UNIT 7 MULTIPLE LINEAR REGRESSION

Structure

7.1 INTRODUCTION

In Unit 5, we have discussed the simple linear regression model that could be used for explaining linear relationship between a response variable and a single regressor variable. In Unit 6, we have discussed the testing of the significance and computation of the confidence intervals of regression coefficients for simple linear regression. In many real life situations, we may require extending such considerations to two or more regressor variables. There are many situations in which we need to explore the relationship between one response variable with more than one regressor variable. Some examples are given below.

1. The response to a drug may depend upon the dosage, age, body weight and other related factors.
2. The length of a patient's stay in a hospital may depend on the disease condition, age, income, etc.
3. The height or weight of a person may vary with age, gender, nutritional intake, etc.
4. The blood pressure or heart-beats may be related to age, weight, physical activities, diet, etc.
5. The lean body mass may be estimated with the help of age, height, BMI (Body Mass Index), fat component of the body, etc.

In all these examples and in many more real life situations, multiple regression analysis is useful in quantifying the relative roles of various regressor variables. Multiple regression examines the effect of more than one regressor variable on the response variable at the same time.

Regression models that are not linear are called non-linear models.

Therefore, in this unit, we shall explain the regression model for determining the relationship between a response variable and more than one regressor variable. In Sec. 7.2, we consider the case of two regressor variables and then generalise it to the case of k regressor variables. We also explain how to construct the scatter plot matrix. In Sec. 7.3, we explain how to estimate the regression coefficients by manually solving the normal equations, which are obtained through the method of least squares. Sec. 7.4 deals with the matrix

approach to fit the multiple regression model in case of more than two regressor variables. Finally, we explain residual analysis of the fitted multiple linear regression model in Sec. 7.5.

Objectives

After studying this unit, you should be able to:

- estimate the regression coefficients of the multiple regression model using the method of least squares;
- fit a multiple linear regression model;
- predict the value of the response variable for given values of the regressor variables;
- apply the matrix approach to estimate the regression coefficients; and
- validate some assumptions of the multiple linear regression model using residual and normal probability plots.

7.2 MULTIPLE LINEAR REGRESSION

Recall the simple linear regression model explained in Unit 5 which is given by

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad \ldots (1)$$

Suppose that the response of a drug depends on the dosage and age. Let Y denote the response of the drug, X_1, the dosage and X_2, the age.

We define a multiple linear regression model for approximating the relationship between a quantitative response variable Y and two quantitative regressor variables (X_1 and X_2) as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \quad \ldots (2)$$

where ε is an error term which follows normal distribution with mean 0 and variance σ^2 as in the case of simple linear regression model. The parameters β_0, β_1 and β_2 are unknown and are called the partial regression coefficients. The parameter β_0 is known as intercept and β_1 and β_2, the slopes corresponding to the respective regressor variables.

This is a multiple linear regression model with two regressor variables. It is a linear model because equation (2) is linear in the parameters β_0, β_1 and β_2 as well as in both the variables.

Here β_0 represents the intercept of the regression plane. β_1 indicates the expected change in the response of drug due to per unit change in the dose when age (X_2) is held to be fixed. Similarly, β_2 represents the expected change in response due to per unit change in age (X_2) when the effect of dose (X_1) is held to be fixed.

7.2.1 Scatter Matrix Plot

In Unit 5, you have learnt how to plot a scatter diagram to get an approximate idea about the relationship between the response and regressor variables. You

know that the scatter plot is a two-dimensional diagram. Therefore, we cannot plot it for multiple regression analysis. However, many computer softwares are available to plot a three-dimensional scatter plot to check the dependence of response variable on two regressor variables. Here, we will use the **matrix** of scatter plots (scatterplot matrix) to get a rough idea of relationship among the variables. A **scatterplot matrix** contains all pairwise scatter plots of the variables in a matrix format. For example, let us consider the case of two regressor variables. Then we create scatter plots for the combinations of two variables, e.g., (Y, X_1), (Y, X_2), (X_1, X_2), etc., to assess the relationship between them. After creating all scatter plots, we put them together in the form of a 3 × 3 matrix so that they look like a scatterplot matrix as shown in Fig. 7.1 for Example 1.

If there are k regressor variables, the scatterplot matrix will have (k + 1) rows and (k + 1) columns. In this way, the scatterplot matrix is a two dimensional array [(k +1)×(k + 1)] of the scatter plots where every cell except the diagonal cells contains the respective scatter plot. The first row gives an idea about the relationship of the response variable with the $(1^{st,}\ 2^{nd}, ..., k^{th})$ regressor variables, respectively. The second row shows the relationship of the first regressor variable with the response variable and the $(2^{nd}, 3^{rd}, ..., k^{th})$ regressor variables, while the $(k+1)^{th}$ row illustrates the relationship of the k^{th} regressor variable with the response variable and the $(1^{st,}\ 2^{nd}, ..., (k-1)^{th})$ regressor variables.

Let us consider the following example to learn how to plot a scatterplot matrix.

Example 1: A researcher is interested in analysing the relationship of systolic blood pressure (SBP) with age and weight of women, and selects a sample of 15 women of age group 25-40 years. The data on SBP, age and weight are given below:

S. No.	SBP (mm/Hg)	Age (years)	Weight (kg)
1	124	30	71
2	134	38	82
3	135	39	98
4	121	26	72
5	122	29	70
6	119	27	72
7	128	32	76
8	118	25	54
9	120	26	58
10	123	31	68
11	129	37	63
12	117	25	62
13	131	35	92
14	126	34	75
15	134	40	89

Check whether there is a linear relationship among SBP, age and weight using the scatterplot matrix.

Solution: We construct the scatterplot matrix considering pairs of variables and arrange them as shown in Fig.7.1.

