Chapter VI

Correlation
and
Simple Linear Regression Analysis

D. 6. 1. (***Correlation Analysis***)
The *correlation analysis* means the study of existence, magnitude and direction of the relation between two or more variables.

D. 6. 2. (***Linear, Non-Linear Correlation***)
The correlation between two variables is said to be *linear* if the ratio of change is constant.
It is *non-linear* if the ratio of change is not constant.

D. 6. 3. (***Positive, Negative Correlation***)
If two variables change in the same direction, then this is called a *positive correlation*.
(For example: price and supply)
If two variables change in the opposite direction, then the correlation is called a *negative correlation.*
(For example: price and demand)

R. 6. 1. (***Degree of Correlation***)

Degrees	Positive	Negative
Absence of correlation	0	0
Perfect correlation	+1	-1
High degree	$]0.75,\ 1[$	$]-1,\ -0.75[$
Moderate degree	$]0.25,\ 0.75[$	$]-0.75,\ -0.25[$
Low degree	$]0,\ 0.25[$	$]-0.25,\ 0[$

R. 6. 2. (***Methods of Determining Correlation***)
We shall consider the following most commonly used methods:

(1) Scatter Plot
(2) Karl Pearson's coefficient of correlation
(3) Spearman's rank correlation coefficient.

R. 6. 3. (***Scatter Plot or Dot Diagram***)
In this method the values of the two variables are plotted on a graph paper. One is taken along the horizontal ($x-$) axis and the other along the vertical ($y-$) axis. By plotting the data, we get points (dots) on the graph which are generally scattered and hence the name 'scatter plot'. The manner in which these points are scattered, suggest the degree and the direction of correlation.

Let the degree of correlation be denoted by r. Its direction is given by the signs positive and negative.

i) If all points lie on a rising straight line the correlation is perfectly positive and $r = +1$
ii) If all points lie on a falling straight line the correlation is perfectly negative and $r = -1$
iii) If the points lie in a narrow strip, rising upwards, the correlation is high degree of positive.
iv) If the points lie in a narrow strip, falling downwards, the correlation is high degree of negative.
v) If the points are spread widely over a broad strip, rising upwards, the correlation is low degree positive.
vi) If the points are spread widely over a broad strip, falling downwards, the correlation is low degree negative.
vii) If the points are spread (scattered) without any specific pattern, the correlation is absent, i.e. $r = 0$

Though this method is simple and gives a rough idea about the existence and the degree of correlation, it is not reliable. As it is not an exact mathematical method, it cannot measure the degree of correlation.

D. 6. 4. (***Karl Pearson's Coefficient of Correlation***)
Let X and Y be two variates.

$$r := \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\cdot\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \cdot \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}.$$

R. 6. 4.
It can be shown:

$$r = \frac{n\cdot\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left(n\cdot\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right).\left(n\cdot\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right)}}.$$

R. 6. 5.
Correlation measures the *linear relationship* between two variables. If there is a nonlinear relationship, the correlation value may be deceptive.

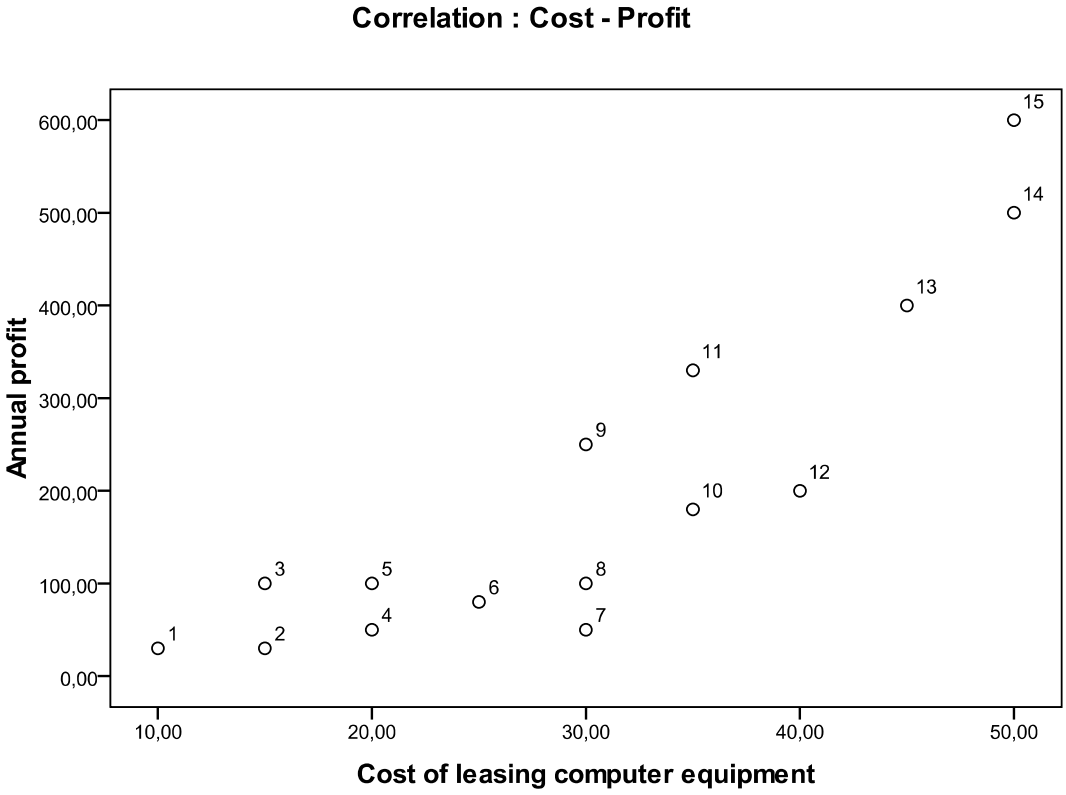
Ex. 6. 1.
The following table shows the annual profit [Mio €] and annual costs of leasing computer equipment [1000 €] of 15 firms:

