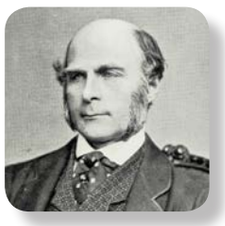
CHAPTER

5 REGRESSION ANALYSIS



Francis Galton

Francis Galton (1822-1911) was born in a wealthy family. The youngest of nine children, he appeared as an intelligent child. Galton's progress in education was not smooth. He dabbled in medicine and then studied Mathematics at Cambridge. In fact he subsequently freely acknowledged his weakness in formal Mathematics, but this weakness was compensated by an exceptional ability to understand the meaning of data. Many statistical terms, which are in current usage were coined by Galton. For example, correlation is due to him, as is regression, and he was the originator of terms and concepts such as quartile, decile and percentile, and of the use of median as the midpoint of a distribution.

The concept of regression comes from genetics and was popularized by Sir Francis Galton during the late 19th century with the publication of regression towards mediocrity in hereditary stature. Galton observed that extreme characteristics (e.g., height) in parents are not passed on completely to their offspring. An examination of publications of Sir Francis Galton and Karl Pearson revealed that Galton's work on inherited characteristics of sweet peas led to the initial conceptualization of linear regression. Subsequent efforts by Galton and Pearson brought many techniques of multiple regression and the product-moment correlation coefficient.

LEARNING OBJECTIVES

The student will be able to

- ❖ know the concept of regression, its types and their uses.
- ❖ fit best line of regression by applying the method of least squares.
- ❖ calculate the regression coefficient and interpret the same.
- ❖ know the uses of regression coefficients.
- ❖ distinguish between correlation analysis and regression analysis.

2PUID

Introduction

The correlation coefficient is an useful *statistical tool for describing the type (positive or negative or uncorrelated) and intensity of linear relationship* (such as moderately or highly) between two variables. But it fails to give a *mathematical functional* relationship for prediction purposes. Regression analysis is a vital statistical method for obtaining functional relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one to understand how the typical value of the dependent variable (or 'response variable') changes when any one of the independent variables (regressor(s) or predictor(s)) is varied, while the other independent variables are held fixed. It helps to determine the impact of changes in the value(s) of the the independent variable(s) upon changes in the value of the dependent variable. Regression analysis is widely used for prediction.

5.1 DEFINITION

Regression analysis is a statistical method of determining the mathematical functional relationship connecting independent variable(s) and a dependent variable.

Types of 'Regression'

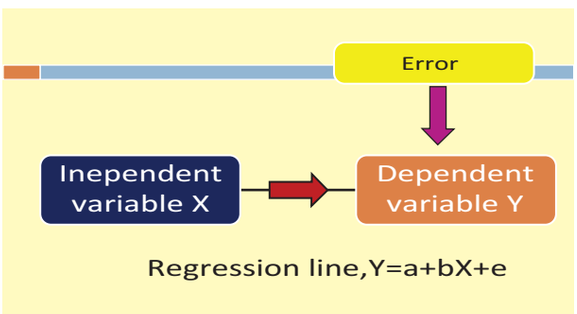
Based on the kind of relationship between the dependent variable and the set of independent variable(s), there arises two broad categories of regression *viz.*, linear regression and non-linear regression.

If the relationship is linear and there is only one independent variable, then the regression is called as simple linear regression. On the other hand, if the relationship is linear and the number of independent variables is two or more, then the regression is called as multiple linear regression. If the relationship between the dependent variable and the independent variable(s) is not linear, then the regression is called as non-linear regression.

5.1.1 Simple Linear Regression

It is one of the most widely known modeling techniques. In this technique, the dependent variable is continuous, independent variable(s) can be continuous or discrete and nature of relationship is linear. This relationship can be expressed using a straight line equation (linear regression) that best approximates all the individual data points.

Simple linear regression establishes a relationship between a **dependent variable (Y)** and one **independent variable (X)** using a **best fitted straight line** (also known as regression line).



NOTE

There are many reasons for the presence of the error term in the linear regression model. It is also known as measurement error. In some situations, it indicates the presence of several variables other than the present set of regressors.

The general form of the simple linear regression equation is $Y = a + bX + e$, where 'X' is independent variable, 'Y' is dependent variable, a' is intercept, 'b' is slope of the line and 'e' is error term. This equation can be used to estimate the value of response variable (Y) based on the given values of the predictor variable (X) within its domain.

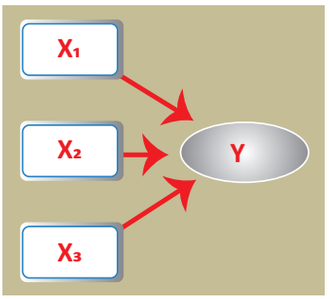
5.1.2 Multiple Linear Regression

In the case of several independent variables, regression analysis also allows us to compare the effects of independent variables measured on different scales, such as the effect of price changes and the number of promotional activities.

Multiple linear regression uses two or more independent variables to estimate the value(s) of the response variable (Y).

The general form of the multiple linear regression equation is

$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_tX_t + e$



Here, Y represents the dependent (response) variable, X_i represents the i^{th} independent variable (regressor), a and b_i are the regression coefficients and e is the error term.

