat{a} - \hat{b}X_i)(-X_i) = \sum_{i=1}^{n}(\hat{a} + \hat{b}X_i - Y_i) = 0$$

$$\sum Y_iX_i = \hat{a}\sum X_i + \hat{b}\sum X_i^2$$

$$\sum Y_iX_i = \hat{a}\sum X_i + \hat{b}\sum X_i^2$$

Equation 25.4 and 25.5 are known as two normal equations. Solving these two normal equations, we get

$$\sum X_i \sum Y_i - n\sum X_i Y_i = \hat{b}\left[\left(\sum X_i\right)^2 n\sum X_i^2\right]$$

$$\Longrightarrow \hat{b} = \frac{n\sum X_i Y_i - (\sum X_i)(\sum Y_i)}{n\sum X_i{}^2 - (\sum X_i)^2} = \frac{Cov(X,Y)}{V(X)} = b_{YX}$$

Putting the value of $\hat{b}$ in either of the normal equation we get

$$\hat{a} = \frac{(\sum X_i{}^2)(\sum Y_i) - (\sum X_i)(\sum X_i Y_i)}{n\sum X_i{}^2 - (\sum X_i)^2}$$

Substituting these values of $\hat{a}$ & $\hat{b}$ from equation (25.7) and (25.6) respectively in equation (25.1) we get required equation of line regression of Y on X.

Dividing both the equation by number of pairs of observation we get

$$\frac{\sum Y_i}{n} = \frac{\sum \hat{a}}{n} + \frac{b\sum X_i}{n} \Longrightarrow \bar{Y} = \hat{a} + \hat{b}\bar{X}$$

This implies that line of best fit passes through the point $\bar{X}, \bar{Y}$ or in other words points $\bar{X}, \bar{Y}$ lies on the line of regression of Y on X. The required equation of the line of regression of Y on X can be written as:

$$(Y - \bar{Y}) = b_{YX}(X - \bar{X}) = \frac{Cov(x,y)}{V(x)}(X - \bar{X})$$

But we know tha $r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \Longrightarrow Cov.(X,Y) = r\sigma_X\sigma_Y$

Substituting the value of Cov. (X,Y) in equation we get

Line of Regression of X on Y

Similarly we can have a line of X on Y i.e., X = a + bY

$$\hat{b}' = b_{XY} = \frac{Cov(X,Y)}{V(Y)} = \frac{r\sigma_Y}{\sigma_X}$$

The required equation of the line of regression of X on Y can be written as :

$$(X-\bar{X}) = b_{XY}(Y-\bar{Y}) = \frac{Cov(X,Y)}{V(X)}(Y-\bar{Y}) = \frac{r\sigma_X}{\sigma_Y}(Y-\bar{Y}) \quad \text{(Eq. 25.11)}$$

From equations (25.9) and (25.11) it is evident that both the lines of regression X on Y and Y on X pass through the point $\bar{X}, \bar{Y}$. Hence $(\bar{X}, \bar{Y})$ is a point of intersection of Y on X and X on Y. The above procedure of fitting of regression equation is illustrated through the following example:

Example 1 : The following data pertains to spoilage of milk (in %)(X) and the temperature ($^{\circ}$C) (Y) of storage of milk in a dairy plant.

Spoilage of Milk (X)	27.3	29.5	26.8	29.5	30.5	29.7	25.6	25.4	24.6	23.6
Temperature($^{\circ}$C) (Y)	33.9	34.6	34.5	36.9	37.1	37.3	28.8	29.6	31.2	30.7

Fit a linear regression for spoilage of milk (in %)(X) on the temperature ($^{\circ}$C) (Y)and vice versa . Also predict the spoilage of milk when temperature is 40°C and value of temperature when spoilage in milk is 35 %.

Solution

$$\bar{X} = \frac{\sum X_i}{n} = \frac{272.5}{10} = 27.25, \bar{Y} = \frac{\sum Y_i}{n} = \frac{334.6}{10} = 33.46$$

Spoilage of Milk (X_i)	Temperature ($^{\circ}$C) (Y_i)	$(X_i-\bar{X})$	$(Y_i-\bar{Y})$	$(X_i-\bar{X})^2$	$(Y_i-\bar{Y})^2$	$(X_i-\bar{X})(Y_i-\bar{Y})$
27.3	33.9	0.05	0.44	0.0025	0.1936	0.022
29.5	34.6	2.25	1.14	5.0625	1.2996	2.565
26.8	34.5	-0.45	1.04	0.2025	1.0816	-0.468
29.5	36.9	2.25	3.44	5.0625	11.8336	7.7400
30.5	37.1	3.25	3.64	10.5625	13.2496	11.830
29.7	37.3	2.45	3.84	6.0025	14.7456	9.408
25.6	28.8	-1.65	-4.66	2.7225	21.7156	7.689
25.4	29.6	-1.85	-3.86	3.4225	14.8996	7.141
24.6	31.2	-2.65	-2.26	7.0225	5.1076	5.989
23.6	30.7	-3.65	-2.76	13.3225	7.6176	10.074
Total 272.5	334.6			53.3850	91.7440	61.990