Working Table

i	x_i	$(x_i - \bar{x})$	y_i	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	10	-20	30	-170	3400	400	28900
2	15	-15	30	-170	2550	225	28900
3	15	-15	100	-100	1500	225	10000
4	20	-10	50	-150	1500	100	22500
5	20	-10	100	-100	1000	100	10000
6	25	-5	80	-120	600	25	14400
7	30	0	50	-150	0	0	22500
8	30	0	100	-100	0	0	10000
9	30	0	250	50	0	0	2500
10	35	5	180	-20	-100	25	400
11	35	5	330	130	650	25	16900
12	40	10	200	0	0	100	0
13	45	15	400	200	3000	225	40000
14	50	20	500	300	6000	400	90000
15	50	20	600	400	8000	400	160000
Sum	450	0	3000	0	28100	2250	457000

$$\bar{x} = \frac{450}{15} = 30, \qquad \bar{y} = \frac{3000}{15} = 200$$

$$r = \frac{28100}{\sqrt{2250 \cdot 457000}} \approx 0.88.$$



D. 6. 5. (***Spearman's Rank Coefficient of Correlation***)
Let

$R_i, i = 1, 2, ..., n$: ranks of the characteristic X

$R_i^{'}, \ i = 1, 2, ..., n$: ranks of the characteristic Y,

$$\rho := 1 - \frac{6 \cdot \sum_{i=1}^{n} \left(R_i - R_i^{'} \right)^2}{(n-1) \cdot n \cdot (n+1)} :$$

Ex. 6. 2.
The following table shows the advertisement costs (Y) and the revenues (X) of a firm:

Advertisement Costs	Revenue
1.4	210
1.8	220
1.9	240
2.4	240
2.8	320
3.2	400
3.6	410
4.0	480

Find and interpret the Spearman's rank coefficient of correlation.

Solution:

Advertisement Costs	Revenue	R_i	$R_i^{'}$	$(R_i - R_i^{'})^2$
1.4	210	1	1.0	0.00
1.8	220	2	2.0	0.00
1.9	240	3	3.5	0.25
2.4	240	4	3.5	0.25
2.8	320	5	5.0	0.00
3.2	400	6	6.0	0.00
3.6	410	7	7.0	0.00
4.0	480	8	8.0	0.00
				0.50

$$n = 8, \qquad \sum_{i=1}^{8} \left(R_i - R_i^{'} \right)^2 = 0.50,$$

$$\rho := 1 - \frac{6 \cdot 0.50}{7.8.9 \cdot} \approx 0.994.$$

There is therefore a strong degree of correlation between X and Y.

D. 6. 6. (***Simple Linear Regression Model***)
A *simple linear regression model* is defined as follows:

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

β_0 and β_1 are referred to as the *parameter* of the model, and ε is a random variable referred to as the *error* term.