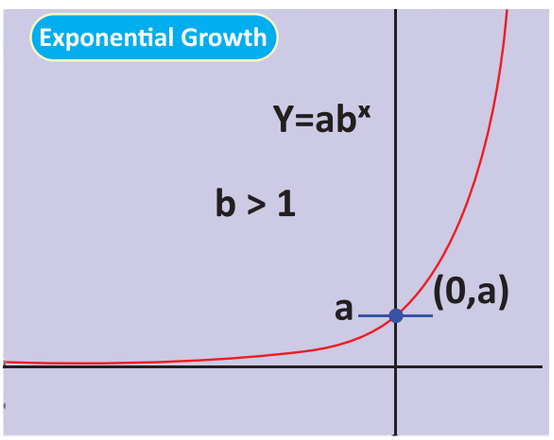
Suppose that price of a product (Y) depends mainly upon three promotional activities such as discount (X_1), instalment scheme (X_2) and free installation (X_3). If the price of the product has linear relationship with each promotional activity, then the relationship among Y and X_1, X_2 and X_3 may be expressed using the above general form as

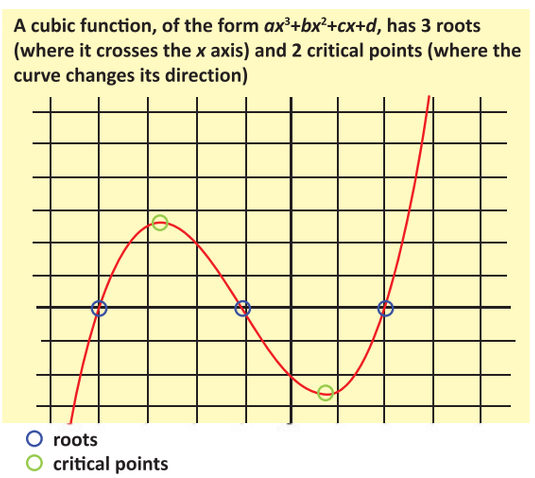
$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + e\,.$$

These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building regression models for predictive purposes.

5.1.3 Non-Linear Regression

If the regression is not linear and is in some other form, then the regression is said to be non-linear regression. Some of the non-linear relationships are displayed below.





5.2 USES OF REGRESSION

Benefits of using regression analysis are as follows:

1. It indicates the **significant mathematical** relationship between independent variable (X) and dependent variable(Y). *(i.e)* Model construction
2. It indicates the **strength of impact (b)** of independent variable on a dependent variable.
3. It is used to estimate (interpolate) the value of the response variable for different values of the independent variable from its range in the given data. It means that extrapolation of the dependent variable is not generally permissible.

NOTE

Multiple linear regression and Curvilinear relationships (non-linear regression) are out of the syllabus. Basic information about them are given here, for enhancing the knowledge.

4. In the case of several independent variables, regression analysis is a way of mathematically sorting out which of those variables indeed have an impact (It answers the questions: Which independent variable matters most? Which can we ignore? How do those independent variables interact with each other?

5.3 WHY ARE THERE TWO REGRESSION LINES?

There may exist two regression lines in certain circumstances. When the variables X and Y are interchangeable with related to causal effects, one can consider X as independent variable and Y as dependent variable (or) Y as independent variable and X as dependent variable. As the result, we have (1) **the regression line of Y on X** and (2) **the regression line of X on Y**.

Both are valid regression lines. But we must judicially select the one regression equation which is suitable to the given environment.

Note: If, X only causes Y, then there is only one regression line, of Y on X.

5.3.1 Simple Linear Regression

In the general form of the simple linear regression equation of Y on X

$$Y = a + bX + e$$

the constants 'a' and 'b' are generally called as the regression coefficients.

The coefficient 'b' represents the rate of change in the value of the mean of Y due to every unit change in the value of X. When the range of X includes '0', then the intercept 'a' is $E(Y|X = 0)$. If the range of X does not include '0', then 'a' does not have practical interpretation.

If (x_i, y_i), $i = 1, 2, \ldots, n$ is a set of n-pairs of observations made on (X, Y), then fitting of the above regression equation means finding the estimates '$\hat{a}$' and '$\hat{b}$' for 'a' and 'b' respectively. These estimates are determined based on the following general assumptions:

i) the relationship between Y and X is linear (approximately).

ii) the error term 'e' is a random variable with mean zero.

iii) the error term 'e' has constant variance.

There are other assumptions on 'e', which are not required at this level of study.

Before going for further study, the following points are to be kept in mind.

- Both the independent and dependent variables must be measured at the interval scale.
- There must be **linear relationship** between independent and dependent variables.
- Linear Regression is very sensitive to **Outliers** (extreme observations). It can affect the regression line extremely and eventually the estimated values of Y too.

Meaning of line of "best fit"

Based on the assumption (ii), the response variable Y is also a random variable with mean

$$E(Y|X=x) = a + bx$$

In regression analysis, the main objective is finding the line of best fit, which provides the fitted equation of Y on X.

The line of 'best fit' is the line (straight line equation) which minimizes the error in the estimation of the dependent variable Y, for any specified value of the independent variable X from its range.

The regression equation $E(Y|X=x) = a + bx$ represents a family of straight lines for different values of the coefficients 'a' and 'b'. The problem is to determine the estimates of 'a' and 'b' by minimizing the error in the estimation of Y so that the line is a best fit. This necessitates to find the suitable values of the estimates of 'a' and 'b'.

5.4 METHOD OF LEAST SQUARES

In most of the cases, the data points do not fall on a straight line (not highly correlated), thus leading to a possibility of depicting the relationship between the two variables using several different lines. Selection of each line may lead to a situation where the line will be closer to some points and farther from other points. We cannot decide which line can provide best fit to the data.