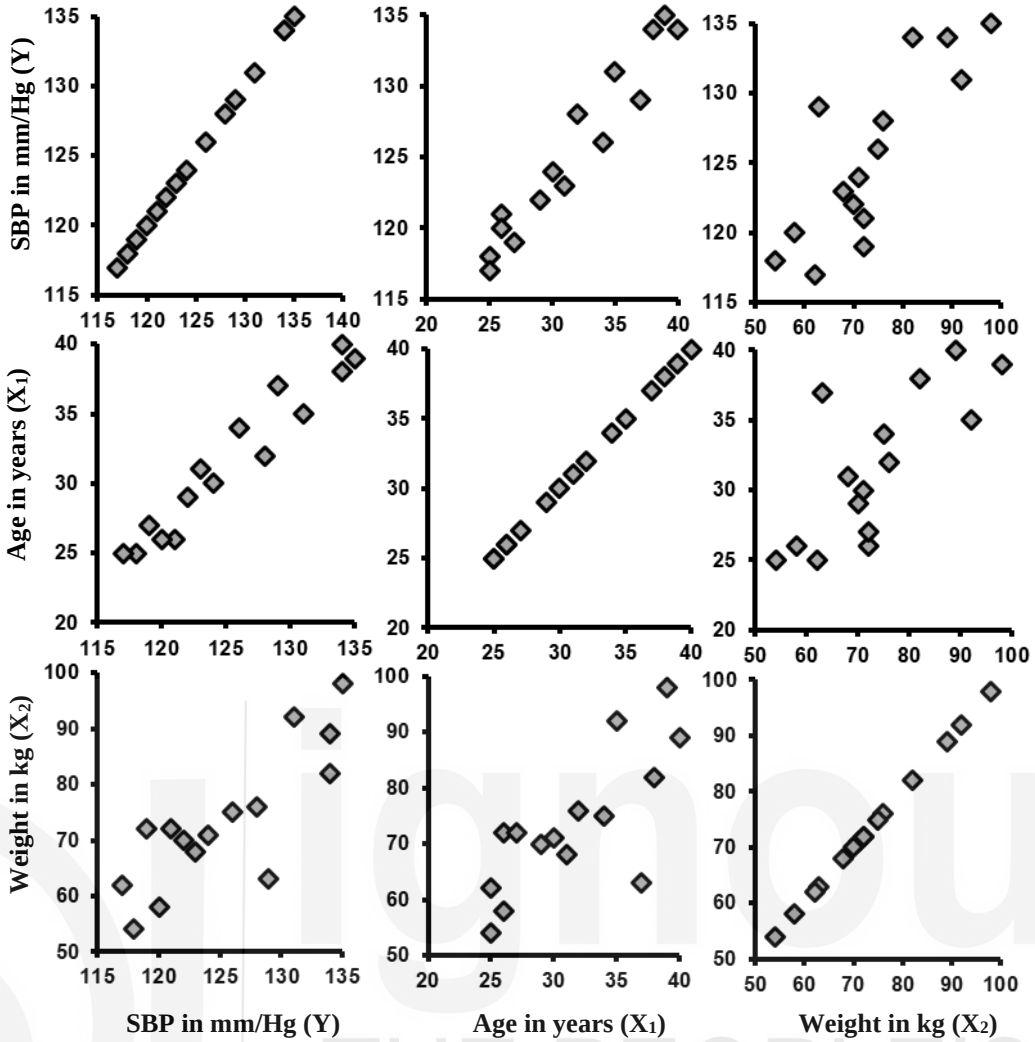

Fig. 7.1: Scatterplot matrix.

Notice that the scatterplot matrix shown in Fig. 7.1 depicts the approximate linear relationship among all three variables.

7.2.2 Generalisation of Multiple Regression Model

In many real life situations, the response variable depends on more than two regressor variables. For example, the response of a drug may be related to the dosage, age, weight, gender, condition of the patient, etc.

As a matter of fact, we may assume that as we increase the number of regressor variables in the model, the prediction is likely to be better. However, some other side effects may occur if we consider more regressor variables in the model. Therefore, before finalising the model, it is necessary to check the appropriate number of regressor variables to be included in the model for making it the best predictor.

In this subsection, we generalise the regression model defined in equation (2) for k (k > 2) regressor variables. In the multiple regression models, it is assumed that a linear relationship exists between response variable Y and k regressor variables, say, $X_1, X_2, ..., X_k$. We define a multiple regression model for approximating the relationship between a response variable Y and a set of k regressor variables as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon \quad \ldots (3)$$

where $\beta_0, \beta_1, \beta_2, ..., \beta_k$ are known as the regression coefficients.

This is called a multiple linear regression model with k-regressor variables. The parameter β_i (i = 1, 2, …, k) represents the expected change in the response Y corresponding to per unit change in X_i when all other remaining regressor variables $X_j's \left(j \neq i, j = 1, 2, ..., k \right)$ are held to be fixed.

Being the coefficients of regressor variables, these regression coefficients are assumed to be unknown. So we need to estimate them from empirical data. We will discuss this aspect in the next section.

Now, the regressor variable X may not necessarily have a linear relationship with variable Y and could be any function of X. Thus, X may appear in the model as a quadratic, cubic or inverse function. For example, we can define k^{th} order polynomial regression model as:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_k X^k + \varepsilon \quad ... (4)$$

Then we use the transformations $X_1 = X$, $X_2 = X^2$, ..., $X_k = X^k$, and rewrite the model (4) as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon \quad ... (5)$$

So by choosing appropriate transformations such as the one in equation (5), we can transform a k^{th} degree polynomial regression model defined in equation (4), into a multiple linear regression model. However, we shall not consider such transformations in this block.

Before studying the next section, you may like to solve the following exercises.

E1) Explain the assumptions of multiple linear regression model.

E2) Construct the scatterplot matrix for the data on the average duration of walk (minutes), serum creatinine (mg/dL) and random blood sugar level (mg/dL) of 15 diabetic patients given in the following table:

S. No.	Random Blood Sugar (mg/dL)	Duration of Walk (minutes)	Serum Creatinine (mg/dL)
1	430	20	0.35
2	420	25	0.35
3	410	30	0.45
4	400	30	0.5
5	390	45	0.65
6	395	35	0.55
7	420	30	0.55
8	410	35	0.6
9	400	35	0.6
10	390	40	0.65
11	380	45	0.75
12	370	50	0.65
13	390	35	0.65
14	365	55	0.6
15	325	45	0.65

7.3 ESTIMATION OF THE REGRESSION COEFFICIENTS

In this section, we explain how to estimate the unknown regression coefficients of the multiple linear regression model using the method of least squares. You

have learnt this method for the simple linear regression model in Sec. 5.4 of Unit 5. You know that this method minimises the error sum of squares, i.e., sum of the squares of the differences between observed value and true value of the response variable.

For the sake of simplicity, we first consider the multiple regression model with two regressor variables. We shall discuss the case of more than two regressor variables in the next sub-section.

7.3.1 Multiple Regression Model with Two Regressor Variables

Suppose we have response variable Y and two regressor variables, say, X_1 and X_2. Then equation (2) defines the multiple linear regression model for two regressor variables. Since the parameters β_0, β_1 and β_2 are unknown, we estimate the values of these regression coefficients using the method of least squares.

For fitting a multiple linear regression model, we select a random sample of n observations, in which each observation consists of an observed value of the response variable along with the values of each regressor variable. We can also represent the observed sample values of the variables Y, X_1 and X_2 in a tabular form with n rows as shown in Table 1.

Table 1: Samples values of Y, X_1 and X_2

Y	X_1	X_2
y_1	x_{11}	x_{21}
y_2	x_{12}	x_{22}
.	.	.
.	.	.
.	.	.
y_n	x_{1n}	x_{2n}

For n pairs of data, we can rewrite the multiple linear regression model for the i^{th} observation, i.e., $(y_i, x_{1i}, \text{and } x_{2i})$ as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \quad ; i = 1, 2, ..., n$$

where ε_i is the i^{th} error term, i.e., the deviation of the observed value of the response variable from its actual value. Here also we assume that the error terms (ε_i's) are independently and normally distributed with mean zero and variance σ^2 (constant), i.e., $E(\varepsilon_i) = 0, V(\varepsilon_i) = \sigma^2$ and $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$; $i \neq j = 1, 2, ..., n$. Recall that we made the same assumptions for simple linear regression.

For the i^{th} observation, we write the error as:

$$\varepsilon_i = (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})$$

$$\therefore \quad \varepsilon_i^2 = (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2$$

Then we take the sum for all values of i and write

$$E = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2 \qquad \ldots (6)$$

We now successively differentiate the error sum of squares (E), with respect to the unknown parameters (β_0, β_1 and β_2) and equate the derivatives to zero as

explained in Sec. 5.3. In this way, we obtain three normal equations, each corresponding to unknown parameters β_0, β_1 and β_2, respectively, as shown below:

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_{1i} + \beta_2 \sum_{i=1}^{n} x_{2i} = \sum_{i=1}^{n} y_i \quad \ldots (7)$$

$$\beta_0 \sum_{i=1}^{n} x_{1i} + \beta_1 \sum_{i=1}^{n} x_{1i}^2 + \beta_2 \sum_{i=1}^{n} x_{2i} x_{1i} = \sum_{i=1}^{n} y_i x_{1i} \quad \ldots (8)$$

$$\beta_0 \sum_{i=1}^{n} x_{2i} + \beta_1 \sum_{i=1}^{n} x_{1i} x_{2i} + \beta_2 \sum_{i=1}^{n} x_{2i}^2 = \sum_{i=1}^{n} y_i x_{2i} \quad \ldots (9)$$

