UNIT 9 SIMPLE LINEAR REGRESSION

Structure

9.1 INTRODUCTION

In Blocks 1 and 2, you have learnt some basic methods of optimisation of various problems such as LPP, transportation problem, assignment problem, queueing problem, scheduling and sequencing problem and inventory problem. In this unit, we discuss the concepts of regression modelling. Regression analysis is a statistical tool for investigating and analysing the average relationship between two or more variables. In regression analysis, one variable is referred to as the dependent variable or response variable, whereas the other variable is referred to as independent variable, predictor variable or regressor variable. The regression analysis with only one predictor variable is known as **simple regression analysis** and with two or more predictor variables is known as **multiple regression analysis**.

The term simple linear regression refers to a regression equation with only one predictor variable and the equation is linear. This means that if Y is the dependent variable and X, the independent variable, the regression equation is of the form $Y = a + b X$. Such an analysis is useful in many situations, e.g., in analysing the relationship between the number of customers and monthly sales of a product. In regression analysis, we use the method of least squares to fit the regression equation to the data. We discuss the concept of simple linear regression, the method of least squares and fitting of regression line in Secs. 9.2 to 9.4. In Sec. 9.5, we present residual analysis.

In the next unit, we shall discuss the test of significance of regression coefficients, the coefficient of determination and confidence intervals of regression coefficients.

Objectives

After studying this unit, you should be able to:

- describe the simple linear regression model;
- estimate the model parameters using the method of least squares;
- describe the properties of the estimates;
- fit the simple linear regression line; and
- check the model assumptions with the help of residual analysis.

9.2 SIMPLE LINEAR REGRESSION

Regression analysis may broadly be defined as the analysis of relationships among variables. It is popular because it explores the power of statistics as a tool for establishing average relationships among variables. This relationship is given as an equation that helps to predict the dependent variable Y through one or more independent variables. In regression analysis, the variable whose values vary with the variations in the values of the other variable(s) is called the **dependent variable** or **response variable.** The other variables which are independent in nature and influence the response variable are called **independent variables**, **predictor variables** or **regressor variables**. A regression equation with a single independent variable is called a **simple regression equation** and it is linear when the model parameters have a linear relation.

The equation $Y = a + b X^2$ may also be classified as simple linear regression because it has only one independent variable X and by change of variable, e.g., $X^2=T$, we can get a straight line of the form $Y = a + b T$ even though Y bears a quadratic relationship with X. Its conceptual usefulness exists in analysing the relationship between a variable of interest and a set of related predictor variables using an equation.

Let us take an example where regression technique may be useful. Suppose a statistician employed by a cold drink bottler is analysing the product delivery and service operation for vending machines. He would like to find how the delivery time taken by the delivery man to load and service a machine is related to the volume of delivery cases. The statistician visits 50 randomly chosen retailer shops having vending machines and observes the delivery time (in minutes) and the volume of delivery cases for each shop. He plots those 50 observations on a graph, which shows that an approximate linear relationship exists between the delivery time and delivery volume. If Y represents the delivery time and X, the delivery volume, the equation of a straight line relating these two variables may be given as

$$Y = a + bX \quad \ldots (1)$$

where a is the intercept and b, the slope.

In such cases, we draw a straight line in the form of equation (1) so that the data points generally fall near the straight line. Now, suppose the points do not fall exactly on the straight line. Then we should modify equation (1) to minimise the difference between the observed value of Y and that given by the straight line (a + b X). This is known as error.

The error e, which is the difference between the observed value and the predicted value of the variable of interest Y, may be conveniently assumed as a statistical error. This error term accounts for the variability in Y that cannot be explained by the linear relationship between X and Y. It may arise due to the effects of other factors. Thus, a more plausible model for the variable of interest (Y) may be given as

$$Y = a + b X + e \quad \ldots (2)$$

where the intercept a and the slope b are unknown constants and e is a random error component.

Equation (2) is called a **linear regression model.** Since this model involves only one regressor or independent variable, it is called a **simple linear regression model**. In case of two or more than two independent variables, it is known as a **multiple linear regression model**.

The following assumptions are made about the error term e in the regression model given in equation (2):

i) the error term e is a random variable with mean zero; i.e., E (e) = 0;

ii) the variance of the error term e, denoted by σ^2, is the same for all values of X;

iii) all the error terms, $e_1, e_2, \ldots, e_n$, are independent. This means that the value of e for a particular value of X is not related to the value of e for any other value of X. Thus, the value of Y for a particular X is independent of any other value of X; and

iv) the error term e is a normally distributed random variable.

These assumptions provide the theoretical basis for the t-test and F-test used to determine whether the relationship between X and Y is significant and for the estimation of confidence and prediction interval.

You may now like to answer the following questions, which will help you assess your understanding of simple linear regression.

E1) What is simple linear regression ? Write the equation for simple linear regression model.

E2) Answer the following questions for the error term e in the regression model $Y = a + b X + e$:

i) What are the mean and variance of the random error e?

ii) Which distribution does it follow?

iii) Are the errors $e_1, e_2, \ldots$ independent?

9.3 LEAST SQUARES ESTIMATION OF PARAMETERS

In the previous section, we have discussed the simple regression equation with only one regressor variable X and the variable of interest Y. We have also discussed the simple linear regression model with a single regressor variable X. The simple linear regression model has two unknown parameters a and b, which are known as **intercept** and **regression coefficient**, respectively. Their values are unknown. Therefore, they must be estimated using sample data. The estimation of the parameters a and b is done by minimising the error term e.

Let $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$ be n pairs of values in the data. The equation of the simple linear regression model may be written as

$$Y = a + bX + e \qquad \ldots (3)$$

where e represents the error term which arises due to the difference of the observed Y and the fitted line $\hat{Y} = \hat{a} + \hat{b} X$. We use the method of least squares to minimise the error term e. From equation (3), we may write a simple regression model as

$$Y_i = a + b X_i + e_i \qquad i = 1, 2, \ldots, n \qquad \ldots (4)$$

for a sample data of n pairs of values given in terms of (Y_i, X_i), $(i = 1, 2, \ldots n)$.