Method of least squares can be used to determine the line of best fit in such cases. It determines the line of best fit for given observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line.

5.4.1 Method of Least Squares

To obtain the estimates of the coefficients 'a' and 'b', the least squares method minimizes the sum of squares of residuals. The residual for the i^{th} data point e_i is defined as the difference between the observed value of the response variable, y_i, and the estimate of the response variable, $\hat{y}_i$, and is identified as the error associated with the data. *i.e.*, $e_i = y_i - \hat{y}_i$, $i = 1, 2, \ldots, n$.

The method of least squares helps us to find the values of unknowns 'a' and 'b' in such a way that the following two conditions are satisfied:

- Sum of the residuals is zero. That is $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$.
- Sum of the squares of the residuals $E(a,b) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the least.

5.4.2 Fitting of Simple Linear Regression Equation

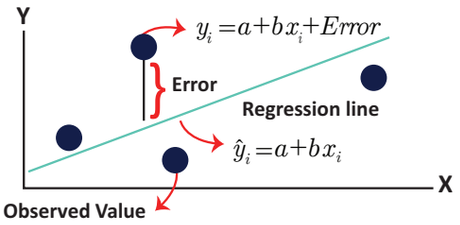
The method of least squares can be applied to determine the estimates of 'a' and 'b' in the simple linear regression equation using the given data (x_1,y_1), (x_2,y_2), ..., (x_n,y_n) by minimizing

$$E(a,b) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

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

Here, $\hat{y}_i = a + bx_i$ is the expected (estimated) value of the response variable for given x_i.

Simple Linear Regression Model



It is obvious that if the expected value ($\hat{y}_i$) is close to the observed value (y_i), the residual will be small. Since the magnitude of the residual is determined by the values of '*a*' and *'b'*, estimates of these coefficients are obtained by minimizing the sum of the squared residuals, *E(a,b).*

Differentiation of *E(a,b)* with respect to *'a'* and *'b'* and equating them to zero constitute a set of two equations as described below:

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

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

These give

$$na + b\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}y_i$$

$$a\sum_{i=1}^{n}x_i + b\sum_{i=1}^{n}x_i^2 = \sum_{i=1}^{n}x_iy_i$$

These equations are popularly known as **normal equations.** Solving these equations for '*a*' and *'b'* yield the estimates $\hat{a}$ and $\hat{b}$.

$$\hat{a} = \overline{y} - \hat{b}\overline{x}$$

and

$$\hat{b} = \frac{\frac{1}{n}\sum_{i=1}^{n}x_iy_i - \overline{x}\,\overline{y}}{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2}$$

It may be seen that in the estimate of '*b*', the numerator and denominator are respectively the sample covariance between *X* and *Y,* and the sample variance of *X*. Hence, the estimate of *'b'* may be expressed as

$$\hat{b} = \frac{Cov(X,Y)}{V(X)}$$

Further, it may be noted that for notational convenience the denominator of $\hat{b}$ above is mentioned as variance of *X*. But, the definition of sample variance remains valid as defined in Chapter I, that is, $\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}^2\right)$.

From Chapter 4, the above estimate can be expressed using, r_{XY}, Pearson's coefficient of the simple correlation between *X* and *Y,* as

$$\hat{b} = r_{XY}\frac{SD(Y)}{SD(X)}.$$

Important Considerations in the Use of Regression Equation:

1. Regression equation exhibits only the relationship between the respective two variables. Cause and effect study shall not be carried out using regression analysis.
2. The regression equation is fitted to the given values of the independent variable. Hence, the fitted equation can be used for prediction purpose corresponding to the values of the regressor within its range. Interpolation of values of the response variable may be done corresponding to the values of the regressor from its range only. The results obtained from extrapolation work could not be interpreted.

Example 5.1

Construct the simple linear regression equation of Y on X if $n = 7, \sum_{i=1}^{n} x_i = 113$, $\sum_{i=1}^{n} x_i^2 = 1983$, $\sum_{i=1}^{n} y_i = 182$ and $\sum_{i=1}^{n} x_i y_i = 3186$.

Solution:

The simple linear regression equation of Y on X to be fitted for given data is of the form

$$\hat{Y} = a + bx \tag{1}$$

The values of 'a' and 'b' have to be estimated from the sample data solving the following normal equations.

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{2}$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \tag{3}$$

Substituting the given sample information in (2) and (3), the above equations can be expressed as

$$7\,a + 113\,b = 182 \tag{4}$$

$$113\,a + 1983\,b = 3186 \tag{5}$$

$(4) \times 113 \Rightarrow 791\,a + 12769\,b = 20566$

$(5) \times 7 \Rightarrow 791\,a + 13881\,b = 22302$

$$(-) \quad (-) \quad (-)$$

$$-1112\,b = -1736$$

$$\Rightarrow b = \frac{1736}{1112} = 1.56$$

$$b = 1.56$$

Substituting this in (4) it follows that,

$$7\,a + 113 \times 1.56 = 182$$

$$7\,a + 176.28 = 182$$

$$7\,a = 182 - 176.28$$

$$= 5.72$$

Hence, $a = 0.82$

Example 5.2

Number of man-hours and the corresponding productivity (in units) are furnished below. Fit a simple linear regression equation $\hat{Y} = a + bx$ applying the method of least squares.