D. 6. 7. (***Simple Linear Regression Equation***)
The equation that describes how the expected value of y, denoted by $E(y)$, is related to x is called the *regression equation*:

$$E(y) = \beta_0 + \beta_1 x.$$

D. 6. 8. (***Estimated Linear Regression Equation***)
Substituting the values of the sample statistics b_0 and b_1 for β_0 and β_1, we obtain the *estimated linear regression equation*:

$$y^* = b_0 + b_1 x.$$

R. 6. 6. (***Least Square Method***)
The coefficients of a regression function can be estimated by using the *Least Square Method*:

$$S(...) = \sum_{i=1}^{n} \left(y_i - y^* \right)^2 \rightarrow Min!$$

R. 6. 7.

Applying least squares method to the simple linear regression function, we obtain the following "*normal equations*" .

$$\begin{cases} n \cdot b_0 + b_1 \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ b_0 \cdot \sum_{i=1}^{n} x_i + b_1 \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot y_i \end{cases}$$

Ex. 6. 3.

Suppose the following data were collected from a sample of 10 pizzerias of a certain firm:

Restaurant	**Student Population (1000s)**	**Quarterly Sales(1000s)**
i	x_i	y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Find a regression function describing the quarterly sales as a linear function of the student population.

Solution:

Working Table

x_i	y_i	x_i^2	$x_i \cdot y_i$	y_i^2
2	58	4	116	3364
6	105	36	630	11025
8	88	64	704	7744
8	118	64	944	13924
12	117	144	1404	13689
16	137	256	2192	18769
20	157	400	3140	24649
20	169	400	3380	28561
22	149	484	3278	22201
26	202	676	5252	40804
140	1300	2528	21040	184730

$$\begin{cases} 10b_0 \quad 140b_1 = 1300 \\ 140b_0 + 2528b_1 = 21040 \end{cases} \qquad b_0 = 60, \quad b_1 = 5$$

$$y^* = 60 + 5x$$

R. 6. 8.

The coefficients b_0 and b_1 can also be calculated by the following formulas:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

where

$\bar{x}$ = mean value for the independent variable

$\bar{y}$ = mean value for the dependent variable

Ex. 6. 3. (***cont'd***)

Working Table

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	-12	-72	864	144
2	6	105	-8	-25	200	64
3	8	88	-6	-42	252	36
4	8	118	-6	-12	72	36
5	12	117	-2	-13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Total	140	1300	0	0	2840	568

$$b_1 = \frac{2840}{680} = 5, \qquad b_0 = 130 - 5 \cdot 14 = 60.$$

D. 6. 9. (***Coefficient of Determination***)
The *coefficient of determination* is defined as:

$$r^2 := \frac{SSR}{SST}$$

where:

$$SST = SSR + SSE$$

$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$: total sum of squares

$SSR = \sum_{i=1}^{n} (y_i^* - \bar{y})^2$: sum of square due to regression

$SSE = \sum_{i=1}^{n} (y_i - y_i^*)^2$: sum of squares due to error

R. 6. 9.
r^2 $(0 \le r^2 \le 1)$ can be interpreted as the percentage of the total sum of squares that can be explained by using the estimated regression equation.

Ex. 6. 3. (***cont'd***)

Working Table

i	x_i	y_i	y_i^*	$y_i - y_i^*$	$(y_i - y_i^*)^2$	$(y_i - \bar{y})^2$
1	2	58	70	-12	144	5184
2	6	105	90	15	225	625
3	8	88	100	-12	144	1764
4	8	118	100	18	324	144
5	12	117	120	-3	9	169
6	16	137	140	-3	9	49
7	20	157	160	-3	9	729
8	20	169	160	9	81	1521
9	22	149	170	-21	441	361
10	26	202	190	12	144	5184
	140	1300	1300	0	1530	15730

$$SSR = SST - SSE = 15730 - 1530 = 14200$$

$$r^2 = \frac{14200}{15730} = 0.9027.$$

We can, therefore, conclude that 90.27% of the total sum of squares can be explained by using the regression equation

$$y^* = 60 + 5x$$

to predict quarterly sales. In other words, 90.27% of the variability in sales can be explained by the linear relationship between the size of the student population and sales. It is a good fit for the estimated regression equation.

D. 6. 10. (***Simple Correlation Coefficient***)

$$r := (\text{sign of } b_1)\sqrt{r^2}, \quad -1 \le r \le 1.$$

Ex. 6. 3. (***cont'd***)

$$r = \sqrt{0.9027} = 0.9501.$$

R. 6. 10.

The coefficient of determination for a simple linear regression function can alternatively be calculated according to the following formula:

$$r = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\left(n \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i\right)\left(n \cdot \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} y_i\right)}}, \quad -1 \le r \le 1$$

Ex. 6. 3. (***cont'd***)

$$r = \frac{10 \cdot 21040 - 140 \cdot 1300}{\sqrt{\left(10 \cdot 2528 - 140^2\right) \cdot \left(10 \cdot 184730 - 1300^2\right)}} \approx 0.9501.$$