We estimate a and b so that the sum of the squares of the differences between the observed values (Y_i) and the points lying on the straight line is minimum, i.e., the sum of squares of the error terms given by

$$E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - a - bX_i)^2 \quad \ldots (5)$$

is minimum. To find the values of a and b for which the sum of squares of the error terms, i.e., E is minimum, we differentiate it with respect to the parameters a and b and equate the results to zero:

$$\frac{\partial E}{\partial a} = -2\sum_{i=1}^{n} (Y_i - a - bX_i) = 0 \quad \ldots (6)$$

and

$$\frac{\partial E}{\partial b} = -2\sum_{i=1}^{n} (Y_i - a - bX_i) X_i = 0 \quad \ldots (7)$$

Simplifying equations (6) and (7), we get

$$n\,a + b\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \quad \ldots (8)$$

$$a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i X_i \quad \ldots (9)$$

Equations (8) and (9) are called the **least-squares normal equations**. The solution to these normal equations is

$$\hat{a} = \overline{Y} - \hat{b}\overline{X} \quad \ldots (10)$$

where $\overline{Y}$ and $\overline{X}$ are the averages of Y_i and X_i, respectively. On putting the value of $\hat{a}$ from equation (10) in equation (9), we get

$$\hat{b} = \frac{\sum_{i=1}^{n} Y_i X_i - \left(\sum Y_i\right)\left(\sum X_i\right)/n}{\sum_{i=1}^{n} X_i^2 - \left(\sum X_i\right)^2/n} \quad \ldots (11)$$

Since the denominator of equation (11) is the corrected sum of squares of X_i, we may rewrite it as

$$SS_X = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum X_i\right)^2}{n} = \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Similarly, the numerator is the corrected sum of the cross product of X_i and Y_i and may be rewritten as:

$$SS_{XY} = \sum Y_i X_i - \frac{\left(\sum Y_i\right)\left(\sum X_i\right)}{n} = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

Therefore, the expression for $\hat{b}$ may be rewritten as

$$\hat{b} = \frac{SS_{XY}}{SS_X} \quad \ldots (12)$$

Thus, $\hat{a}$ and $\hat{b}$ are the least squares estimates of the intercept a and slope b, respectively. Therefore, the fitted simple linear regression model is given by

$$\hat{Y} = \hat{a} + \hat{b}\,X \qquad \ldots (13)$$

Equation (13) gives a point estimate of the mean of Y for a particular X. The difference between the fitted value $\hat{Y}_i$ and Y_i is known as the **residual** and is denoted by r_i:

$$r_i = Y_i - \hat{Y}_i, \qquad i = 1, 2, \ldots, n \qquad \ldots(14)$$

The role of the residuals and its analysis is very important in regression modelling. We shall discuss it in detail in Sec. 9.5.

Properties of Estimates of Parameters

So far you have learnt that the estimates of parameters of a simple linear regression model are obtained with the help of the method of least squares. On putting the values of the estimates of the parameters in the equation, we get a fitted simple linear regression model. The least squares estimates $\hat{a}$ and $\hat{b}$ have several important properties, which are given below:

a) The means of $\hat{a}$ and $\hat{b}$ (the estimates of the parameters a and b) of a simple linear regression model

$$E(Y_i) = a + b\,X_i$$

are given by

$$E(\hat{a}) = a \text{ and } E(\hat{b}) = b,$$

respectively. Thus, if we assume that the model is correct, then $\hat{a}$ and $\hat{b}$ are unbiased estimators of a and b.

The variance of $\hat{b}$ is given as

$$\text{Var}(\hat{b}) = \frac{\sigma^2}{SS_x} \qquad \ldots(15)$$

The variance of $\hat{a}$ is given as

$$\text{Var}(\hat{a}) = \text{Var}(\bar{Y} - \hat{b}\,\bar{X})$$
$$= \text{Var}(\hat{\bar{Y}}) + \bar{X}^2\,\text{Var}(\hat{b}) - 2\bar{X}\,\text{Cov}(\bar{Y}, \hat{b})$$

Now the variance of $\hat{\bar{Y}}$ is just $\text{Var}(\bar{Y}) = \frac{\sigma^2}{n}$ and the covariance between $\hat{\bar{Y}}$ and $\hat{b}$ would be zero. Thus, we get

$$\text{Var}(\hat{a}) = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{SS_X}\right) \qquad \ldots(16)$$

b) The sum of the residuals in any regression model that contains an intercept a is always zero, i.e.,

$$\sum_{i=1}^{n} r_i = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right) = \sum (Y_i - \hat{a} - \hat{b}X_i)$$
$$= \sum [Y_i - \bar{Y} - \hat{b}\left(X_i - \bar{X}\right)] \qquad \because \hat{a} = \bar{Y} - \hat{b}\,\bar{X}$$
$$= \sum (Y_i - \bar{Y}) - \hat{b}\sum \left(X_i - \bar{X}\right) = 0 \qquad \ldots(17)$$

c) The sum of the observed values (Y_i) is equal to the sum of the fitted values $(\hat{Y}_i)$:

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$$

d) The least square regression lines always pass through the centroid $(\bar{X}, \bar{Y})$ of the data since we know that the two lines of regression intersect at this point.

e) The sum of the residuals weighted by the corresponding value of the regressor variables is always zero, i.e., $\sum_{i=1}^{n} X_i r_i = 0$

f) The sum of the residuals weighted by the corresponding fitted values is always equal to zero, i.e., $\sum_{i=1}^{n} \hat{Y}_i r_i = 0$

Now, you should answer the following questions to assess your understanding.

E3) Describe the method of obtaining the least squares estimates of the parameters a and b for the regression model Y = a + bX + e.

E4) Describe the properties (in brief) of the estimates of the parameters a and b for the regression model Y = a + bX + e.

9.4 FITTING OF REGRESSION LINE

In Sec. 9.3, you have learnt the method of least squares for estimating the parameters a and b in the simple linear regression model Y= a + b X +e. You have also learnt the properties of the least squares estimates. In this section, we discuss the method of fitting the simple linear regression equation for the given data of n pairs of observations on (X, Y).

Let the given data of n pairs of observations on X and Y be as follows:

X:	X_1	X_2	X_3	. …	…	X_i	….. …	X_n
Y:	Y_1	Y_2	Y_3	. …	…	Y_i	….. …	Y_n

where Y is the dependent variable and X, the independent variable. Suppose, we wish to fit the following simple regression equation to the data:

$$Y = a + bX \qquad \ldots(18)$$

where a is the intercept and b is the slope of the equation. For fitting equation (18) to the data on (X, Y), we follow the steps given below:

Step 1: We draw a scatter diagram by plotting the (X, Y) points given in data.

Step 2: We construct a table as given below and take the sum of the values of X_i, X_iY_i, and X_i^2. We write the values of $\sum X$, $\sum Y$, $\sum XY$ and $\sum X^2$ in the last row:

X	Y	XY	X^2
X_1	Y_1	$X_1 Y_1$	X_1^2
X_2	Y_2	$X_2 Y_2$	X_2^2
X_3	Y_3	$X_3 Y_3$	X_3^2
.	.	.	.
.	.	.	.
X_i	Y_i	$X_i Y_i$	X_i^2
.	.	.	.
.	.	.	.
X_n	Y_n	$X_n Y_n$	X_n^2
$\sum X$	$\sum Y$	$\sum XY$	$\sum X^2$

Step 3: We express of $\hat{a}$ given in equation (10) as follows:

$$\hat{a} = \bar{Y} - b\bar{X} = \frac{1}{n}\left[\sum Y - b\sum X\right] \quad \ldots (19)$$

Step 4: We substitute the values of $\sum X$, $\sum Y$, $\sum XY$ and $\sum X^2$ in equation (11) and calculate the optimum value of $\hat{b}$ as

$$\hat{b} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2} \quad \ldots (20)$$

Step 5: Now for obtaining the value of $\hat{a}$, we substitute the value of $\sum X$, $\sum Y$ and slope $\hat{b}$ in equation (19) and obtain the optimum value of intercept $\hat{a}$ as

$$\hat{a} = \frac{1}{n}\left[\sum Y - \hat{b}\sum X\right]$$

Step 6: We put these values of $\hat{a}$ and $\hat{b}$ in the regression equation and get

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

Let us obtain the regression line for a given set of data.