Man-hours	3.6	4.8	7.2	6.9	10.7	6.1	7.9	9.5	5.4
Productivity (in units)	9.3	10.2	11.5	12	18.6	13.2	10.8	22.7	12.7

Solution:

The simple linear regression equation to be fitted for the given data is

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

Here, the estimates of a and b can be calculated using their least squares estimates

$$\hat{a} = \overline{y} - \hat{b}\overline{x}$$

i.e.,

$$\hat{a} = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{b}\frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{b} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - (\overline{x} \times \overline{y})}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2}$$

or equivalently $$\hat{b} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

From the given data, the following calculations are made with n=9

Man-hours x_i	Productivity y_i	x_i^2	$x_i y_i$
3.6	9.3	12.96	33.48
4.8	10.2	23.04	48.96
7.2	11.5	51.84	82.8
6.9	12	47.61	82.8
10.7	18.6	114.49	199.02
6.1	13.2	37.21	80.52
7.9	10.8	62.41	85.32
9.5	22.7	90.25	215.65
5.4	12.7	29.16	66.42
$\sum_{i=1}^{9} x_i = 62.1$	$\sum_{i=1}^{9} y_i = 121$	$\sum_{i=1}^{9} x_i^2 = 468.97$	$\sum_{i=1}^{9} x_i y_i = 894.97$

Substituting the column totals in the respective places in the of the estimates $\hat{a}$ and $\hat{b}$, their values can be calculated as follows:

$$\hat{b} = \frac{(9\times 894.97)-(62.1\times 121)}{(9\times 468.97)-(62.1)^2}$$

$$= \frac{8054.73-7514}{4220.73-3856.41}$$

$$= \frac{540.73}{364.32}$$

Thus, $\hat{b} = 1.48$.

Now $\hat{a}$ can be calculated using $\hat{b}$ as

$$\hat{a} = \frac{121}{9} - \left(1.48\times \frac{62.1}{9}\right)$$

= 13.40 – 10.21

Hence, $\hat{a}$ = 3.19

Therefore, the required simple linear regression equation fitted to the given data is

$$\hat{Y} = 3.19 + 1.48x$$

It should be noted that the value of Y can be estimated using the above fitted equation for the values of x in its range *i.e.*, 3.6 to 10.7.

In the estimated simple linear regression equation of Y on X

$$\hat{Y} = \hat{a} + \hat{b}x$$

we can substitute the estimate $\hat{a} = \overline{y} - \hat{b}\overline{x}$. Then, the regression equation will become as

$$\hat{Y} = \overline{y} - \hat{b}\overline{x} + \hat{b}x$$

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

It shows that the simple linear regression equation of Y on X has the slope $\hat{b}$ and the corresponding straight line passes through the point of averages $(\overline{x}, \overline{y})$. The above representation of straight line is popularly known in the field of Coordinate Geometry as 'Slope-Point form'. The above form can be applied in fitting the regression equation for given regression coefficient $\hat{b}$ and the averages $\overline{x}$ and $\overline{y}$.

As mentioned in Section 5.3, there may be two simple linear regression equations for each X and Y. Since the regression coefficients of these regression equations are different, it is essential to distinguish the coefficients with different symbols. The regression coefficient of the simple linear regression equation of Y on X may be denoted as b_{YX} and the regression coefficient of the simple linear regression equation of X on Y may be denoted as b_{XY}.

Using the same argument for fitting the regression equation of Y on X, we have the simple linear regression equation of X on Y with best fit as

$$\hat{X} = \hat{c} + b_{XY} y$$

where $\hat{c} = \overline{x} - b_{XY}\overline{y}$

$$b_{XY} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}}{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2}$$

The slope-point form of this equation is

$$\hat{X} - \overline{x} = b_{XY}(y - \overline{y}).$$

Also, the relationship between the Karl Pearson's coefficient of correlation and the regression coefficient are

$$b_{XX} = r_{XY}\frac{SD(X)}{SD(Y)} \text{ and } b_{YX} = r_{XY}\frac{SD(Y)}{SD(X)}.$$

5.5 PROPERTIES OF REGRESSION COEFFICIENTS

1. Correlation coefficient is the geometric mean between the regression coefficients.

$$r_{XY} = \sqrt{b_{XY} \times b_{YX}}$$

2. It is clear from the property 1, both regression coefficients must have the same sign. *i.e.,* either they will positive or negative.
3. If one of the regression coefficients is greater than unity, the other must be less than unity.
4. The correlation coefficient will have the same sign as that of the regression coefficients.
5. Arithmetic mean of the regression coefficients is greater than the correlation coefficient.

$$\frac{b_{XY} + b_{YX}}{2} \geq r_{XY}$$

6. Regression coefficients are independent of the change of origin but not of scale.

Properties of regression equation

1. If $r = 0$, the variables are uncorrelated, the lines of regression become perpendicular to each other.
2. If $r = 1$, the two lines of regression either coincide or parallel to each other.
3. Angle between the two regression lines is $\theta = \tan^{-1}\left(\frac{m_1 - m_2}{1 + m_1 m_2}\right)$ where m_1 and m_2 are the slopes of regression lines X on Y and Y on X respectively.
4. The angle between the regression lines indicates the degree of dependence between the variable.
5. Regression equations intersect at $(\overline{X}, \overline{Y})$

Example 5.3

Calculate the regression equation of X on Y from the data given below, taking deviations from actual means of X and Y.

x	12	14	15	14	18	17
y	42	40	45	47	39	45

Estimate the likely demand when the $X = 25$.