We rewrite equation (7) as:

$$\beta_0 = \bar{y} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2 \quad \ldots (10)$$

For the sake of simplicity, we substitute the value of β_0 given in equation (10) in equations (8) and (9). We then obtain

$$\left(\bar{y} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2\right) \sum_{i=1}^{n} x_{1i} + \beta_1 \sum_{i=1}^{n} x_{1i}^2 + \beta_2 \sum_{i=1}^{n} x_{2i} x_{1i} = \sum_{i=1}^{n} y_i x_{1i}$$

$$\beta_1 \left(\sum_{i=1}^{n} x_{1i}^2 - n\bar{x}_1^2 \right) + \beta_2 \left(\sum_{i=1}^{n} x_{1i} x_{2i} - n\bar{x}_1 \bar{x}_2 \right) = \sum_{i=1}^{n} y_i x_{1i} - n\bar{y}\,\bar{x}_1$$

$$\sum_{i=1}^{n} x_{1i}'^2 = \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2$$
$$= \sum_{i=1}^{n} x_{1i}^2 - 2\sum_{i=1}^{n} x_{1i}\bar{x}_1 + n\bar{x}_1^2$$
$$= \sum_{i=1}^{n} x_{1i}^2 - n\bar{x}_1^2$$

Suppose $y_i' = y_i - \bar{y}$, $x_{1i}' = x_{1i} - \bar{x}_1$, $x_{2i}' = x_{2i} - \bar{x}_2$. Then we have

$$\beta_1 \sum_{i=1}^{n} x_{1i}'^2 + \beta_2 \sum_{i=1}^{n} x_{1i}' x_{2i}' = \sum_{i=1}^{n} y_{1i}' x_{1i}' \quad \ldots (11)$$

$$\left(\bar{y} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2\right) \sum_{i=1}^{n} x_{2i} + \beta_1 \sum_{i=1}^{n} x_{1i}\, x_{2i} + \beta_2 \sum_{i=1}^{n} x_{2i}^2 = \sum_{i=1}^{n} y_i x_{2i}$$

$$\beta_1 \left(\sum_{i=1}^{n} x_{1i} x_{2i} - n\bar{x}_1 \bar{x}_2 \right) + \beta_2 \left(\sum_{i=1}^{n} x_{2i}^2 - n\bar{x}_2^2 \right) = \sum_{i=1}^{n} y_i x_{2i} - n\bar{y}\,\bar{x}_2$$

$$\beta_1 \sum_{i=1}^{n} x_{1i}' x_{2i}' + \beta_2 \sum_{i=1}^{n} x_{2i}'^2 = \sum_{i=1}^{n} y_i' x_{2i}' \quad \ldots (12)$$

Note that equations (11) and (12) are simultaneous linear equations. We solve them using the standard method and obtain the least square estimators of β_1 and β_2 as follows:

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^{n} y_i' x_{1i}'\right)\left(\sum_{i=1}^{n} x_{2i}'^2\right) - \left(\sum_{i=1}^{n} y_i' x_{2i}'\right)\left(\sum_{i=1}^{n} x_{1i}' x_{2i}'\right)}{\left(\sum_{i=1}^{n} x_{1i}'^2\right)\left(\sum_{i=1}^{n} x_{2i}'^2\right) - \left(\sum_{i=1}^{n} x_{1i}' x_{2i}'\right)^2} \quad \ldots (13)$$

$$\hat{\beta}_2 = \frac{\left(\sum_{i=1}^{n} y_i' x_{2i}'\right)\left(\sum_{i=1}^{n} x_{1i}'^2\right) - \left(\sum_{i=1}^{n} y_i' x_{1i}'\right)\left(\sum_{i=1}^{n} x_{1i}' x_{2i}'\right)}{\left(\sum_{i=1}^{n} x_{1i}'^2\right)\left(\sum_{i=1}^{n} x_{2i}'^2\right) - \left(\sum_{i=1}^{n} x_{1i}' x_{2i}'\right)^2} \quad \ldots (14)$$

We can rewrite equations (13) and (14) as:

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^{n} y_i x_{i1} - n\bar{y}\,\bar{x}_1\right)\left(\sum_{i=1}^{n} x_{2i}^2 - n\bar{x}_2^2\right) - \left(\sum_{i=1}^{n} y_i x_{2i} - n\bar{y}\,\bar{x}_2\right)\left(\sum_{i=1}^{n} x_{1i}x_{2i} - n\bar{x}_1\bar{x}_2\right)}{\left(\sum_{i=1}^{n} x_{1i}^2 - n\bar{x}_1^2\right)\left(\sum_{i=1}^{n} x_{2i}^2 - n\bar{x}_2^2\right) - \left(\sum_{i=1}^{n} x_{1i}x_{2i} - n\bar{x}_1\bar{x}_2\right)^2} \quad \text{... (15)}$$

$$\hat{\beta}_2 = \frac{\left(\sum_{i=1}^{n} y_i x_{2i} - n\bar{x}_2\bar{y}\right)\left(\sum_{i=1}^{n} x_{1i}^2 - n\bar{x}_1^2\right) - \left(\sum_{i=1}^{n} y_i x_{1i} - n\bar{y}\,\bar{x}_1\right)\left(\sum_{i=1}^{n} x_{1i}x_{2i} - n\bar{x}_1\bar{x}_2\right)}{\left(\sum_{i=1}^{n} x_{1i}^2 - n\bar{x}_1^2\right)\left(\sum_{i=1}^{n} x_{2i}^2 - n\bar{x}_2^2\right) - \left(\sum_{i=1}^{n} x_{1i}x_{2i} - n\bar{x}_1\bar{x}_2\right)^2} \quad \text{... (16)}$$

Substituting the values of $\hat{\beta}_1$ and $\hat{\beta}_2$ from equations (15) and (16) in equation (10), we obtain

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x}_2 \quad \text{... (17)}$$

We define the fitted multiple linear regression model as:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \quad \text{... (18)}$$

Let us solve the following example so that you may understand the fitting of multiple regressor model.

Example 2: Consider the data on systolic blood pressure (SBP), age and weight of 15 women given in Example 1. For this data:

- Fit the multiple linear regression model.
- Also, estimate the value of systolic blood pressure of a woman if her age is 30 years and weight is 65 kg.

Solution: Suppose the multiple regression model is given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

where X_1 (age) and X_2 (weight) are the two regressor variables, and Y (SBP) is a response variable. To estimate the regression coefficients of the multiple regression model, we construct the following table:

Table 2: Computation of $\sum_{i=1}^{15} y_i, \sum_{i=1}^{15} x_{1i}, \sum_{i=1}^{15} x_{2i}, \sum_{i=1}^{15} x_i^2, \sum_{i=1}^{15} x_{2i}^2, \sum_{i=1}^{15} y_i x_{1i}, \sum_{i=1}^{15} y_i x_{2i}$ and $\sum_{i=1}^{15} x_{1i}x_{2i}$

S. No.	y	x_1	x_2	x_1^2	x_2^2	yx_1	yx_2	x_1x_2
1	124	30	71	900	5041	3720	8804	2130
2	134	38	82	1444	6724	5092	10988	3116
3	135	39	98	1521	9604	5265	13230	3822
4	121	26	72	676	5184	3146	8712	1872
5	122	29	70	841	4900	3538	8540	2030
6	119	27	72	729	5184	3213	8568	1944
7	128	32	76	1024	5776	4096	9728	2432
8	118	25	54	625	2916	2950	6372	1350
9	120	26	58	676	3364	3120	6960	1508
10	123	31	68	961	4624	3813	8364	2108

11	129	37	63	1369	3969	4773	8127	2331
12	117	25	62	625	3844	2925	7254	1550
13	131	35	92	1225	8464	4585	12052	3220
14	126	34	75	1156	5625	4284	9450	2550
15	134	40	89	1600	7921	5360	11926	3560
Total	1881	474	1102	15372	83140	59880	139075	35523