Example 1: Sales data of 10 months for a coffee house situated near a prime location of a city comprising the number of customers (in hundreds) and monthly sales (in Thousand Rupees) are given below:

S. No.	No. of Customers (in hundreds)	Monthly Sales (in thousand Rs.)
1	6.0	01
2	6.1	06
3	6.2	08
4	6.3	10
5	6.5	11
6	7.1	20
7	7.6	21
8	7.8	22
9	8.0	23
10	8.1	25

Find the simple linear regression equation that fits the given data.

Solution: It is given that n = 10. Let us assume an equation of simple linear regression as follows:

$$Y = a + bX$$

where X is the independent variable and Y, the dependent variable. Now for finding the fitted regression equation, we follow the procedure described in Steps 1 to 6.

We first draw a scatter diagram by plotting the (X, Y) points as shown in Fig. 9.1.

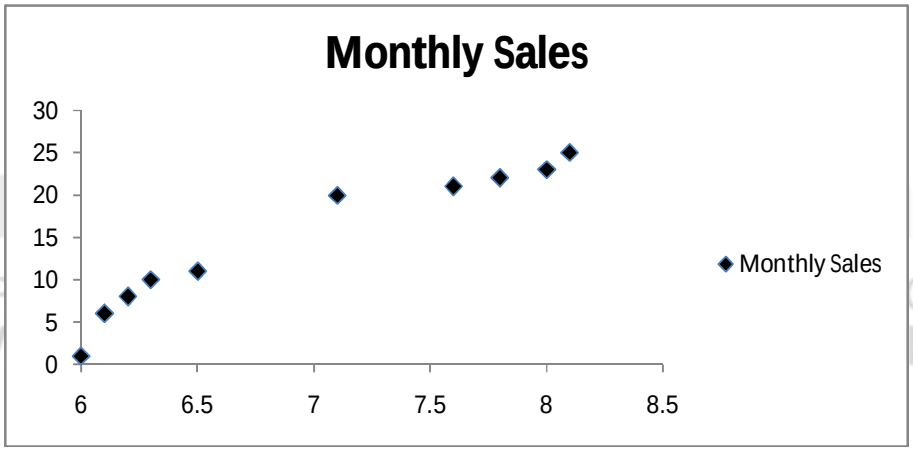


Fig. 9.1: Scatter diagram for the (X, Y) points given in the data of monthly sales.

We construct a table for finding the values of ΣX, ΣY, ΣXY and ΣX^2 as follows:

X	Y	XY	X^2
6.0	01	6.0	36.00
6.1	06	36.6	37.21
6.2	08	49.6	38.44
6.3	10	63.0	39.69
6.5	11	71.5	42.25
7.1	20	142.0	50.41
7.6	21	159.6	57.76
7.8	22	171.6	60.84
8.0	23	184.0	64.00
8.1	25	202.5	65.61
$\sum X = 69.7$	$\sum Y = 147$	$\sum XY = 1086.4$	$\sum X^2 = 492.21$

From the table,

$$\sum X = 69.7 \quad \sum Y = 147 \quad \sum XY = 1086.4 \quad \text{and} \quad \sum X^2 = 492.21$$

We obtain the intercept a and slope b from equations (19) and (20) by substituting the values of $\sum X$, $\sum Y$, $\sum XY$ and $\sum X^2$ in them. Thus,

$$\hat{b} = \frac{10 \times 1086.4 - (147)(69.7)}{10 \times 492.21 - (69.7)(69.7)}$$

$$= \frac{1086.40 - 10245.9}{4922.10 - 4858.09} = \frac{618.10}{64.01} = 9.656$$

and

sg

$$= \frac{1}{10}[147 - 9.656 \times 69.7] = \frac{1}{10}[147 - 673.02] = -52.6$$

The best fitted regression equation for the given data using the method of least squares is given as

$$\hat{Y} = -52.6 + 9.656\ X$$

We may also show the fitted regression line for the given data as in Fig. 9.2.

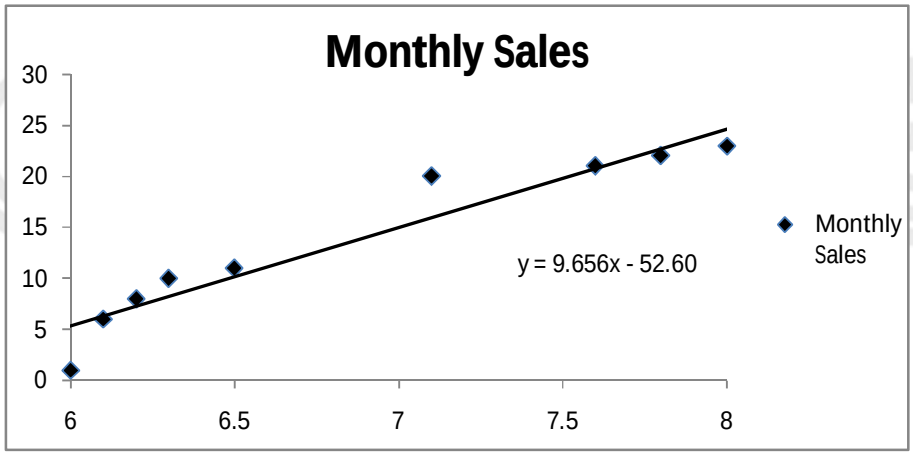


Fig. 9.2: Fitted regression equation for Example 1.

You may now like to solve the following problem to check your understanding:

E5) A statistician collected the sales data of a restaurant near a metro station for 12 months. The data comprising the number of customers (in hundreds) and monthly sales (in thousand Rupees) is given as follows:

Month	No. of Customers (in hundreds)	Monthly Sales (in thousand Rs.)
1	04	1.8
2	06	3.5
3	06	5.8
4	08	7.8
5	10	8.7
6	14	9.8
7	18	10.7
8	20	11.5
9	22	12.9
10	26	13.6
11	28	14.2
12	30	15.0

Find the fitted simple linear regression equation for the given data.

9.5 RESIDUAL ANALYSIS

In the previous section, we have discussed how to obtain the estimated values of the parameters a and b for a simple linear regression model given by equation (1):

$$Y = a + bX$$

We have also explained how to get the following fitted model after minimising the error term e:

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

We are now interested in knowing: How well does this equation fit the data? Is the model likely to be useful as a predictor? Can the fitted model be used to predict the value of the dependent variable?

All these issues must be investigated before adopting the model for use. A simple but efficient method of detecting model deficiencies is to examine the residuals. Residuals are the differences between the observed values and the corresponding fitted values. Mathematically, if Y_i is the i^{th} observation of the dependent variable Y and $\hat{Y}_i$, the corresponding fitted value, then

$$r_i = Y_i - \hat{Y}_i, \quad i = 1, 2, \ldots, n \qquad \ldots (21)$$

From equation (21), note that the residual r_i is the deviation between the value given in the data and the corresponding fitted value. Upon substituting $\hat{Y}_i = \hat{a} + \hat{b}X_i$ in equation (21), we get

$$r_i = Y_i - (\hat{a} - \hat{b}X_i), \quad i = 1, 2, \ldots, n \qquad \ldots (22)$$

where $\hat{a}$ and $\hat{b}$ are the estimated values of the parameters a and b. As we have discussed earlier, residuals play a leading role in detecting the adequacy of a model. In other words, residuals may be defined as realised or observed values of the model errors.