Solution:

	x_i	$u_i = x_i - 15$	u_i^2	y_i	$v_i = y_i - 43$	v_i^2	$u_i v_i$
	12	-3	9	42	-1	1	3
	14	-1	1	40	-3	9	3
	15	-0	0	45	2	4	0
	14	-1	1	47	4	16	-4
	18	3	9	39	-4	16	-12
	17	2	4	45	2	4	4
Total	90	0	24	258	0	50	-6

$$\bar{x} = \sum_{i=1}^{6} x_i / 6 = \frac{90}{6} = 15$$

$$\bar{y} = \sum_{i=1}^{6} y_i / 5 = \frac{258}{6} = 43$$

The regression line of U on V is computed as under

$$\hat{b}_{UV} = \frac{n\sum_{i=1}^{n} u_i v_i - \sum_{i=1}^{n} u_i \sum_{i=1}^{n} v_i}{n\sum_{i=1}^{n} v_i^2 - \left(\sum_{i=1}^{n} v_i\right)^2} = \frac{6(-6)}{6 \times 50} = -0.12$$

$$\hat{a} = \bar{u} - \hat{b}_{UV}\,\bar{v} = 0$$

Hence, the regression line of U on V is $U = \hat{b}_{UV}\, v + \hat{a} = -0.12v$

Thus, the regression line of X on Y is $(Y-43) = -0.25(x-15)$

When $x = 25$, $y - 43 = -0.25\ (25-15)$

$$y = 40.5$$

Important Note: If $\overline{X}$, $\overline{Y}$ are not integers then the above method is tedious and time consuming to calculate b_{YX} and b_{XY}. The following modified formulae are easy for calculation.

$$b_{YX} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b_{XY} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}$$

Example 5.4

The following data gives the experience of machine operators and their performance ratings as given by the number of good parts turned out per 50 pieces.

Operators	1	2	3	4	5	6	7	8
Experience (X)	8	11	7	10	12	5	4	6
Ratings (Y)	11	30	25	44	38	25	20	27

Obtain the regression equations and estimate the ratings corresponding to the experience x=15.

Solution:

	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
	8	11	88	64	121
	11	30	330	121	900
	7	25	175	49	625
	10	44	440	100	1936
	12	38	456	144	1444
	5	25	125	25	625
	4	20	80	16	400
	6	27	162	36	729
Total	63	220	1856	555	6780

Regression equation of Y on X,

$$\hat{Y} - \overline{y} = b_{YX}\left(x - \overline{x}\right)$$

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{63}{8} = 7.875$$

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{220}{8} = 27.5$$

The above two means are in decimal places so for the simplicity we use this formula to compute b_{YX}.

$$b_{YX} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{8\times 1856 - 63\times 220}{8\times 555 - 63\times 63}$$

$$= \frac{14848 - 13860}{4440 - 3969}$$

$$= \frac{988}{471}$$

$$b_{YX} = 2.098$$

The regression equation of Y on X,

$$\hat{Y} - \bar{y} = b_{YX}\left(x - \bar{x}\right)$$

$$\hat{Y} - 27.5 = 2.098\ (x - 7.875)$$

$$\hat{Y} - 27.5 = 2.098\ x - 16.52$$

$$\hat{Y} = 2.098x + 10.98$$

When $x = 15$,

$$\hat{Y} = 2.098 \times 15 + 10.98$$

$$\hat{Y} = 31.47 + 10.98$$

$$= 42.45$$

Regression equation of X on Y,

$$\hat{X} - \bar{x} = b_{XY}\left(y - \bar{y}\right)$$

$$b_{XY} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}$$

$$= \frac{8\times 1856 - 63\times 220}{8\times 6780 - 220\times 220}$$

$$= \frac{14848 - 13860}{54240 - 48400}$$

$$= \frac{988}{5840}$$

$$b_{XY} = 0.169$$

The regression equation of X on Y,

$$\hat{X} - 7.875 = 0.169\,(y - 27.5)$$

$$\hat{X} - 7.875 = 0.169y - 0.169 \times 27.5$$

$$\hat{X} = 0.169y + 3.2275$$

Example 5.5

The random sample of 5 school students is selected and their marks in statistics and accountancy are found to be

Statistics	85	60	73	40	90
Accountancy	93	75	65	50	80

Find the two regression lines.

Solution:

The two regression lines are:

Regression equation of Y on X,

$$\hat{Y} - \bar{y} = b_{YX}\left(x - \bar{x}\right)$$

Regression equation of X on Y,

$$\hat{X} - \bar{x} = b_{XY}\left(y - \bar{y}\right)$$

	x_i	y_i	$u_i = x_i - A$ $= x_i - 60$	$v_i = x_i - B$ $= x_i - 75$	$u_i v_i$	u_i^2	y_i^2
	85	93	25	18	450	625	324
	60 A	75 B	0	0	0	0	0
	73	65	13	–10	–130	169	100
	40	50	–20	–25	500	400	625
	90	80	30	5	150	900	25
Total	348	363	48	12	970	2094	1074

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{348}{5} = 69.6$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{363}{5} = 72.6$$

Since the mean values are in decimals format not as integers and numbers are big, we take origins for x and y and then solve the problem.