From the above table, we have

$$\sum_{i=1}^{15} y_i = 1881, \sum_{i=1}^{15} x_{1i} = 474, \sum_{i=1}^{15} x_{2i} = 1102, \sum_{i=n}^{15} x_{1i}^2 = 15372, \sum_{i=1}^{15} x_{2i}^2 = 83140,$$

$$\sum_{i=1}^{15} y_i x_{1i} = 59880, \sum_{i=1}^{15} y_i x_{2i} = 139075 \text{ and } \sum_{i=1}^{15} x_{1i} x_{2i} = 35523$$

$$\bar{y} = \frac{1881}{15} = 125.4, \bar{x}_1 = \frac{474}{15} = 31.60 \text{ and } \bar{x}_2 = \frac{1102}{15} = 73.4667$$

On substituting the values of $\sum_{i=1}^{15} y_i, \sum_{i=1}^{15} x_{1i}, \sum_{i=1}^{15} x_{2i}, \sum_{i=1}^{15} x_i^2, \sum_{i=1}^{15} x_{2i}^2, \sum_{i=1}^{15} y_i x_{1i}, \sum_{i=1}^{15} y_i x_{2i}$ and $\sum_{i=1}^{15} x_{1i} x_{2i}$ from Table 2 in the normal equations (7) to (9), we obtain

$$15\beta_0 + 474\beta_1 + 1102\beta_2 = 1881 \quad \text{... (i)}$$

$$474\beta_0 + 15372\beta_1 + 35523\beta_2 = 59880 \quad \text{... (ii)}$$

$$1102\beta_0 + 35523\beta_1 + 83140\beta_2 = 139075 \quad \text{... (iii)}$$

From equation (i), we have

$$\beta_0 = \frac{(1881 - 474\beta_1 - 1102\beta_2)}{15} \quad \text{... (iv)}$$

On substituting the value of β_0 in equation (ii) and simplifying, we get

$$474\left(\frac{1881 - 474\beta_1 - 1102\beta_2}{15}\right) + 15372\beta_1 + 35523\beta_2 = 59880$$

$$891594 - 224676\beta_1 - 522348\beta_2 + 230580\beta_1 + 532845\beta_2 = 898200$$

$$\Rightarrow \quad 5904\beta_1 + 10497\beta_2 = 6606 \quad \text{... (v)}$$

On substituting the value of β_0 in equation (iii) and simplifying, we determine

$$1102\left(\frac{1881 - 474\beta_1 - 1102\beta_2}{15}\right) + 35523\beta_1 + 83140\beta_2 = 139075$$

$$2072862 - 522348\beta_1 - 1214404\beta_2 + 532845\beta_1 + 1247100\beta_2 = 2086125$$

$$\Rightarrow \quad 10497\beta_1 + 32696\beta_2 = 13263 \quad \text{... (vi)}$$

We now solve the simultaneous linear equations (v) and (vi). For this we multiply equation (v) with 10497 and equation (vi) with 5904 and get

$$61974288\beta_1 + 110187009\beta_2 = 69343182 \quad \text{... (vii)}$$

$$-61974288\beta_1 \pm 193037184\beta_2 = \pm 78304752 \quad \text{... (viii)}$$

On solving equations (vii) and (viii), we get

$$-82850175\beta_2 = -8961570$$

or $$\beta_2 = \frac{-8961570}{-82850175}$$

$$\therefore \quad \beta_2 = 0.1082$$

On substituting the value of β_2 in equation (v), we get

$$5904\beta_1 + 10497 \times 0.1082 = 6606$$

or $$\beta_1 = \frac{5470.5818}{5904}$$

$$\therefore \quad \beta_1 = 0.9266$$

To determine the value of β_0, we substitute the values of β_1 and β_2 in equation (iv) and get

$$\beta_0 = \frac{(1881 - 474 \times 0.9266 - 1102 \times 0.1082)}{15}$$

or $$\beta_0 = \frac{1322.5979}{15} = 88.1732$$

Thus, the fitted model would be

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\hat{Y} = 88.1732 + 0.9266\ X_1 + 0.1082\ X_2$$

The predicted value of SBP of a woman if her age is 30 years and weight is 65 kg is determined to be

$$\hat{Y} = 88.1732 + 0.9266 \times 30 + 0.1082 \times 65$$

$$= 123.0017$$

7.3.2 Multiple Regression Model with k Regressor Variables

The multiple linear regression model defined in equation (3) for k regressor variables is given as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon \quad \text{... (19)}$$

Since the parameters $\beta_0, \beta_1, ..., \beta_k$ are unknown, we estimate the values of these coefficients using the method of least squares. In Sec. 7.3.1, you have learnt this method to estimate the regression parameters β_0, β_1 and β_2 in case of multiple linear regression model for two regressor variables. For n pairs of data, we can define the multiple regression model for the i^{th} observation, i.e., $(y_i, x_{1i}, x_{2i}, ..., x_{ki})$ as:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i \quad ; i = 1, 2, ..., n \qquad ... (20)$$

The sum of squares of the error terms is given by

$$E = \sum_{i=1}^{n} \varepsilon_i^2 = \left(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - ... - \beta_k x_{ki} \right)^2 \qquad \ldots (21)$$

We now differentiate the error sum of squares (E) with respect to the parameters, $\beta_0, \beta_1, ..., \beta_k$, and equate the derivatives to zero. In this way, we obtain (k + 1) normal equations each corresponding to unknown parameters $\beta_0, \beta_1, ..., \beta_k$, respectively, as shown below:

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_{1i} + \beta_2 \sum_{i=1}^{n} x_{2i} + ... + \beta_k \sum_{i=1}^{n} x_{ki} = \sum_{i=1}^{n} y_i$$

$$\beta_0 \sum_{i=1}^{n} x_{1i} + \beta_1 \sum_{i=1}^{n} x_{1i}^2 + \beta_2 \sum_{i=1}^{n} x_{2i} x_{1i} + ... + \beta_k \sum_{i=1}^{n} x_{ki} x_{1i} = \sum_{i=1}^{n} y_i x_{1i}$$

$$\beta_0 \sum_{i=1}^{n} x_{2i} + \beta_1 \sum_{i=1}^{n} x_{1i} x_{2i} + \beta_2 \sum_{i=1}^{n} x_{2i}^2 + ... + \beta_k \sum_{i=1}^{n} x_{ki} x_{2i} = \sum_{i=1}^{n} y_i x_{2i} \qquad ... (22)$$

……

……

……

$$\beta_0 \sum_{i=1}^{n} x_{ki} + \beta_1 \sum_{i=1}^{n} x_{1i} x_{ki} + \beta_2 \sum_{i=1}^{n} x_{2i} x_{ki} + ... + \beta_k \sum_{i=1}^{n} x_{ki}^2 = \sum_{i=1}^{n} y_i x_{ki}$$

We solve these (k + 1) normal equations by the methods used for solving the simultaneous linear equations when k < n. In this way, we obtain the least squares estimators of $\beta_0, \beta_1, ..., \beta_k$ denoted by $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$, respectively.

If the numbers of regression coefficients are large, it will not be so easy to solve these equations manually. In this situation, we use the matrix approach for estimating the multiple regression coefficients as described in Block 3 of MST-001. We can also use some computer programs or software to solve these equations easily.

We illustrate the matrix approach for k regressor variables in the next section.

You may like to solve the following exercises before studying further.

E3) For the data given in **E2**, fit a multiple regression model for the dependence of random blood sugar level (mg/dL) on the average duration of walk (minutes) and serum creatinine (mg/dL) of 15 diabetic patients. Also, estimate the value of blood sugar of a patient who walks for 30 minutes daily and has 0.5 mg/dL serum creatinine.