9.5.1 Scaling of Residuals

Analysis and plotting of residuals is an effective way of discovering the model inadequacies, investigating how well the model fits the data and checking the assumptions listed below:

i) the error term e_i has zero mean;

ii) the error term e_i has constant variance σ^2;

iii) the errors are uncorrelated;

iv) the errors are normally distributed.

In this section, we discuss the standardised method for scaling residuals.

Some observations in the data are separated or very far away from the rest of the data. These are called **outliers** or **extreme values**. The scaled residuals are helpful in finding the outliers or the extreme values.

Residuals have zero mean and their approximate average variance is estimated by

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(r_i - \bar{r})^2}{n-k} = \frac{\sum_{i=1}^{n} r_i^2}{n-k} = SS_{Res} \quad (\because \bar{r} \text{ is zero}) \qquad \ldots (23)$$

where n – k is the degree of freedom. Alternatively $\sum r_i^2$ can be written as

$$\sum r_i^2 = \sum \left(Y_i - \hat{Y}_i\right)^2 = \sum \left[Y_i - \bar{Y} - \hat{b}\left(X_i - \bar{X}\right)\right]^2$$

$$= \sum \left(Y_i - \bar{Y}\right)^2 - 2\hat{b}\sum \left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right) + \hat{b}^2 \sum \left(X_i - \bar{X}\right)^2$$

Thus,

$$SS_{Res} = \left[\sum \left(Y_i - \bar{Y}\right)^2 - 2\hat{b}\sum \left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right) + \hat{b}^2 \sum \left(X_i - \bar{X}\right)^2\right] / (n - k)$$

$$= SS_Y - 2SS_{XY}^2 / SS_X + SS_{XY}^2 / SS_X$$

$$= SS_Y - SS_{XY}^2 / SS_X$$

Residuals are not independent because estimates of k parameters impose k constraints on them.

A logical method of scaling the residuals is to divide the i^{th} error term by its standard deviation that will provide the i^{th} standardised residual d_i, i.e.,

$$d_i = \frac{r_i}{\sqrt{SS_{Res}}}, \qquad \forall i = 1, 2, 3...n \qquad \ldots (24)$$

$\forall$ is the symbol used to denote the phrase "for every".

This method is known as the **standardised method of scaling the residuals** and d_i is called the i^{th} **standardised residual**. The standardised residuals have zero mean and their variance is approximately 1, i.e.,

$$E\left(d_i\right) = \frac{E(r_i)}{\sqrt{SS_{Res}}} = 0$$

and

$$Var\left(d_i\right) = \left(\frac{1}{\sqrt{SS_{Res}}}\right)^2 Var\left(r_i\right) \approx 1 \qquad \ldots (25)$$

The values of the standardised residuals indicate the positions of the observed data. A large value of a standardised residual indicates its distance from the fitted value of the model. Consequently a very large value of a standardised residual, i.e., $d_i \geq 3$ indicates an extreme observation or an outlying observation.

9.5.2 Residual Plots

A very effective way to investigate the adequacy of the fit of a regression model and to check the underlying assumptions is the graphical analysis of residuals. The basic residual plots are generated by plotting the residual values against the predicted Y values. These should be examined in all regression modelling problems.

A satisfactory residual plot should be more or less a horizontal band of points shown in Fig. 9.3.

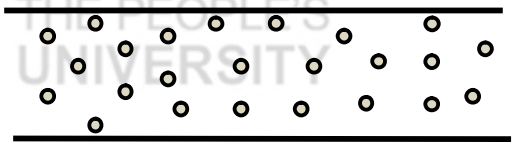


Fig. 9.3

A Heteroscadastic data (Non-constant variance) will have a residual plot as shown in Fig. 9.4.

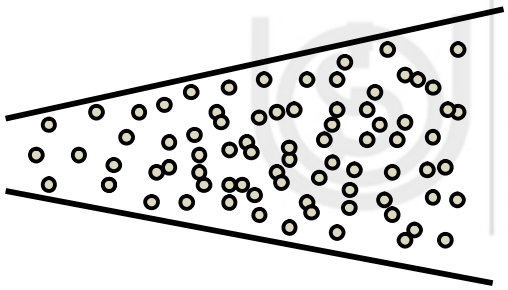


Fig. 9.4

The trend shown in Fig. 9.5 may be exhibited for data for which there exists an error in the regression calculation or some additional regression is needed in the models.

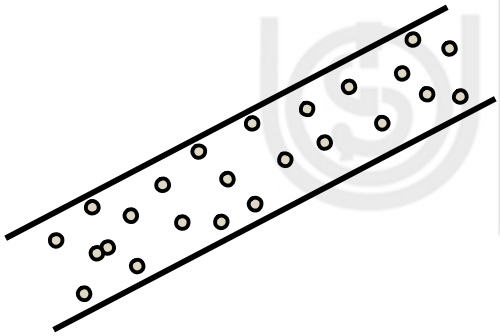


Fig. 9.5

If the relationship between X and Y is nonlinear, the pattern of residuals shown in Fig. 9.6 will be observed, i.e., a curvilinear relationship is suggested.

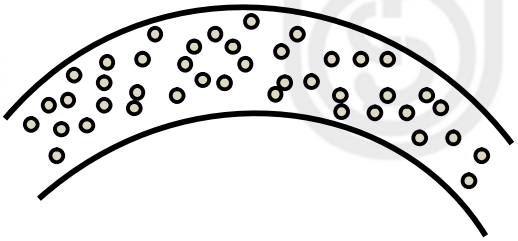


Fig. 9.6

9.5.3 Normal Probability Plot

A small departure from the normality assumptions does not affect the model extensively. Since the confidence intervals depend on the normality assumptions, the gross normality is potentially more serious. Construction of the normal probability plot of the residuals is a very simple method to check the normality assumptions. The method of constructing a normal probability plot is as follows:

1. We arrange the standardised residuals in increasing order.
2. We rank the residuals, which are arranged in increasing order. Let $r_1, r_2, r_3, \ldots, r_n$ be the calculated standardised residuals ranked in increasing order.
3. After ranking the residuals, we calculate the cumulative probability (p_i) as

$$p_i = \frac{\left(i - \frac{1}{2}\right)}{n} \qquad i = 1, 2, ..., n. \qquad \ldots (26)$$

for the corresponding ranked residual values.

4. Then we plot the cumulative probability by multiplying p_i by 100. It is known as percentile cumulative probability (P_i), i.e.,

$$P_i = \frac{\left(i - \frac{1}{2}\right)}{n} \times 100; \qquad i = 1, 2, ..., n. \qquad \ldots (27)$$

We plot P_i versus the standardised residuals to obtain the normal probability plot. The resulting points would lie approximately on a straight line.