Regression equation of Y on X,

$$\hat{Y} - \bar{y} = b_{YX}\left(x - \bar{x}\right)$$

Calculation of b_{YX}

$$b_{YX} = b_{VU}\frac{n\sum_{i=1}^{n}u_i v_i - \sum_{i=1}^{n}u_i \sum_{i=1}^{n}v_i}{n\sum_{i=1}^{n}u_i^2 - \left(\sum_{i=1}^{n}u_i\right)^2}$$

$$= \frac{5\times 970 - 48\times 9(-12)}{5\times 2094 - (48)^2}$$

$$= \frac{4850 + 576}{10470 - 2304}$$

$$= \frac{5426}{8126} = 0.664$$

$$b_{YX} = b_{VU} = 0.664$$

$$\hat{Y} - 72.6 = 0.664\ (x - 69.6)$$

$$\hat{Y} - 72.6 = 0.64x - 46.214$$

$$\hat{Y} = 0.664x + 26.386$$

Regression equation of X on Y,

$$\hat{X} - \bar{x} = b_{XY}\left(y - \bar{y}\right)$$

Calculation of b_{XY}

$$b_{XY} = b_{UV}\frac{n\sum_{i=1}^{n}u_i v_i - \sum_{i=1}^{n}u_i \sum_{i=1}^{n}v_i}{n\sum_{i=1}^{n}v_i^2 - \left(\sum_{i=1}^{n}v_i\right)^2}$$

$$= \frac{5\times 970 - 48\times(-12)}{5\times 1074 - (-12)^2}$$

$$= \frac{4850+576}{5370-144} = \frac{5426}{5226}$$

$$b_{UV} = 1.038$$

$$\hat{X} - 69.6 = 1.038\ (y - 72.6)$$

$$\hat{X} - 69.6 = 1.038y - 75.359$$

$$\hat{X} = 1.038y - 5.759$$

Example 5.6

Is there any mistake in the data provided about the two regression lines $Y = -1.5\,X + 7$, and $X = 0.6\,Y + 9$? Give reasons.

Solution:

The regression coefficient of Y on X is $b_{YX} = -1.5$

The regression coefficient of X on Y is $b_{XY} = 0.6$

Both the regression coefficients are of different sign, which is a contrary. So the given equations cannot be regression lines.

Example: 5.7

	mean	S.D
Yield of wheat (kg. unit area)	10	8
Annual Rainfall (inches)	8	2

Correlation coefficient: 0.5

Estimate the yield when rainfall is 9 inches

Solution:

Let us denote the dependent variable yield by Y and the independent variable rainfall by X.

Regression equation of Y on X is given by

$$Y - \bar{y} = r_{XY} \frac{SD(Y)}{SD(X)} (x - \bar{x})$$

$\bar{x} = 8,\ SD(X) = 2, \bar{y} = 10,\ SD(Y) = 8,\quad r_{XY} = 0.5$

$$Y - 10 = 0.5 \times \frac{8}{2} (x - 8)$$

$$= 2\,(x - 8)$$

When $x = 9$,

$$Y - 10 = 2\,(9 - 8)$$

$$Y = 2 + 10$$

$$= 12 \text{ kg (per unit area)}$$

Corresponding to the annual rain fall 9 inches the expected yield is 12 kg (per unit area).

Example 5.8

For 50 students of a class the regression equation of marks in Statistics (X) on marks in Accountancy (Y) is $3Y - 5X + 180 = 0$. The mean marks in of Accountancy is 50 and variance of marks in statistics is $\frac{16}{25}$ of the variance of marks in Accountancy.

Find the mean marks in statistics and the coefficient of correlation between marks in the two subjects when the variance of Y is 25.

Solution:

We are given that:

$n = 50$, Regression equation of X on Y as $3Y - 5X + 180 = 0$

$\bar{y} = 50$, $V(X) = \frac{16}{25} V(Y)$, and $V(Y) = 25$.

We have to find (i) $\bar{x}$ and (ii) r_{XY}

(i) Calculation for $\bar{x}$

Since $(\bar{x}, \bar{y})$ is the point of intersection of the two regression lines, they lie on the regression line $3Y - 5X + 180 = 0$

Hence, $3\bar{y} - 5\bar{x} + 180 = 0$

$3(50) - 5\bar{x} + 180 = 0$

$-5\bar{x} = -180 - 150$

$= -330$

$\bar{x} = \frac{-330}{-5} = 66$

$\bar{x} = 66$

(ii) Calculation for coefficient of correlation.

$3Y - 5X + 180 = 0$

$-5X = -180 - 3Y$

$X = 36 + 0.6\ Y$

$b_{XY} = 0.6$

Also $b_{XY} = r_{XY} \frac{SD(X)}{SD(Y)}$

$0.6 = r_{XY} \frac{SD(X)}{SD(Y)}$

$r_{XY} = \frac{0.6 \times SD(Y)}{SD(X)}$

$r_{XY}^2 = 0.36 \times \frac{V(Y)}{V(X)}$ (1)

Given that:

$V(Y) = 25$

$V(X) = \frac{16}{25} V(Y)$

$= \frac{16}{25} \times 25$

$V(X) = 16$

Substituting in (1) we have

$$r_{XY}^2 = \frac{0.36 \times 25}{16}$$

$$r_{XY} = \sqrt{\frac{0.36 \times 25}{16}} = 0.75$$

Example 5.9

If two regression coefficients are $b_{YX} = \frac{5}{6}$ and $b_{XY} = \frac{9}{20}$, what would be the value of r_{XY}?

Solution:

The correlation coefficient $r_{XY} = \pm\sqrt{(b_{YX})(b_{XY})}$

$$= \pm\sqrt{\frac{5}{6} \times \frac{9}{20}} = \sqrt{0.375} = 0.6124$$

Since both the signs in b_{YX} and b_{XY} are positive, correlation coefficient between X and Y is positive.

Example 5.10

Given that $b_{YX} = -\frac{8}{7}$ and $b_{XY} = -\frac{5}{6}$. Find r?

NOTE

The sign of the corelation coefficient will be the signs of the regression coefficients.