E4) The following table provides the information of birth weight of neonate, gestational age of fetus and increase in mother's weight during pregnancy:

S. No.	Birth Weight (Y)	Gestational Age (X_1)	Increase in Mother's Weight (X_2)
1	2.4	34.0	7
2	2.3	34.7	5
3	2.0	29.8	7
4	2.9	38.2	9

5	2.7	36.1	10.5
6	3.2	42.8	14
7	3.4	40.8	10
8	2.8	37.8	8
9	3.2	38.4	13.5
10	3.7	41.3	13
11	4.0	42.0	11
12	3.4	40.1	12

Fit the multiple regression model to check the relationship of birth weight with the other variables given in above data.

7.4 ESTIMATION OF REGRESSION COEFFICIENTS USING THE MATRIX APPROACH

It is difficult to follow the procedure described in Sec. 7.3 when we have more than two regressor variables in the model. In such a situation, it is more convenient to express all related computations of the multiple regression model in matrix form. We can express the multiple regression model given in equation (20) in matrix form as:

$$\underset{\sim}{Y} = \mathbf{X}\ \underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \qquad \ldots (23)$$

where, $\underset{\sim}{Y}$ is an $(n \times 1)$ vector of n observations of the dependent variable,

X is an (n×(k+1)) matrix of n observations of the unit vector and each of the k response variables, $\underset{\sim}{\beta}$ is a ((k+1) × 1) vector of the regression coefficients, and $\underset{\sim}{\varepsilon}$ is an (n×1) vector of the error terms, such that

$$\underset{\sim}{Y}_{n\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \underset{\sim}{1} & \underset{\sim}{X}_1 & \underset{\sim}{X}_2 & \ldots & \underset{\sim}{X}_k \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \ldots & x_{k1} \\ 1 & x_{12} & x_{22} & \ldots & x_{k2} \\ 1 & x_{13} & x_{23} & \ldots & x_{k3} \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & x_{1n} & x_{2n} & \ldots & x_{kn} \end{bmatrix},$$

$$\underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_k \end{bmatrix} \text{ and } \underset{\sim}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

We can also express (k + 1) normal equations given in equation (22) in matrix form. But before expressing the normal equations in matrix form, we shall explain how to determine the values of $X'X$ and $X'\underset{\sim}{Y}$:

$$X'X = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{k1} & x_{k2} & x_{k3} & \dots & x_{kn} \end{bmatrix} \times \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ 1 & x_{13} & x_{23} & \dots & x_{k3} \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix}$$

$$X'X = \begin{bmatrix} n & \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{2i} & \dots & \sum_{i=1}^{n} x_{ki} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & \dots & \sum_{i=1}^{n} x_{1i}x_{ki} \\ \sum_{i=1}^{n} x_{2i} & \sum_{i=1}^{n} x_{2i}x_{1i} & \sum x_{2i}^2 & \dots & \sum x_{2i}x_{ki} \\ . & . & . & . & . \\ . & . & . & . & . \\ \sum_{i=1}^{n} x_{ki} & \sum_{i=1}^{n} x_{ki}x_{1i} & \sum_{i=1}^{n} x_{ki}x_{2i} & \dots & \sum_{i=1}^{n} x_{ki}^2 \end{bmatrix} \quad \dots (24)$$

$$\text{and } X'\underset{\sim}{Y} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{k1} & x_{k2} & x_{k3} & \dots & x_{kn} \end{bmatrix} \times \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i x_{1i} \\ \sum_{i=1}^{n} y_i x_{2i} \\ . \\ . \\ \sum_{i=1}^{n} y_i x_{ki} \end{bmatrix} \quad \dots (25)$$

You can see that the (k + 1) normal equations given in equation (22) can be expressed in matrix form as:

$$= \begin{bmatrix} n & \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{2i} & \dots & \sum_{i=1}^{n} x_{ki} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & \dots & \sum_{i=1}^{n} x_{1i}x_{ki} \\ \sum_{i=1}^{n} x_{2i} & \sum_{i=1}^{n} x_{2i}x_{1i} & \sum x_{2i}^2 & \dots & \sum x_{2i}x_{ki} \\ . & . & . & \dots & . \\ . & . & . & \dots & . \\ \sum_{i=1}^{n} x_{ki} & \sum_{i=1}^{n} x_{ki}x_{1i} & \sum_{i=1}^{n} x_{ki}x_{2i} & \dots & \sum_{i=1}^{n} x_{ki}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i x_{1i} \\ \sum_{i=1}^{n} y_i x_{2i} \\ . \\ . \\ \sum_{i=1}^{n} y_i x_{ki} \end{bmatrix} \quad \dots (26)$$

$$\text{i.e., } X'X\,\underset{\sim}{\beta} = X'\,\underset{\sim}{Y} \quad \dots (27)$$

We multiply both sides of equation (27) by the inverse of $X'X$, i.e., $(X'X)^{-1}$ to solve the normal equations as follows:

$$(X'X)^{-1} X'X \underset{\sim}{\beta} = (X'X)^{-1} X' \underset{\sim}{Y}$$

$$\Rightarrow \quad I\underset{\sim}{\beta} = (X'X)^{-1} X' \underset{\sim}{Y}, \text{ Since } (X'X)^{-1} X'X = I_{k+1}$$

where I_{k+1} is an identity matrix of order $((k+1)\times(k+1))$.

$$\Rightarrow \quad \underset{\sim}{\hat{\beta}} = (X'X)^{-1} X' \underset{\sim}{Y} \qquad \ldots (28)$$

where $\underset{\sim}{\hat{\beta}} = \begin{bmatrix} \hat{\beta}_0 & \hat{\beta}_1 & \hat{\beta}_2 & \ldots & \hat{\beta}_k \end{bmatrix}$

Thus, we can obtain the least squares estimates $(\underset{\sim}{\hat{\beta}})$ of the regression coefficients $(\underset{\sim}{\beta})$ from equation (28). Substituting the values of the estimated regression coefficients obtained from equation (28) in equation (23), we obtain the fitted multiple linear regression model as:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k \qquad \ldots (29)$$

We can then determine the predicted value of Y for the i^{th} observation as:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_k x_{ki} \qquad ; i = 1, 2, \ldots, n$$

We write the fitted multiple linear regression model in matrix form as:

$$\underset{\sim}{\hat{Y}} = \mathbf{X}\underset{\sim}{\hat{\beta}} \qquad \ldots (30)$$

where $\underset{\sim}{\hat{Y}} = \begin{bmatrix} \hat{y}_1 & \hat{y}_2 & \ldots & \hat{y}_n \end{bmatrix}$ is an $(n\times 1)$ vector of n predicted values of the response variable (Y).

Let us solve a couple of examples so that you become familiar with the matrix approach for two and three regressor variables.

Example 3: Consider the data of Example 1 given in Sec. 7.2 to fit a multiple linear regression model with two regressor variables using the matrix approach. Also estimate the predicted value of systolic blood pressure at $X_1 = 28$ years and $X_2 = 60$ kg.