In the normal probability plot, we usually determine the straight line visually with emphasis on the central values of all the points rather than the extreme points at both ends. You should remember the following points about the normal probability plot for the fitted model:

1. A normal probability plot is called **ideal** if all points lie approximately along a straight line as shown in Fig. 9.7.

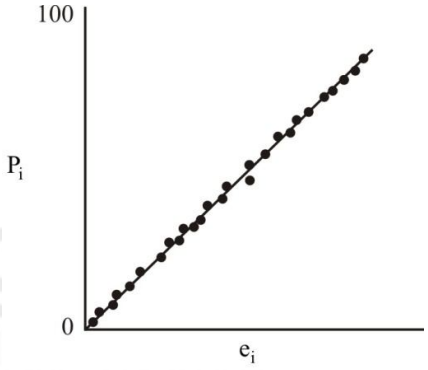
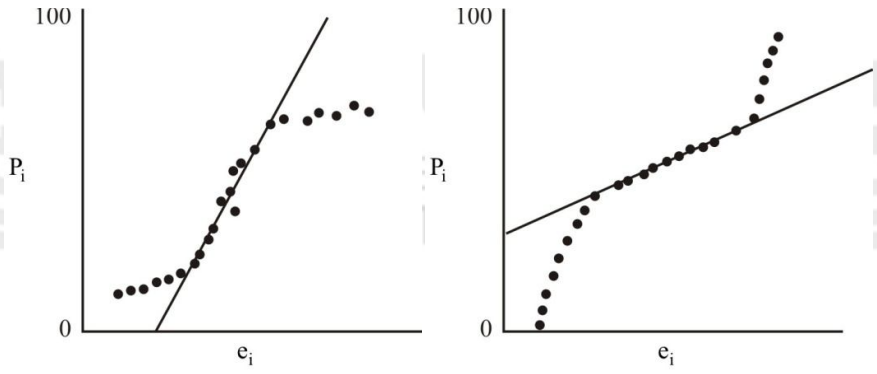


Fig. 9.7

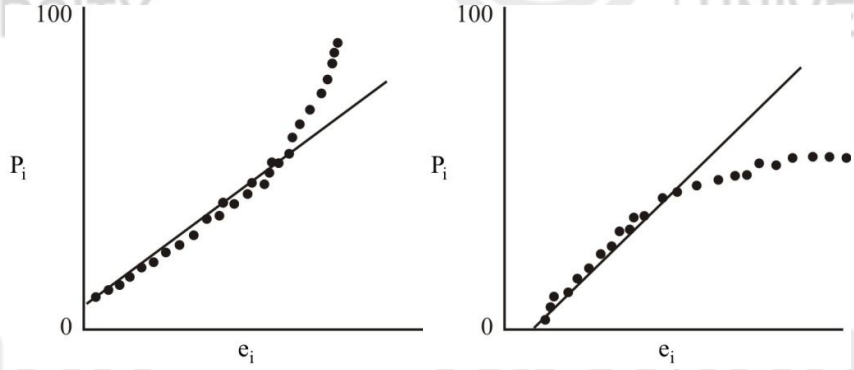
2. A sharp upward and downward curve at both ends indicates that the end points of this distribution are too light for it to be considered normal. These do not constitute the normal probability plot for light-tailed distribution as shown in Fig. 9.8a. A heavy-tailed distribution shows flattening at the extremes, which is a pattern of samples from a distribution with heavier tails than normal as shown in Fig. 9.8b.



(a) **(b)**

Fig. 9.8

3. The positive and negative skewed normal probability plot shows the pattern of an upward trend (Fig. 9.9a) and downward trend (Fig. 9.9b) of the distribution, respectively, at the ends of the distribution.



(a) **(b)**

Fig. 9.9

Let us take up an example to explain the concepts of residual analysis.

Example 2: For the data given in Example 1, calculate the residuals and determine the standardised residuals for the model. Draw the corresponding residual plot and normal probability plot.

Solution: In Example 1, we have fitted the regression equation Y = a + b X for the given set of data. We have obtained the best fitted regression equation as

$$\hat{Y} = -52.6 + 9.656X$$

We now calculate the fitted value of the dependent variable from the best fitted regression line given above. We also calculate the values of residuals and their squares for all values of Y given in the data by substituting the values of X, and arrange them in a table as follows:

Y	$\hat{\mathbf{Y}}$	$\mathbf{r_i = Y_i - \hat{Y}}$	$\mathbf{r_i^2}$
1	5.34	− 4.34	$(-4.34)^2 = 18.83$
6	6.30	− 0.30	$(-0.30)^2 = 0.09$
8	7.26	0.74	$(+0.74)^2 = 0.55$
10	8.23	1.77	$(+1.77)^2 = 3.13$
11	10.16	+0.84	$(+0.84)^2 = 0.70$
20	15.95	+4.05	$(+4.05)^2 = 16.40$
21	20.78	+0.22	$(+0.22)^2 = 0.48$
22	22.72	− 0.72	$(-0.72)^2 = 0.52$
23	24.65	−1.65	$(-1.65)^2 = 2.72$
25	25.61	− 0.61	$(-0.61)^2 = 0.37$
Total	**147.00**	**0.00**	$\sum(r_i - \bar{r})^2$ **= 43.79**

Let us check how well the regression line fits the given data.

We calculate the standard deviation of residuals from equation (23) to calculate the standardised residuals for the given data as follows:

$$SS_{Res}^{1/2} = \sqrt{\frac{\sum r_i^2}{n-k}} = \sqrt{\frac{43.79}{8}} = \sqrt{5.47} = 2.34$$

Here $n - k = 10 - 2 = 8$.

We now construct a table which contains the values of residuals and the standardised residuals as follows:

S. No.	Residuals r_i	$d_i = r_i / \sqrt{SS_{Res}}$
1	−4.34	−4.34/2.34 = − 1.855
2	−0.30	−0.30/2.34 = − 0.128
3	−0.74	+0.74/2.34 = + 0.316
4	+1.77	+1.77/2.34 = + 0.756
5	+0.84	+0.84/2.34 = + 1.036
6	+4.05	+4.05/2.34 = + 1.730
7	+0.22	+0.22/2.34 = + 0.094
8	−0.72	−0.72/2.34 = − 0.307
9	−1.65	−1.65/2.34 = − 0.705
10	−0.61	−0.61/2.34 = − 0.260
Total	$\sum r_i = 0.00$	$\sum d_i =$ **0.000**

Note that the sum of the residuals is zero. We now plot the standardised residuals against the fitted values of Y and obtain a residual plot as shown in Fig. 9.10.