Solution:

$$r_{XY} = \pm\sqrt{(b_{YX})(b_{XY})}$$

$$= \sqrt{-\frac{8}{7} \times -\frac{5}{6}} = \sqrt{\frac{20}{21}} = -0.9759.$$

Since both the signs in b_{YX} and b_{XY} are negative, correlation coefficient between X and Y is negative.

5.6 DIFFERENCE BETWEEN CORRELATION AND REGRESSION

Correlation	Regression
1. It indicates only the nature and extent of linear relationship	It is the study about the impact of the independent variable on the dependent variable. It is used for predictions.
2. If the linear correlation is coefficient is positive / negative , then the two variables are positively / or negatively correlated	The regression coefficient is positive, then for every unit increase in x, the corresponding average increase in y is b_{YX}. Similarly, if the regression coefficient is negative , then for every unit increase in x, the corresponding average decrease in y is b_{YX}.
3. One of the variables can be taken as x and the other one can be taken as the variable y.	Care must be taken for the choice of independent variable and dependent variable. We can not assign arbitrarily x as independent variable and y as dependent variable.
4. It is symmetric in x and y, *ie.*, $r_{XY}=r_{YX}$	It is not symmetric in x and y, that is, b_{XY} and b_{YX} have different meaning and interpretations.

POINTS TO REMEMBER

- There are several types of regression - Simple linear correlation , multiple linear correlation and non-linear correlation.
- In simple linear regression there are two linear regression lines Y on X and X on Y.
- In the linear regression line $Y = a + bX + e$, where 'X' is independent variable, 'Y' is dependent variable, 'a' is intercept, 'b' is slope of the line and 'e' is error term.
- The point $(\overline{X},\overline{Y})$ passes through the regression lines.
- The " Method of least squares" gives the line of best fit.
- Both the regression lines have the same sign either positive of negative.
- The sign of the regression coefficient and the sign of the correlation coefficient is same.

EXERCISE 5

I. Choose the best answer.

1. _____ is widely used for prediction

   a) regression analysis b) correlation analysis

   c) analysis of variance d) analysis of covariance

2. The linear regression analysis can be classified in to

   a) 4 types b) 3 types c) 2 types d) none of the above

3. The linear equation $Y = a + bx$ is called as regression equation of

   a) X on Y b) Y on X c) between X and Y d) 'a' on 'b'

4. In regression equation $X = a + by + e$ is

   a) correlation coefficient of Y on X b) correlation coefficient of X on Y

   c) regression coefficient of Y on X d) regression coefficient of X on Y

5. $b_{YX} =$

   a) $r_{XY}\frac{SD(X)}{SD(Y)}$ b) $r_{XY}\frac{SD(Y)}{SD(X)}$ c) $\frac{SD(X)}{SD(Y)}$ d) $\frac{SD(Y)}{SD(X)}$

6. If $b_{XY} > 1$ then b_{YX} is

   a) 1 b) 0 c) > 1 d) < 1

7. In the Regression equation $\hat{Y} - \bar{y} = r_{XY}\frac{SD(Y)}{SD(X)}(x - \bar{x})$, $r_{XY}\frac{SD(Y)}{SD(X)}$ is

   a) b_{YX} b) b_{XY} c) r_{XY} d) $\text{cov}(X, Y)$

8. Using the regression coefficients we can calculate

   a) cov(X, Y) b) $SD(X)$

   c) correlation coefficient d) coefficient of variance

9. Arithmetic mean of the regression coefficients b_{XY} and b_{YX} is

   a) $> r_{XY}$ b) $\geq r_{XY}$ c) $\leq r_{XY}$ d) $< r_{XY}$

10. Regression analysis helps in establishing a functional relationship between _____ variables.

    a) 2 or more variables b) 2 variables

    c) 3 variables d) none of these

11. _____ is the Father of mental tests

    a) R.A. Fisher b) Croxton and Cowden

    c) Francis Galton d) A.L. Bowley

12. Correlation coefficient is the _____ between the regression coefficients

a) arithmetic mean b) geometric mean

c) harmonic mean d) none of the above

13. If the two lines of regression are perpendicular to each other then r_{XY} =

a) 0 b) 1 c) –1 d) 0.5

14. If the two regression lines are parallel then

a) $r_{XY} = 0$ b) $r_{XY} = +1$ c) $r_{XY} = -1$ d) $r_{XY} = \pm 1$

15. Angle between the two regression lines is

a) $\tan^{-1}\left(\frac{m_1 + m_2}{1 - m_1 m_2}\right)$ b) $\tan^{-1}\left(\frac{m_1 m_2}{1 + m_1 m_2}\right)$