Solution: From equations (24) and (25), we have

$$X'X = \begin{bmatrix} n & \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{2i} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} \\ \sum_{i=1}^{n} x_{2i} & \sum_{i=1}^{n} x_{2i}x_{1i} & \sum_{i=1}^{n} x_{2i}^2 \end{bmatrix}, \; X'\underset{\sim}{Y} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i x_{1i} \\ \sum_{i=1}^{n} y_i x_{2i} \end{bmatrix}$$

From Table 2 constructed in Example 2, we obtain $X'X$ and $X'\underset{\sim}{Y}$ as follows:

$$X'X = \begin{bmatrix} 15 & 474 & 1102 \\ 474 & 15372 & 35523 \\ 1102 & 35523 & 83140 \end{bmatrix} \text{ and } X'\underset{\sim}{Y} = \begin{bmatrix} 1881 \\ 59880 \\ 139075 \end{bmatrix}$$

Thus, we determine the regression coefficients from

$$\underset{\sim}{\hat{\beta}} = (X'X)^{-1} X'\underset{\sim}{Y}$$

$$\underset{\sim}{\hat{\beta}} = \begin{bmatrix} 15 & 474 & 1102 \\ 474 & 15372 & 35523 \\ 1102 & 35523 & 83140 \end{bmatrix}^{-1} \begin{bmatrix} 1881 \\ 59880 \\ 139075 \end{bmatrix}$$

You have learnt how to determine the inverse of a matrix in Unit 10 of "MST-01: Foundation in Mathematics and Statistics", using adjoints. To compute $(X'X)^{-1}$, we need to obtain the inverse of the matrix $(X'X)$. So, we first determine the determinant of $(X'X)$, i.e., $|X'X|$ and then the adjoint matrix of $X'X$, i.e., adj $(X'X)$ as explained below:

$$|X'X| = 15\begin{vmatrix} 15372 & 35523 \\ 35523 & 83140 \end{vmatrix} + 470\begin{vmatrix} 474 & 35523 \\ 1102 & 83140 \end{vmatrix} + 1102\begin{vmatrix} 474 & 15372 \\ 1102 & 25523 \end{vmatrix}$$

$$= 15\left[15372 \times 83140 - (35523)^2\right] - 474\left[474 \times 83140 - 1102 \times 35523\right]$$

$$+ 1102\left[474 \times 35523 - 1102 \times 35523\right]$$

$$= 15 \times 16144551 - 474 \times 124194636(-102042) + 1102 \times 112450284$$

$$= 242168265 - 124194636 - 112450284 = 5523345$$

We now compute the co-factors A_{ij} for determining the adjoint matrix of $(X'X)$

$$A_{11} = +\begin{vmatrix} 15372 & 35523 \\ 35523 & 83140 \end{vmatrix}$$

$$= \left[15372 \times 83140 - (35523)^2\right]$$

$$= 16144551$$

$$A_{12} = -\begin{vmatrix} 474 & 35523 \\ 1102 & 83140 \end{vmatrix}$$

$$= -\left[474 \times 83140 - 1102 \times 35523\right]$$

$$= -262014$$

$$A_{13} = +\begin{vmatrix} 474 & 15372 \\ 1102 & 35523 \end{vmatrix}$$

$$= [474 \times 35523 - 1102 \times 15372]$$

$$= -102042$$

$$A_{21} = +\begin{vmatrix} 474 & 1102 \\ 35523 & 83140 \end{vmatrix}$$

$$= -\left[474 \times 83140 - 35523 \times 1102\right]$$

$$= -262014$$

$$A_{22} = +\begin{vmatrix} 15 & 1102 \\ 1102 & 83140 \end{vmatrix}$$

$$= [15 \times 83140 - (1102)^2]$$

$$= 32696$$

$$A_{23} = -\begin{vmatrix} 15 & 474 \\ 1102 & 35523 \end{vmatrix}$$

$$= [15 \times 35523 - 1102 \times 474]$$

$$= -10497$$

$$A_{31} = +\begin{vmatrix} 474 & 15372 \\ 1102 & 35523 \end{vmatrix}$$

$$= [474 \times 35523 - 1102 \times 15372]$$

$$= -102042$$

$$A_{32} = -\begin{vmatrix} 15 & 474 \\ 1102 & 35523 \end{vmatrix}$$

$$= [15 \times 35523 - 1102 \times 474]$$

$$= -10497$$

$$A_{33} = +\begin{vmatrix} 15 & 474 \\ 474 & 15372 \end{vmatrix}$$

$$= [15 \times 15372 - (474)^2$$

$$= 59$$

Now, we have

$$\text{adj}(X'X) = \begin{bmatrix} A_{11} & A_{21} & A_{31} \\ A_{12} & A_{22} & A_{32} \\ A_{13} & A_{23} & A_{33} \end{bmatrix} = \begin{bmatrix} 16144551 & -262014 & -102042 \\ -262014 & 32696 & -10497 \\ -102042 & -10497 & 5904 \end{bmatrix}$$

Thus,

$$(X'X)^{-1} = \frac{\text{adj}(X'X)}{|X'X|}$$

$$(X'X)^{-1} = \frac{1}{5523345}\begin{bmatrix} 16144551 & -262014 & -102042 \\ -262014 & 32696 & -10497 \\ -102042 & -10497 & 5904 \end{bmatrix}$$

$$= \begin{bmatrix} 2.9230 & -0.0474 & -0.0185 \\ -0.0474 & 0.0059 & -0.0019 \\ -0.0185 & -0.0019 & 0.0011 \end{bmatrix}$$

Therefore, from equation (28), we have

$$\underset{\sim}{\hat{\beta}} = \begin{bmatrix} 2.9230 & -0.0474 & -0.0185 \\ -0.0474 & 0.0059 & -0.0019 \\ -0.0185 & -0.0019 & 0.0011 \end{bmatrix} \times \begin{bmatrix} 1881 \\ 59880 \\ 139075 \end{bmatrix}$$

$$= \begin{bmatrix} 88.1732 \\ 0.9266 \\ 0.1082 \end{bmatrix}$$

It is to be noted that all calculations were performed up to 15 fixed decimal places for showing accurate results in this block. For the sake of simplicity, we are showing results up to 4 decimal places only. The results may vary if we carry out the calculations by fixing values at various decimal places.

Thus, $\beta_0 = 88.1732$, $\beta_1 = 0.9266$ and $\beta_2 = 0.1082$

Since the values in the matrix are very small fractional values, it would be better if these are reported for as many decimal places as feasible.

The value of $\hat{\beta}_1 = 0.9266$ indicates that the expected increase or decrease will be approximately 0.93 mm/Hg in SBP due to per year increase or decrease in age (X_1) when weight (X_2) is considered as constant. Similarly, one kg increase or decrease in weight will result in $(\hat{\beta}_2 = 0.1082) \cong 0.11 \text{mm/Hg}$ increase or decrease in SBP when age is held constant on the basis of the given data.

Thus, the fitted multiple regression model will be

$$\hat{Y} = 88.1732 + 0.9266\ X_1 + 0.1082\ X_2$$

For $X_1 = 28$ and $X_2 = 60$, the predicted value of SBP is computed as:

$$\hat{Y} = 88.1732 + 0.9266 \times 28 + 0.1082 \times 60$$

$$= 120.6076$$

Now, we consider an example of three regressor variables.