Regression coefficient of Yon X (b_{YX}) :

$$b_{YX} = \frac{\frac{1}{n}\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\frac{1}{n}\sum(X_i-\bar{X})^2.} = \frac{61.990/10}{53.385/10} = 1.1612$$

Regression coefficient of X on Y (b_{XY}) :

$$b_{XY} = \frac{\frac{1}{n}\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\frac{1}{n}\sum(Y_i - \overline{Y})^2.} = \frac{61.990/10}{91.7440/10} = 0.6757$$

Regression equation of Y on X i.e., regression line of temperature on spoilage of milk is

$$(Y - \overline{Y}) = b_{YX}(X - \overline{X}) \Longrightarrow (Y - 33.46) = 1.1612(X - 27.25)$$

The required equation of the line of regression of Y on X i.e. regression equation of temperature on spoilage in milk is Y=1.8173+1.1612X. To predict the value of temperature when spoilage in milk is 35%, we put X=35 in the above equation so we get Y=42.4593. It means when spoilage in milk is 35% the temperature will be 42.4593 $^{\circ}$C.

Regression equation of X on Y i.e., regression line of spoilage of milk on temperature is

$$(X - \overline{X}) = b_{XY}(Y - \overline{Y}) \Longrightarrow (X - 27.25) = 0.6757(X - 33.46)$$

The required equation of the line of regression of X on Y i.e. regression equation of spoilage of milk on temperature is X=4.6416+0.6757Y. To predict the value of spoilage in milk when temperature is 40°C, we put Y=40 in the above equation so we get X=31.6693. It means when temperature is 40°C then spoilage in milk will be 31.6693 %.

25.6.2 Why there are two regression lines

The line of regression of Y on X ($Y = a + b_{YX}X$) is used to estimate/predict the value of Y for any given value of X i.e. Y is a dependent variable and X is an independent/explanatory variable. The estimates so obtained will be best in the sense that it will have the minimum possible error as defined by the principle of the least squares. In order to predict or estimate X for any given value of Y we use the regression equation of X on Y ($X = a + b_{XY}Y$) which is obtained by minimizing sum of squares due to error of estimates in X. Here X is dependent variable and Y is independent/explanatory variable. Two regression equations are not reversible or interchangeable. Regression equation of Y on X is obtained by minimizing the sum of square of errors parallel to the Y-axis, while the regression equation of X on Y is obtained by minimizing the sum of squares of error parallel to X-axis. In a particular case of perfect correlation, positive or negative i.e., r = 1, the equation of line of regression of Y on X becomes:

$$(Y - \overline{Y}) = \pm\frac{\sigma_Y}{\sigma_X}(X - \overline{X}) \Longrightarrow \frac{(Y - \overline{Y})}{\sigma_Y} = \pm\frac{(X - \overline{X})}{\sigma_X}$$

Similarly , the equation of the line of regression of X on Y becomes:

$$(X-\bar{X}) = \pm\frac{\sigma_X}{\sigma_Y}(Y-\bar{Y}) \Rightarrow \frac{(X-\bar{X})}{\sigma_X} = \pm\frac{(Y-\bar{Y})}{\sigma_Y}$$

Above two equations are same. Hence, in case of perfect correlation (r = 1) both the lines coincide. Therefore, in general we always have two lines of regression except in the particular case of perfect correlation when both the lines coincide and we get only one line.

25.6.3 Angle between Two Regression Lines:

The angle between two regression lines is given by

$$\theta = \tan^{-1}\left\{\frac{\sigma_X\sigma_Y}{\sigma_X^2+\sigma_Y^2}\left(\frac{r^2-1}{r}\right)\right\}$$

If r=1, then $\theta = \tan^{-1}(0) \Rightarrow \theta = 0$ or π i.e., two lines are either coincident or they are parallel to each other. But since both the lines of regression intersect at the point $(\bar{X},\bar{Y})$, they cannot be parallel. Hence in case of perfect correlation, positive or negative, the two lines of regression coincide. If r=0, then $\theta = \tan^{-1}(\infty) \Rightarrow \theta = \pi/2$ i.e., if the variables are uncorrelated, the two lines of regression become perpendicular to each other. Hence, for higher degree of correlation between the variables, the angle between the lines is smaller i.e., the two lines of regression are nearer to each other. On the other hand, the angle between the lines increases as the lines of regression move apart and the value of correlation decreases.