c) $\tan^{-1}\left(\frac{m_1 - m_2}{1 + m_1\ m_2}\right)$ d) none of the above

16. b_{XY} =

a) $r_{XY}\frac{SD(Y)}{SD(X)}$ b) $r_{XY}\frac{SD(X)}{SD(Y)}$

c) $r_{XY}\,SD(X)\,SD(Y)$ d) $\frac{1}{b_{YX}}$

17. Regression equation of X on Y is

a) $Y = a + b_{YX}x + e$ b) $Y = b_{XY}x + a + e$

c) $X = a + b_{XY}y + e$ d) $X = b_{YX}y + a + e$

18. For the regression equation $2\hat{Y} = 0.605x + 351.58$. The regression coefficient of Y on X is

a) $b_{XY} = 0.3025$ *b)* $b_{XY} = 0.605$

c) $b_{YX} = 175.79$ d) $b_{YX} = 351.58$

19. If $b_{XY} = 0.7$ and 'a' = 8 then the regression equation of X on Y is

a) $Y = 8 + 0.7\ X$ b) $X = 8 + 0.7\ Y$

c) $Y = 0.7 + 8\ X$ d) $X = 0.7 + 8\ Y$

20. The regression lines intersect at

a) $(\overline{X}, \overline{Y})$ b) (X, Y) c) $(0, 0)$ d) $(1, 1)$

II. Give very short answer to the following questions.

21. Define regression.
22. What are the types of regression?
23. Write the two simple linear regression equations.
24. Write the two simple linear regression coefficients.
25. If the regression coefficient of X on Y is 16 and the regression coefficient of Y on X is 4, then find the correlation coefficient.
26. Find the standard deviation of Y given that $V(X)$ is 36, $b_{XY} = 0.8$, $r_{XY} = 0.5$.

III. Give short answer to the following questions.

27. Define simple linear and multiple linear regressions
28. Distinguish between linear and non-linear regression.
29. Write the regression equation of X on Y and its normal equations.
30. Write the regression equation of Y on X and its normal equations.
31. Write any three properties of regression.
32. Write any three uses of regression.
33. Write any three differences between correlation and regression.
34. If the regression equations are $\hat{X} = 64 - 0.95y$, $\hat{Y} = 7.25 - 0.95x$ then find the correlation coefficient.
35. Given the following lines of regression.

    $8X - 10Y + 66 = 0$ and $40X - 18Y = 214$. Find the mean values of X and Y.

36. Given $\overline{x} = 90$, $\overline{y} = 70$, $b_{XY} = 1.36$, $b_{YX} = 0.61$ when $y = 50$, Find the most probable value of X.
37. Compute the two regression equations from the following data.

x	1	2	3	4	5
y	3	4	5	6	7

    If $x = 3.5$ what will be the value of $\hat{Y}$?

IV Give detailed answer to the following questions.

38. Write in detail the properties of regression.
39. Explain in detail the uses of regression analysis.
40. Distinguish between correlation and regression.
41. Interpret the result for the given information. A simple regression line is fitted for a data set and its intercept and slope respectively are 2 and 3. Construct the linear regression of the form $Y = a + bx$ and offer your interpretation for 'a' and 'b'. If X is increased from 1 to 2, what is the increase in Y value. Further if X is increased from 2 to 5 what would be the increase in Y. Demonstrate your answer mathematically.

42. Using the method of least square, calculate the regression equation of X on Y and Y on X from the following data and estimate X where Y is 16.

x	10	12	13	17	18
y	5	6	7	9	13

Also determine the value of correlation coefficient.

43. The following table shows the age (X) and systolic blood pressure (Y) of 8 persons.

Age (X)	56	42	60	50	54	49	39	45
Blood pressure (Y)	160	130	125	135	145	115	140	120

Fit a simple linear regression model, Y on X and estimate the blood pressure of a person of 60 years.

44. Find the regression equation of X on Y given that $n = 5$, $\Sigma x = 30$, $\Sigma y = 40$, $\Sigma xy = 214$, $\Sigma x^2 = 220$, $\Sigma y^2 = 340$.

45. Given the following data, estimate the marks in statistics obtained by a student who has scored 60 marks in English.

Mean of marks in Statistics = 80, Mean of marks in English = 50, S.D of marks in Statistics = 15, S.D of marks in English = 10 and Coefficient of correlation = 0.4.

46. Find the linear regression equation of percentage worms (Y) on size of the crop (X) based on the following seven observations.

Size of the crop (X)	16	15	11	27	39	22	20
Percentage worms (Y)	24	25	34	40	35	20	23

47. In a correlation analysis, between production (X) and price of a commodity (Y) we get the following details.

Variance of $X = 36$.

The regression equations are:

$12X - 15Y + 99 = 0$ and $60X - 27Y = 321$

Calculate (a) The average value of X and Y.

(b) Coefficient of correlation between X and Y.

ANSWERS

I. 1. a) 2. c) 3. b) 4. d) 5. b)
6. d) 7. a) 8. c) 9. b) 10. a)
11. c) 12. b) 13. a) 14. d) 15. c)
16. b) 17. c) 18. a) 19. b) 20. a)

II. 25) $r_{XY} = 8$

26) $SD(Y) = 3.75$

III. 34) $r_{XY} = -0.95$

35) $\overline{X} = 13$, $\overline{Y} = 17$

36) when $Y = 50$, $\hat{X} = 62.8$

37) Regression equation X on Y: $\hat{X} = Y - 2$
Regression equation Y on X: $\hat{Y} = X + 2$
when $X = 3.5$, $\hat{Y} = 5.5$

IV. 41) (1) If X increases by 1 unit then Y increases by 3 units
(2) If X increases by 3 units then Y increases by 9 units

42) (1) Regression equation of X on Y is $X = Y + 6$; when $Y = 16$, $X = 22$
(2) Regression equation of Y on X is $Y = 0.89\,X - 2.59$
(3) $b_{xy} = 1$, $b_{yx} = 0.87$, $r = 0.93$

43) $Y = 0.45\,X + 111.53$, $Y = 138.53$ when age is 60 years.

44) $a = 16.4$, $b = -1.3$
Regression equation of X on Y is : $X = 16.4 - 1.3Y$

45) $X = 86$ when $Y = 60$

46) $Y = 0.32\,X + 21.84$

47) (a) Mean of $X = 13$ and mean of $Y = 17$. (b) $r = 0.6$