Example 4: Consider the following data to fit the multiple regression model for examining the relationship of systolic blood pressure with three regressor variables, age, weight and height:

S. No.	SBP (mm/Hg)	Age (year)	Weight (kg)	Height (cm)
1	124	30	71	150
2	134	38	82	176
3	135	39	98	165
4	121	26	72	151
5	122	29	70	157
6	119	27	72	154
7	128	32	76	160
8	118	25	54	160
9	120	26	58	158
10	123	31	68	153
11	129	37	63	159
12	117	25	62	158
13	131	35	92	165
14	126	34	75	155
15	134	40	89	161

Solution: We determine the values required for the computation of the regression coefficients as shown in the following table:

Table 3: Computation of $\sum_{j=1}^{15} y_j, \sum_{j=1}^{15} x_{ij}, \sum_{j=1}^{15} x_{ij}^2, \sum_{j=1}^{15} y_j x_{ij}$ and $\sum_{k=1}^{15} x_{ik} x_{jk}; (i, j = 1, 2, 3)$

S. No.	Y	X_1	X_2	X_3	X_1^2	X_2^2	X_3^2	YX_1	YX_2	YX_3	X_1X_2	X_2X_3	X_1X_3
1	124	30	71	150	900	5041	22500	3720	8804	18600	2130	10650	4500
2	134	38	82	176	1444	6724	30976	5092	10988	23584	3116	14432	6688
3	135	39	98	165	1521	9604	27225	5265	13230	22275	3822	16170	6435
4	121	26	72	151	676	5184	22801	3146	8712	18271	1872	10872	3926
5	122	29	70	157	841	4900	24649	3538	8540	19154	2030	10990	4553
6	119	27	72	154	729	5184	23716	3213	8568	18326	1944	11088	4158
7	128	32	76	160	1024	5776	25600	4096	9728	20480	2432	12160	5120
8	118	25	54	160	625	2916	25600	2950	6372	18880	1350	8640	4000
9	120	26	58	158	676	3364	24964	3120	6960	18960	1508	9164	4108
10	123	31	68	153	961	4624	23409	3813	8364	18819	2108	10404	4743
11	129	37	63	159	1369	3969	25281	4773	8127	20511	2331	10017	5883
12	117	25	62	158	625	3844	24964	2925	7254	18486	1550	9796	3950
13	131	35	92	165	1225	8464	27225	4585	12052	21615	3220	15180	5775
14	126	34	75	155	1156	5625	24025	4284	9450	19530	2550	11625	5270
15	134	40	89	161	1600	7921	25921	5360	11926	21574	3560	14329	6440
Total	1881	474	1102	2382	15372	83140	378856	59880	139075	299065	35523	175517	75549

Using equations (24) and (25), we write the matrices $X'X$ and $X'Y$ as:

$$X'X = \begin{bmatrix} 15 & 474 & 1102 & 2382 \\ 474 & 15372 & 35523 & 75549 \\ 1102 & 35523 & 83140 & 175517 \\ 2382 & 75549 & 175517 & 378856 \end{bmatrix} \text{ and } X'\underset{\sim}{Y} = \begin{bmatrix} 1881 \\ 59880 \\ 139075 \\ 299065 \end{bmatrix}$$

Next, we obtain the determinant of $(X'X)$, i.e., $|X'X|$ and then the adjoint matrix of $X'X$, i.e., $\text{adj}(X'X)$ as we did in Example 3.

$$|X'X| = \begin{vmatrix} 15 & 474 & 1102 & 2382 \\ 474 & 15372 & 35523 & 75549 \\ 1102 & 35523 & 83140 & 175517 \\ 2382 & 75549 & 175517 & 378856 \end{vmatrix} = 2196283446$$

$$\text{adj}(X'X) = \begin{bmatrix} A_{11} & A_{21} & A_{31} & A_{41} \\ A_{12} & A_{22} & A_{32} & A_{42} \\ A_{13} & A_{23} & A_{33} & A_{43} \\ A_{14} & A_{24} & A_{34} & A_{44} \end{bmatrix}$$

Therefore, we can compute the inverse matrix of $X'X$ as

$$(X'X)^{-1} = \frac{\text{adj}(X'X)}{|X'X|}$$

$$(X'X)^{-1} = \frac{1}{2196283446} \begin{bmatrix} 108618667326 & 389698194 & -20107638 & -751319082 \\ 389698194 & 15387857 & -4075077 & 3630807 \\ -20107638 & -4075077 & 2351745 & -150471 \\ -751319082 & 3630807 & -150471 & 5523345 \end{bmatrix}$$

$$= \begin{bmatrix} 49.4557 & 0.1774 & -0.0092 & -0.3421 \\ 0.1774 & 0.0070 & -0.0019 & 0.0017 \\ -0.0092 & -0.0019 & 0.0011 & -0.0001 \\ -0.3421 & 0.0017 & -0.0001 & 0.0025 \end{bmatrix}$$

Note that here we have performed all computations using the matrix approach up to fifteen decimal places of accuracy. The results so obtained may vary with the results obtained from varying decimal places of accuracy.

Thus, we obtain the regression coefficients as:

$$\underset{\sim}{\hat{\beta}} = (X'X)^{-1} X'\underset{\sim}{Y}$$

$$\underset{\sim}{\hat{\beta}} = \begin{bmatrix} 15 & 474 & 1102 & 2382 \\ 474 & 15372 & 35523 & 75549 \\ 1102 & 35523 & 83140 & 175517 \\ 2382 & 75549 & 175517 & 378856 \end{bmatrix}^{-1} \begin{bmatrix} 1881 \\ 59880 \\ 139075 \\ 299065 \end{bmatrix}$$

$$\underset{\sim}{\hat{\beta}} = \begin{bmatrix} 49.4557 & 0.1774 & -0.0092 & -0.3421 \\ 0.1774 & 0.0070 & -0.0019 & 0.0017 \\ -0.0092 & -0.0019 & 0.0011 & -0.0001 \\ -0.3421 & 0.0017 & -0.0001 & 0.0025 \end{bmatrix} \times \begin{bmatrix} 1881 \\ 59880 \\ 139075 \\ 299065 \end{bmatrix}$$

$$= \begin{bmatrix} 71.5436 \\ 0.8462 \\ 0.1048 \\ 0.1223 \end{bmatrix}$$

Thus, we have

$$\beta_0 = 71.5436, \quad \beta_1 = 0.8462, \beta_2 = 0.1048 \text{ and } \beta_3 = 0.1223$$

The fitted multiple regression model can be written as:

$$\hat{Y} = 71.5436 + 0.8462\ X_1 + 0.1048\ X_2 + 0.1223\ X_3$$

You should now solve the following exercises to practice the matrix approach.

E5) To study the impact of duration of daily walk and increase in the serum creatinine on the sugar level of diabetic patients, consider the data given in **E2**. Fit a multiple regression model using the matrix approach and compare the results with the model obtained in **E2**.

E6) Fit a multiple regression model on the following data to check the relationship of birth weight of neonate on three variables: gestational age of fetus, increase in mother's weight during pregnancy and mother's height:

S. No.	Birth Weight (Y)	Gestational Age (X_1)	Increase in Mother's Weight (X_2)	Mother's Height (X_3)
1	2.4	34.0	7	152
2	2.3	34.7	5	174
3	2.0	29.8	7	165
4	2.9	38.2	9	151
5	2.7	36.1	10.5	157
6	3.2	42.8	14	154
7	3.4	40.8	10	160
8	2.8	37.8	8	160
9	3.2	38.4	13.5	158
10	3.7	41.3	13	153
11	4.0	42.0	11	159
12	3.4	40.1	12	165

7.5 RESIDUAL ANALYSIS

In Sec. 7.4, we have obtained the fitted value of the response variable (Y) for the i^{th} observation as:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_k x_{ki}; \qquad i = 1, 2, ..., n \qquad ... (31)$$

The residuals possibly explain the irregularity of the fitted model that might have occurred and misled us.

In matrix form, we can define the vector of predicted values as:

$$\underset{\sim}{\hat{Y}} = X\underset{\sim}{\hat{\beta}} \qquad \ldots (32)$$

where $\underset{\sim}{\hat{Y}} = [\hat{y}_1, \hat{y}_2, ..., \hat{y}_n]$ is an $(n \times 1)$ vector of n predicted values of the response variable (Y).

You have learnt in Sec. 6.2 of Unit 6 that the residual is a difference between the observed value and the corresponding fitted value of the response variable. Therefore, the i^{th} residual is defined as:

$$r_i = y_i - \hat{y}_i \qquad ; i = 1, 2, ..., n \qquad ... (33)$$

In matrix form, we can write

$$\underset{\sim}{R} = \underset{\sim}{Y} - \underset{\sim}{\hat{Y}} \qquad ... (34)$$

where $\underset{\sim}{R} = [r_1, r_2, ..., r_n]$ is a vector of n residuals.