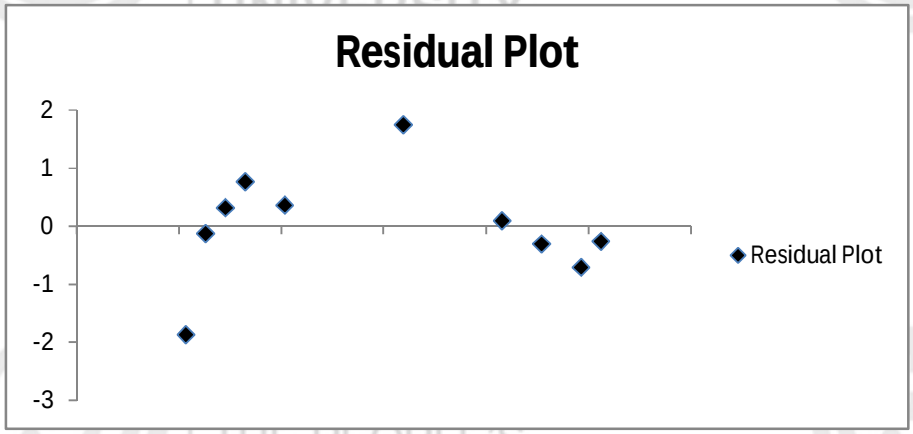


Fig. 9.10: Residual plot for the given data.

We follow the method explained in Sec.9.5.3 to construct the normal probability plot. We arrange the standardised residuals in increasing order and rank the residuals, which are also arranged in increasing order. After ranking the residuals, we calculate the percentile cumulative probability using equation (27) for the corresponding ranked residual values.

Ordered Residuals	Ranks (i)	Cumulative Probability $p_i = (i-1/2)/n$	Percentile Cumulative Probability $P_i = \{(i-1/2)/n\} \times 100$
−1.88807	1	0.041667	04.1667
−1.27982	2	0.125	12.5
−0.53461	3	0.20833	20.833
−0.48012	4	0.291667	29.1667
−0.27835	5	0.375	37.5
0.106038	6	0.45833	45.833
0.16053	7	0.541667	54.1667
0.413844	8	0.625	62.5
0.493373	9	0.70833	70.833
0.784978	10	0.791667	79.1667
1.243004	11	0.875	87.5
1.26215	12	0.95833	95.833

Then we plot the percentile cumulative probability against the standardised residuals to obtain the normal probability plot. The resulting points would lie approximately on a straight line as shown in Fig. 9.11.

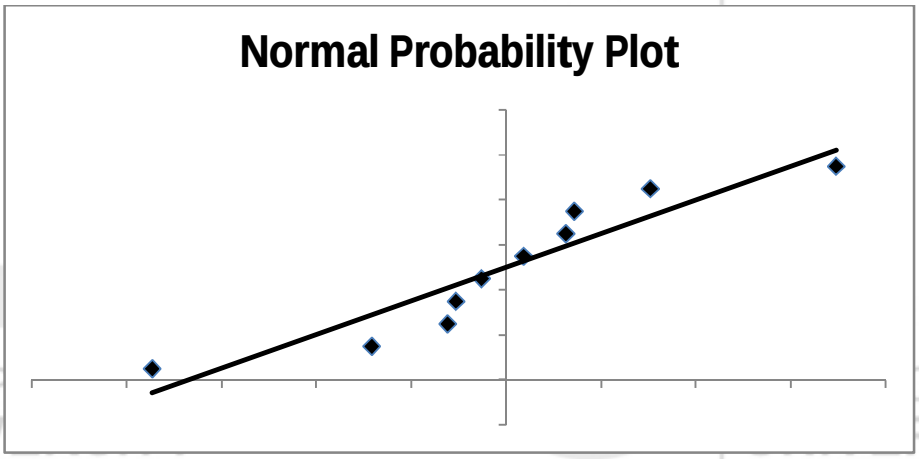


Fig. 9.11: Normal probability plot for the given data.

From Fig. 9.11, note that some points of the distribution deviate slightly from the straight line, but do not lie very far from the central points.

On the basis of the above residual analysis for the given data, the best fitted regression equation is obtained as shown in Fig. 9.12.

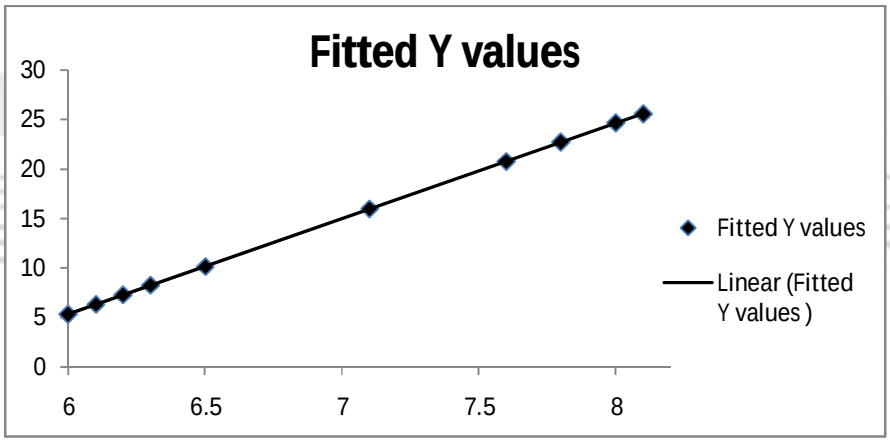


Fig. 9.12: Best fitted regression equation after residual analysis for the given data.

You may now like to solve the following questions to assess your understanding.

E6) Describe the method of scaling the residuals for the simple linear regression model $Y = a + b X + e$.

E7) Calculate the residuals and determine the standardised residuals for the simple linear regression model obtained for the data given in E2 of this unit. Generate the corresponding residual plot and the normal probability plot.

We now summarise the concepts that we have discussed in this unit.

9.6 SUMMARY

1. Simple linear regression fits a straight line through the set of n points in such a way that makes the sum of squared residuals of the model (that is, vertical distances between the points of the data set and the fitted line) as small as possible.
2. The simple linear regression model has two unknown parameters a and b, which are known as **intercept** and **regression coefficient**, respectively. Their values are unknown. Therefore, they must be estimated using the sample data. The estimation of the parameters a and b is done by minimising the error term e.
3. An essential part of regression analysis includes a careful examination of the residuals to guarantee that the assumptions in the least squares theory are met. Some observations in the data are separated or very far away from the rest of the data. These are called **outliers** or **extreme values**. The scaled residuals are helpful in finding the outliers or the extreme values.
4. A very effective way to investigate the adequacy of the fit of a regression model and to check the underlying assumptions is the graphical analysis of the residuals. The basic residual plots are generated by plotting the residual values against the predicted Y values. These should be examined in all regression modelling problems.
5. The values of the standardised residuals indicate the positions of the observed data. Large values of standardised residuals indicate their distance from the fitted values of the model. Consequently a very large value of standardised residuals, i.e., $d_i \geq 3$, indicates an extreme observation or an outlying observation.
6. A small departure from the normality assumptions does not affect the model extensively. Since the confidence intervals depend on the normality assumptions, the gross normality is potentially more serious. Construction of the normal probability plot of the residuals is a very simple method to check the normality assumptions.
7. To construct the normal probability plot, we arrange the standardised residuals in increasing order and rank the residuals, which are also arranged in increasing order. After ranking the residuals, we calculate the percentile cumulative probability for the corresponding ranked residual values.

9.7 SOLUTIONS/ANSWERS

E1) Refer to Sec. 9.2.

E2) Refer to Sec. 9.2.

E3) Refer to Sec. 9.3.