R. 6. 11. (***Model Assumptions***)

1. The error term ε is a random variable with an expected value of zero: $E(\varepsilon) = 0$.
   Implication: β_0 and β_1 are constants, therefore $E(\beta_0) = \beta_0$ and $E(\beta_1) = \beta_1$; thus for a given value of x, the expected value of y is

   $$E(y) = \beta_0 + \beta_1 x \text{ (regression equation)}$$

2. The variance of ε, denoted by σ^2, is the same for all values of x.
   Implication: The variance of y about the regression line equals σ^2 and is the same for all values of x.
3. The values of ε are independent.
   Implication: The value of ε for a particular value of x is not related to the value of ε for any other value of x; thus the value of y for a particular value of x is not related to the value of y for any other value of x.
4. The error term ε is a normally distributed random variable.
   Implication: Because y is a linear function of, y is also a normally distributed random variable.

R. 6. 12. (***Testing for Significance***)
In a simple linear equation, the expected value of y is a linear function of $y: E(y) = \beta_0 + \beta_1 x$. If the value of β_1 is zero, $E(y) = \beta_0 + 0 \cdot x = \beta_0$. In this case, the mean value of y does not depend on the value of x and hence we would conclude that x and y are not linearly related. Alternatively, if the value of β_1 is not equal to zero, we would conclude that the two variables are related.
Thus, to test for a significant regression relationship, we must conduct a hypothesis test to determine whether the value of β_1 is zero.
Two tests are commonly used:

1. t Test
2. F Test

Both tests require an estimate of σ^2, the variance of ε in the regression model.

R. 6. 13. (***t Test for Significance in Simple Linear Regression***)

Step 1:
Formulate the test hypotheses:

$$H_0: \beta_1 = 0, \qquad H_1: \beta_1 \neq 0$$

Step 2:
Calculate the test statistic:

$$t = \frac{b_1}{s_{b_1}}$$

where

$$s_{b_1} = \frac{s}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}},$$

$$s = \sqrt{\frac{SSE}{n-2}}$$

Step 3:
Conclusion:

$$\text{Reject } H_0 \text{ if } \begin{cases} p - value \leq \alpha & p - value \text{ approach} \\ t \leq -t_{\alpha/2} \text{ or } t \geq t_{\alpha/2} & \text{critical value approach} \end{cases}$$

Ex. 6. 3. (***cont'd***)
Let the significance level be equal to $\alpha = 0.01$.

Step 1:

$$H_0: \beta_1 = 0, \qquad H_1: \beta_1 \neq 0$$

Step 2:

$$s = \sqrt{\frac{1530}{10-2}} = 13.829$$

$$s_{b_1} = \frac{13.829}{\sqrt{568}} = 0.580$$

$$t = \frac{5}{0.5803} = 8.62$$

Step 3:

p-value approach: $p-value < 0.0001 < 0.01 = \alpha \quad \Rightarrow$ reject H_0

critical value approach: $t = 8.62 \geq 3.355 = t_{8;0.01} \quad \Rightarrow$ reject H_0

R. 6. 14. (***F Test for Significance in Simple Linear Regression***)

Step 1:
Formulate the test hypotheses:

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0$$

Step 2:
Calculate the test statistic:

$$F = \frac{MSR}{MSE}$$

where

$$MSR = \frac{SSR}{\text{Number of independent variables}}$$

$$MSE = s^2$$

Step 3:
Conclusion:

$$\text{Reject } H_0 \text{ if} \begin{cases} p-value \leq \alpha & p-value \text{ approach} \\ F \geq F_\alpha & \text{critical value approach} \end{cases}$$

(F_α is based on an F distribution with 1 degree of freedom in the numerator and $n-2$ degree of freedom in the denominator.)

Ex. 6. 3. (***cont'd***)
Let the significance level be equal to $\alpha = 0.01$.