25.7 Coefficients of Regression

Just as there are two regression equations, similarity there are two regression coefficients. Regression coefficients measure the average change in the value of one variable for a unit change in the value of another variable. Regression coefficient, infact, represents the slope of a regression line. For two variables X and Y, there are two regression coefficients which are given as follows

25.7.1 Regression Coefficient of Y on X

This coefficient shows that with a unit change in the value of X variable, what will be the average change in the value of Y variable. This is represented by b_{YX} .

$$b_{YX} = \frac{Cov(x,y)}{V(x)} = \frac{r\sigma_Y}{\sigma_X}$$

25.7.2 Regression Coefficient of X on Y

This coefficient shows that with a unit change in the value of Y variable, what will be the average change in the value of Y variable. This is represented by b_{XY} .

$$b_{XY} = \frac{Cov(x,y)}{V(y)} = \frac{r\sigma_X}{\sigma_Y}$$

25.8 Properties of Regression Coefficient

The important properties of the regression coefficients are:

1) The correlation coefficient is the geometric mean between the regression coefficients i.e., $r^2 = b_{YX}.b_{XY} \Rightarrow r = \sqrt{b_{YX}.b_{XY}}$
2) Both the regression coefficients must have the same algebraic signs. This means that either both regression coefficients will be positive or negative i.e., when one regression coefficient is negative, the other would also be negative and if one regression coefficient is positive, the other would be also positive. It is never possible that one regression coefficient is negative while the other is positive.
3) The coefficient of correlation will have the same sign as that of regression coefficients.
4) If one of the regression coefficients is greater than unity (one), the other must be less than unity.
5) The arithmetic mean of the regression coefficients is greater than the correlation coefficient.
6) Regression coefficients are independent of change of origin but not of scale which is illustrated below.

If X_i and Y_i are the given variables and these variables are transformed to new variables U_i and V_i by the change of origin and scale

$$U_i = \frac{X_i - A}{h} \Rightarrow X_i = A + hU_i \quad \text{and} \quad V_i = \frac{Y_i - A}{K} \Rightarrow Y_i = B + kV_i$$

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

$$X_i - \bar{X} = h(U_i - \bar{U}) \qquad Y_i - Y = k(V_i - \bar{V})$$

$$b_{YX} = \frac{\frac{1}{n}\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum(X_i - \bar{X})^2} = \frac{hk\frac{1}{n}\sum(U_i - \bar{U})\cdot(V_i - \bar{V})}{\frac{1}{n}\sum h^2(U_i - \bar{U})^2}$$

$$= \frac{hk\frac{1}{n}\sum(U_i - \bar{U})\cdot(V_i - \bar{V})}{h^2\frac{1}{n}\sum(U_i - \bar{U})^2} = \frac{k}{h}\frac{\frac{1}{n}\sum(U_i - \bar{U})(V_i - \bar{V})}{\frac{1}{n}\sum(U_i - \bar{U})^2}$$

$$b_{YX} = \frac{k}{h}\frac{n\sum U_i V_i - (\sum U_i)(\sum V_i)}{n\sum U_i^2 - (\sum U_i)^2} = \frac{k}{h}\frac{Cov\,(U_i, V_i)}{\sigma_{U_i} \cdot \sigma_{V_i}} = \frac{k}{h} b_{U_i V_i}$$

Hence regression coefficient is independent of change of origin and but not of scale. The procedure is illustrated in the following example by taking the data from example 25.1.

Example 2: Solve example 1 by change of origin and scale:

Solution: Let us define $U_i = X_i$ 29.7 and $V_i = Y_i$ 34.5

(X)	(Y)	$U_i = X_i - 29.7$	$V_i = Y_i - 34.5$	U_i^2	V_i^2	$U_i V_i$
27.3	33.9	-2.4	-0.7	5.76	0.49	1.68
29.5	34.6	-0.2	0	0.04	0	0
26.8	34.5	-2.9	-0.1	8.41	0.01	0.29
29.5	36.9	-0.2	2.3	0.04	5.29	-0.46
30.5	37.1	0.8	2.5	0.64	6.25	2
29.7	37.3	0	2.7	0	7.29	0
25.6	28.8	-4.1	-5.8	16.81	33.64	23.78
25.4	29.6	-4.3	-5	18.49	25	21.5
24.6	31.2	-5.1	-3.4	26.01	11.56	17.34
23.6	30.7	-6.1	-3.9	37.21	15.21	23.79
Total 272.5	334.6	-24.5	-11.4	113.41	104.74	89.92

$$b_{YX} = \frac{k}{h}\frac{n\sum U_i V_i - (\sum U_i)(\sum V_i)}{n\sum U_i^2 - (\sum U_i)^2} = \frac{10x89.92 - (-24.5)(-11.4)}{10x113.41 - (-24.5)^2} = \frac{619.9}{533.85} = 1.1612$$

$$b_{XY} = \frac{h}{k}\frac{n\sum U_i V_i - (\sum U_i)(\sum V_i)}{n\sum V_i^2 - (\sum V_i)^2} = \frac{10x89.92 - (-24.5)(-11.4)}{10x104.74 - (-11.4)^2} = \frac{619.9}{917.44} = 0.6757$$