E4) Refer to Sec. 9.3.

E5) Refer to Sec. 9.4.

E6) We are given n = 12. Let us assume an equation of a simple linear regression model as

$$Y = a + b X$$

For finding the fitted regression equation, we first draw a scatter diagram by plotting the (X, Y) points as shown Fig. 9.13.

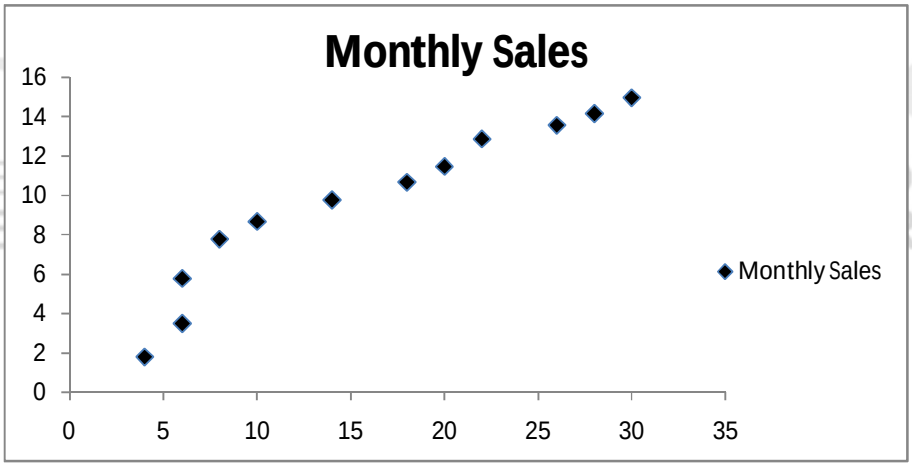


Fig. 9.13: Scatter diagram for the (X, Y) points given in the data.

We construct the following table for calculating the values of $\hat{a}$ and $\hat{b}$:

S. No.	X	Y	XY	X^2
1	4	1.8	7.2	16
2	6	3.5	21.0	36
3	6	5.8	34.8	36
4	8	7.8	62.4	64
5	10	8.7	87.0	100
6	14	9.8	137.2	196
7	18	10.7	192.6	324
8	20	11.5	230.0	400
9	22	12.9	283.8	444
10	26	13.6	353.6	676
11	28	14.2	397.6	784
12	30	15.0	450.0	900
	$\sum X = 192$	$\sum Y = 115.3$	$\sum XY = 2257.2$	$\sum X^2 = 4016$

Substituting the values of $\sum X$, $\sum Y$, $\sum XY$ and $\sum X^2$ in equations (19) and (20), we get

$$\hat{b} = \frac{12 \times 2257.2 - (192)(115.3)}{12 \times 4016 - (192)(192)}$$

$$= \frac{27086.4 - 22137.6}{48192 - 36864} = 0.4369$$

and $\hat{a} = \frac{1}{12}\left[115.3 - (0.437)(192)\right]$

$= 2.6185$

Therefore, the best fitted regression equation (Fig. 9.14) is

$\hat{Y} = 2.6185 + 0.4369X$

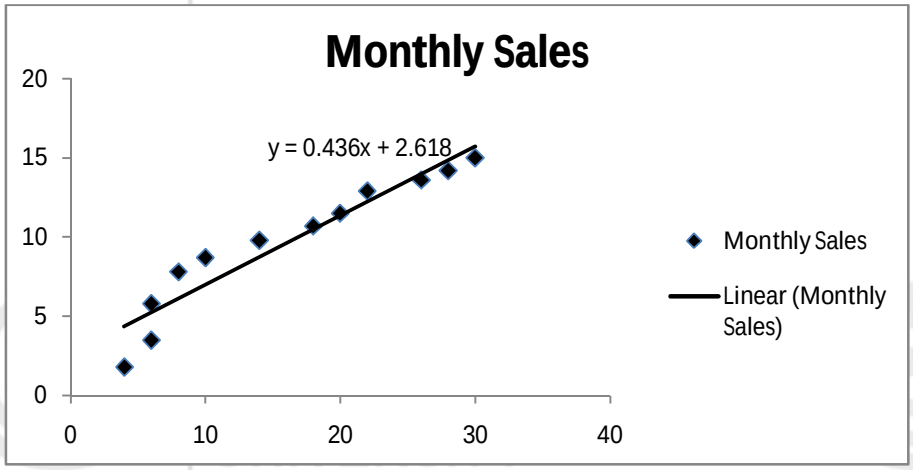


Fig. 9.14: Best fitted regression equation for the given set of (X, Y) values.

E7) Refer to Sec. 9.5.

E8) In Example 2, we have fitted the simple linear regression equation $Y = a + bX + e$ for the given set of data and obtained the best fitted regression equation as

$$\hat{Y} = 2.6185 + 0.4369X$$

From the above equation of regression, we calculate the fitted values of the dependent variable Y, i.e., $\hat{Y}_i$ for all i, and obtain the values of residuals, i.e., r_i for all i, as given in the following table:

S. No.	Y Values	$\hat{Y}$ Values	r_i	r_i^2
1	1.8	4.36	−2.56	$(-2.56)^2 = 6.5536$
2	3.5	5.24	−1.74	$(-1.74)^2 = 3.0276$
3	5.8	5.24	+0.56	$(+0.56)^2 = 0.3136$
4	7.8	6.11	+1.69	$(+1.69)^2 = 2.8561$
5	8.7	6.99	+1.71	$(+1.71)^2 = 2.9241$
6	9.8	8.73	+1.07	$(+1.07)^2 = 1.1449$
7	10.7	10.48	+0.22	$(+0.22)^2 = 0.0484$
8	11.5	11.36	+0.14	$(+0.14)^2 = 0.0196$
9	12.9	12.23	+0.67	$(+0.67)^2 = 0.4489$
10	13.6	13.98	−0.38	$(-0.38)^2 = 0.1444$
11	14.2	14.85	−0.65	$(-0.65)^2 = 0.4225$
12	15.0	15.73	−0.73	$(-0.73)^2 = 0.5329$
Total	$\sum Y_i = 115.30$	$\sum \hat{Y}_i = 115.30$	$\sum r_i = 0.00$	$\sum r_i^2 = 18.4366$

To understand how well the regression line fits the given data, we first calculate the standard deviation of the residuals:

$$SS_{Res}^{1/2} = \sqrt{\frac{\sum r_i^2}{n-k}}$$

$$= \sqrt{\frac{18.4366}{10}}$$

$$= \sqrt{1.84366} = 1.3578$$

where $n - k = 10$

We now construct the following table for obtaining the values of standardised residuals for all values of the residuals:

S. No.	Residuals r_i	Standardised Residuals $d_i = r_i / \sqrt{SS_{Res}}$
1	−2.56	−1.8854
2	−1.74	−1.28148
3	+0.56	+ 0.41243
4	+1.69	+ 1.24466
5	+1.71	+ 1.25939
6	+1.07	+ 0.78803
7	+0.22	+ 0.16202
8	+0.14	+ 0.10310
9	+0.67	+ 0.49344
10	−0.38	− 0.27986
11	−0.65	−0.47871
12	−0.73	−0.53763
Total	$\sum r_i = 0.00$	$\sum d_i = 0.00$

The standardised residuals are obtained as above and it is observed that their total is zero. We plot these residuals against the fitted values of Y as shown in Fig. 9.15.