Step 1:

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0$$

Step 2:

$$F = \frac{14200}{191.25} = 74.25$$

Step 3:

p-value approach: $p - value < 0.0001 < 0.01 = \alpha \Rightarrow$ reject H_0

critical value approach: $F = 74.25 \geq 11.26 = F_{0.01} \Rightarrow$ reject H_0

R. 6. 15. (***ANOVA Table***)
The following *ANOVA* table can be used to summarise the results of the F test:

ANOVA Table

Source of variation	**Sum of Squares**	**Degree of Freedom**	**Mean Square**	***F***
Regression	SSR	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Error	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$	
Total	SST	$n-1$		

Ex. 6. 3. (***cont'd***)

ANOVA Table

Source of variation	**Sum of Squares**	**Degree of Freedom**	**Mean Square**	***F***
Regression	14200	1	$\frac{14200}{1} = 14200$	$\frac{14200}{191.25} = 74.25$
Error	1530	8	$\frac{1530}{8} = 191.25$	
Total	15730	9		

R. 6. 16. (***Test of Significance for Correlation***)
A test of significance for a linear relationship between *x* and *y* can also be performed by using the sample correlation r. With ρ denoting the population correlation coefficient, the hypotheses are as follows:

$$H_0: \rho = 0, \quad H_1: \rho \neq 0.$$

A significant relationship can be concluded if H_0 is rejected. However, the t and F tests give the same result as the test for significance using the correlation coefficient.

R. 6. 17. (*Confidence Interval for β_1 *)

The form of *confidence interval* for β_1 is as follows:

$$\beta_1 \in \left[b_1 - t_{\alpha/2} s_{b_1},\ b_1 + t_{\alpha/2} s_{b_1} \right]$$

Ex. 6. 3. (*cont'd*)

$$\beta_1 \in \left[5 - 3.355 \cdot 0.5803,\ 5 + 3.355 \cdot 3.5803 \right] = \left[3.05,\ 6.95 \right].$$

R. 6. 18. (*Using the Estimated Regression Equation for Estimation and Prediction)*

Here we have the following possibilities:

1. Point estimation
2. Interval estimation for the mean value of y
3. Interval prediction for an individual value of y

R. 6. 19. (*Point Estimation)*

By substituting a certain value for x in the regression function, we obtain a point estimate.

Ex. 6. 3. (*cont'd*)

Predict the sales for a university with 10000 students.

Solution:

$$y^*(10) = 60 + 5 \cdot 10 = 110 \text{ (or \$110000).}$$

R. 6. 20. (*Confidence Interval for the Mean of y)*

The form of *confidence interval* for the *mean value of y* is as follows:

$$E(y_p) = \left[y_p^* - t_{\alpha/2} \cdot s_{y_p^*},\ y_p^* + t_{\alpha/2} \cdot s_{y_p^*} \right]$$

where

x_p : the particular given value for x

y_p : the value of y corresponding to the given x_p

$E(y_p)$: the expected value of y

$y_p^* = b_0 + b_1 x_p$: the point estimate of $E(y_p)$, when $x = x_p$

$$s_{y_p^*} = s \cdot \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Ex. 6. 3. (***cont'd***)
With $x_p = 10$ and $\alpha = 0.05$ we obtain:

$$s_{y_p^*} = 13.829 \cdot \sqrt{\frac{1}{10} + \frac{(10-14)^2}{568}} = 4.95$$

$$E(y_p) = [110 - 2.306 \cdot 4.95,\ 110 + 2.306 \cdot 4.95] = [98.585,\ 121.415].$$

R. 6. 21. (***Confidence Interval for an Individual Value of y)***
The form of *confidence interval* for *an individual value of y* is as follows:

$$y_p \in \left[y_p^* - t_{\alpha/2} \cdot s_{ind},\ y_p^* + t_{\alpha/2} \cdot s_{ind} \right]$$

where

$$s_{ind} = s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}.$$

Ex. 6. 1. (***cont'd***)
With $x_p = 10$ and $\alpha = 0.05$ we obtain:

$$s_{ind} = 13.829 \cdot \sqrt{1 + \frac{1}{10} + \frac{(10-14)^2}{568}} = 14.69$$

$$y_p \in [110 - 2.306 \cdot 14.69,\ 110 + 2.306 \cdot 14.69] = [76.125,\ 143.875].$$

R. 6. 22. (***Residual Analysis)***
The residuals, $y_i - y_i^*$, provide the best information about ε ; hence an analysis of the residuals is an important step in determining whether the assumption for ε are appropriate. Much of residual analysis is based on an examination of graphical plots:

1. *Plot of residuals against values of the independent variable x*
   A residual plot against the independent variable x is a graph in which the values of the independent variable are represented by the horizontal axis and the corresponding residual values are represented by the vertical analysis.

2. *Plot of residuals against the predicted values of the dependent variable y**
   A residual plot against y^* represents the residual values on the vertical axis.

3. *Standardised residual plot*
   Here, the standardised residuals will be represented against the independent variable x.

4. *Normal probability Plot*
   The normal probability plot is a graphical technique for assessing whether or not the data set is approximately normally distributed.
   The data are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line. Departures from this straight line indicate departures from normality