These values are same as obtained in example 25.1, which shows that regression coefficients are independent of change of origin.

$$r = \sqrt{b_{YX} \times b_{XY}} = \sqrt{1.1612 \times 0.6757} = 0.8858$$

25.9 Coefficient of Determination

The total variation in the dependent variable Y can be split into two:

a) **Explained variation**: The variation in Y which is explained by the variation in X is known as explained variation in Y

b) **Unexplained variation**: The variation in Y which is unexplained by the variation in variable X and is due to some other factors (a variable) is called unexplained variation in Y.

Symbolically,

Total variation in Y = Explained variation in Y + Unexplained variation in Y

$$\sum(Y-\bar{Y})^2=\sum(\hat{Y}-\bar{Y})^2+\sum(Y-\hat{Y})^2$$

Where $\hat{Y}$ = computed (or estimated) value of Y on the basis of regression equation

$\bar{Y}$ = Mean value of Y series

Y = Original value of Y series

A similar relationship we may have for X variable (Dependent) in terms of Y: $\sum(X-\bar{X})^2=\sum(\hat{X}-\bar{X})^2+\sum(X-\hat{X})^2$

Coefficient of Determination: Based on above expression, the coefficient of determination (r^2) is defined as the ratio of the explained variation to total variation i.e.,

$$r^2=\frac{\text{Explained variation}}{\text{Total variation}}=\frac{\sum(\hat{Y}-\bar{Y})^2}{\sum(Y-\bar{Y})^2}$$

It is clear that the objective of coefficient of determination is to determine the percentage variation in Y which is explained by variation in X. For example, let us suppose that the correlation coefficient between X and Y is +0.8, then coefficient of determination (r^2) = $(.8)^2$ = .64. It means that 64per cent variation in Y is due to variation in X and 36 per cent variation is due to other factors. Thus, explained variations and unexplained variation are 64 and 36 per cent respectively.

Coefficient of Non-Determination: The proportion of unexplained variation to total variation is termed as coefficient of non-determination. It is denoted by k^2,where $k^2 = 1 - r^2$. It is also written as:

$$k^2=\frac{\text{Unexplained variation}}{\text{Total variation}}=1-r^2$$

The square root of k^2 is termed as coefficient of alientation i.e., $k=\sqrt{k^2}=\sqrt{1-r^2}$

Standard error of estimates: Standard error of estimates of Y on X and that of X on Y can also be calculated as:

$$S_{Y.X} = \sqrt{\frac{\sum(Y-\hat{Y})^2}{N}} = \sqrt{\frac{\text{Unexplained variation in Y}}{N}}$$

$$S_{X.Y} = \sqrt{\frac{\sum(X-\hat{X})^2}{N}} = \sqrt{\frac{\text{Unexplained variation in X}}{N}}$$

Example 3: Compute Coefficient of Determination in example 1

Solution

(X)	(Y)	$\hat{Y}$ = 1.8173+1.1612X	$(Y-\bar{Y})^2$	$(\hat{Y}-\bar{Y})^2$	$(Y-\hat{Y})^2$
27.3	33.9	33.5181	0.1936	0.0034	0.1459
29.5	34.6	36.0727	1.2996	6.8262	2.1688
26.8	34.5	32.9375	1.0816	0.2730	2.4415
29.5	36.9	36.0727	11.8336	6.8262	0.6844
30.5	37.1	37.2339	13.2496	14.2423	0.0179
29.7	37.3	36.3049	14.7456	8.0937	0.9901
25.6	28.8	31.5440	21.7156	3.6710	7.5296
25.4	29.6	31.3118	14.8996	4.6148	2.9302
24.6	31.2	30.3828	5.1076	9.4690	0.6678
23.6	30.7	29.2216	7.6176	17.9639	2.1856
Total 72.5	334.6		91.7440	71.9836	19.7620

From above table the different variation are as

Total variation: $\sum(Y-\bar{Y})^2 = 91.7440$

Explained variation in $Y = \sum(\hat{Y}-\bar{Y})^2 = 71.9836$

Unexplained variation in $Y = \sum(Y-\hat{Y})^2 = 19.7620$

Coefficient of determination (r^2)

$$= \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\sum(\hat{Y}-\bar{Y})^2}{\sum(Y-\bar{Y})^2} = \frac{71.9836}{91.7440} = 0.7846$$

It means that 78.46 percent variations in Y are due to variation in X and 21.54 per cent variation is due to other factors. Moreover, coefficient of determination is square of correlation coefficient i.e., $(0.8858)^2=0.7846$ which is same as computed above.