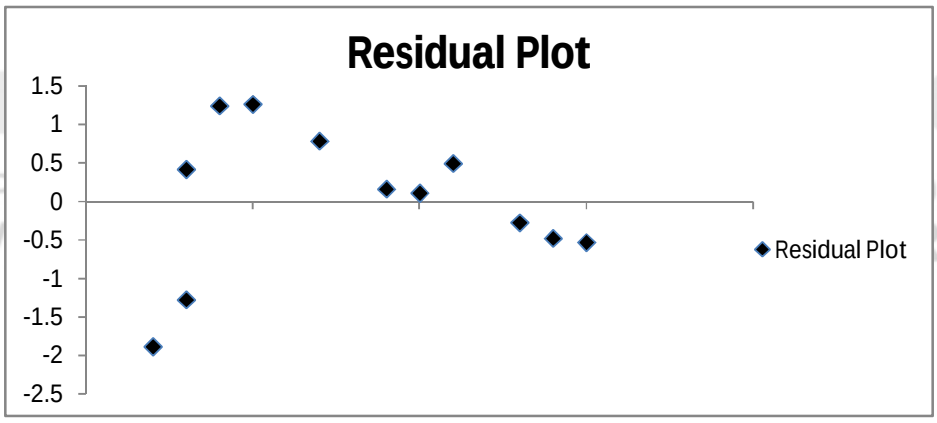


Fig. 9.15: Residual plot for the given data.

Then we follow the method of constructing a normal probability plot as explained in Sec 9.5.3. We arrange the standardised residuals in increasing order and rank the residuals, which are arranged in increasing order. After ranking the residuals, we calculate the cumulative probability using equation (26) and percentile cumulative probability using equation (27) for the corresponding ranked residual values.

Ordered Residual	**Ranks (i)**	**Cumulative Probability** $p_i = (i - 1/2)/n$	**Percentile Cumulative Probability** $P_i = \{(i - 1/2)/n\} \times 100$
–1.86495	1	0.05	05
–0.70882	2	0.15	15
–0.3083	3	0.25	25
–0.26391	4	0.35	35
–0.12972	5	0.45	45
0.092215	6	0.55	55
0.315183	7	0.65	65
0.35957	8	0.75	75
0.760086	9	0.85	85
1.738667	10	0.95	95

Then we plot the percentile cumulative probability versus the standardised residuals to obtain the normal probability plot. The resulting points would lie approximately on a straight line as shown in Fig. 9.16.

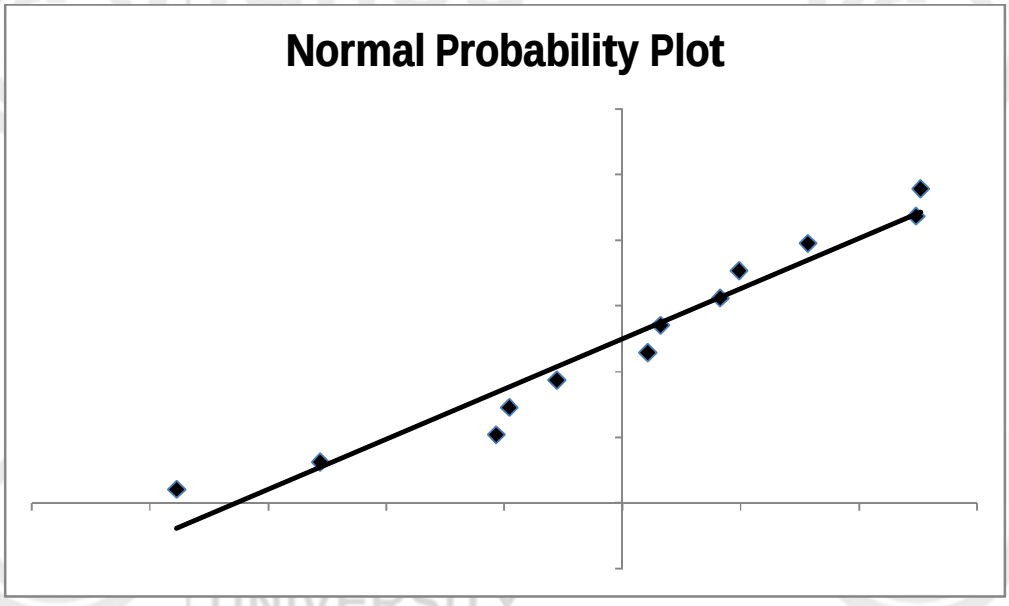


Fig. 9.16: Normal probability plot for the given data.

From the normality probability plot (Fig. 9.16), you may note that some points of the distribution are deviating slightly from the straight line, but do not lie very far from the central points.

On the basis of the above residual analysis for the given data, we obtain the best fitted regression equation shown in Fig. 9.17.

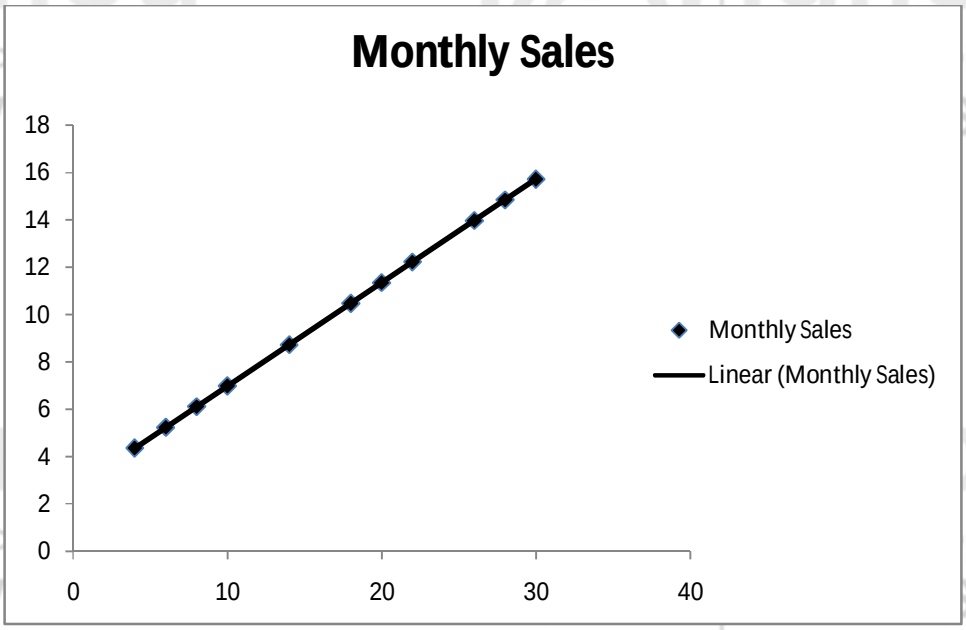


Fig. 9.17: Best fitted regression equation after residual analysis for the given data.