We define the properties of a fitted multiple regression model based on the residuals first as we did for simple linear regression model discussed in Sec. 5.3.1.

Ideally according to the property of residuals, the sum of the residuals for all given observations should always be equal to zero, i.e., $\sum_{i=1}^{n} r_i = 0$. So we can say that r_i is the i^{th} residual which has mean zero and variance σ^2 by the properties of residuals.

As the value of σ^2 is unknown, we estimate it from the given data as follows:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(r_i - \bar{r})^2}{n-k-1} = \frac{\sum_{i=1}^{n} r_i^2}{n-k-1} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1} \quad \ldots (35)$$

where $\bar{r} = \frac{\sum_{i=1}^{n} r_i}{n} = 0$, and

k is the number of regressor variables in the model.

We now compute the i^{th} standardised residual as:

$$s_i = \frac{r_i}{\hat{\sigma}} \qquad ; i = 1, 2, \ldots, n \quad \ldots (36)$$

The residuals plot can be used to detect non-linearity and/or unequal variance. A normal plot of the residuals can be used to detect non-normality.

The standardised residuals have mean zero and variance approximately equal to one. We use the residual and normal probability plots, which are helpful to check the validity of some underlying assumptions as explained in Sec. 6.2 of Unit 6.

We now give two examples to explain how to compute fitted values, residuals, standardised residuals and construct residual and normal probability plots for two and three regressor variables:

Example 5: For the data on systolic blood pressure, age and weight given in Example 2,

i) obtain the residuals and verify that $\sum_{i=1}^{15} r_i = 0$,

ii) determine the standardised residuals, and

iii) construct the residual and normal probability plots.

Solution: From Example 2, the best fitted multiple regression model is given by

$$\hat{Y} = 88.1732 + 0.9266\ X_1 + 0.1082\ X_2$$

We now determine all predicted values of Y, i.e., $\hat{Y}$ by substituting the given values of regressor variables X_1 and X_2 in the fitted multiple regression model.

We calculate values of the residuals using equation (33), squares of the residuals (r_i^2), standardised residuals using equation (36), ordered standardised residuals, percentile cumulative probabilities for all observations given in the data and arrange them in Table 4 as follows:

Table 4: Computation of residuals and standardised residuals

S. No.	Y_i	Predicted Values ($\hat{y}_i$)	Residuals (r_i)	r_i^2	Standardised Residuals $s_i = \frac{r_i}{\hat{\sigma}}$	Ordered Standardised Residuals $s_{(i)}$	Percentiles Cumulative Probabilities (P_i)
1	124	123.6506	0.3494	0.1220	0.2587	– 1.4653	3.3333
2	134	132.2532	1.7468	3.0514	1.2933	– 1.3251	10.0000
3	135	134.9104	0.0896	0.0080	0.0663	– 0.9275	16.6667
4	121	120.0525	0.9475	0.8978	0.7016	– 0.7731	23.3333
5	122	122.6159	– 0.6159	0.3793	– 0.4560	– 0.6394	30.0000
6	119	120.9790	– 1.9790	3.9166	– 1.4653	– 0.4560	36.6667
7	128	126.0447	1.9553	3.8234	1.4477	– 0.2010	43.3333
8	118	117.1789	0.8211	0.6742	0.6080	0.0663	50.0000
9	120	118.5381	1.4619	2.1371	1.0824	0.2587	56.6667
10	123	124.2527	– 1.2527	1.5694	– 0.9275	0.3294	63.3333
11	129	129.2714	– 0.2714	0.0737	– 0.2010	0.6080	70.0000
12	117	118.0442	– 1.0442	1.0904	– 0.7731	0.7016	76.6667
13	131	130.5551	0.4449	0.1980	0.3294	1.0824	83.3333
14	126	127.7897	– 1.7897	3.2029	– 1.3251	1.2933	90.0000
15	134	134.8635	– 0.8635	0.7457	– 0.6394	1.4477	96.6667
Total	1881	1881	0	21.8898	0		

You can see from Table 4 that $\sum_{i=1}^{15} r_i \cong 0$. This verifies the property of the residuals that the sum of the residuals is zero.

We have n – k–1 = 15 – 2–1 = 12 and $\sum_{i=1}^{15} r_i^2 = 21.8898$

From equation (35), the variance of the residuals is estimated as:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{15} r_i^2}{12} = \frac{21.8898}{12} = 1.8242$$

$$\text{or, } \hat{\sigma} = \sqrt{1.8242} = 1.3506$$

To obtain a residual plot, we consider the predicted Y values and standardised residuals on the horizontal and the vertical axes, respectively, computed in Table 1. In this way, we obtain the residual plot shown in Fig. 8.1.

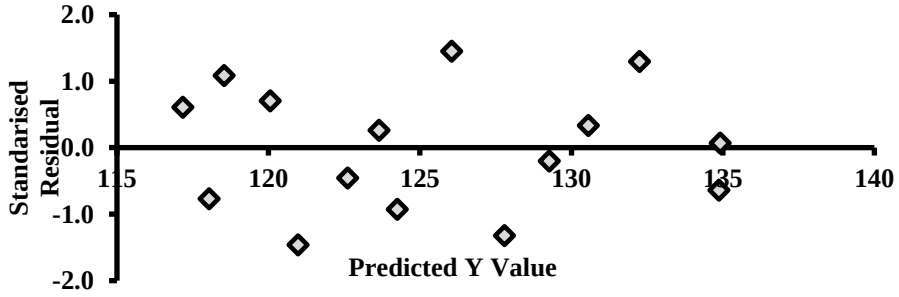

Fig: 7.2: Residual plot.

The standardised residuals shown in Fig. 7.2 appear to be approximately randomly scattered throughout a horizontal band around 0.0 on Y-axis. Hence, the assumption of linear regression is valid or we can say that the multiple linear regression model fits well for the given data.

Then we plot the ordered standardised residuals against the percentiles. We obtain a normal probability plot as shown in Fig. 7.3.

It is to be noted that we can also assign rank to the residuals, without arranging them in increasing order.

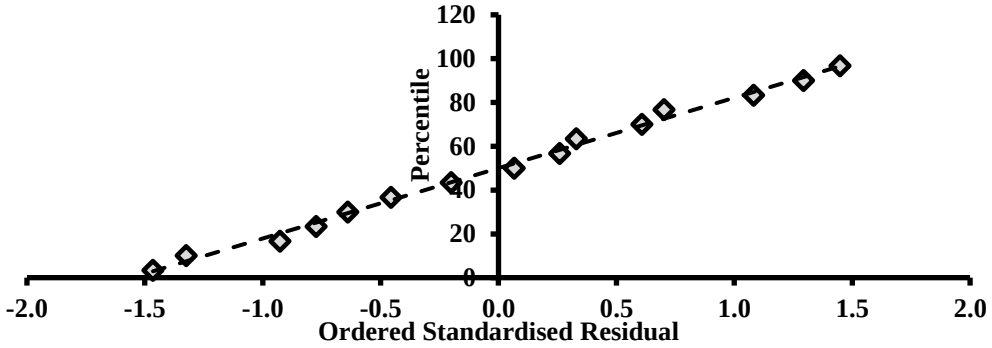

Fig. 7.3: Normal probability plot.

Note that the resulting points lie approximately on a straight line as shown in Fig. 7.3. Notice that some points of the distribution deviate slightly from the straight line, but do not lie very far from the central points. It indicates that the distribution of error terms is approximately normally distributed.

Example 6: For the data on systolic blood pressure, age, weight and height given in Example 3,

i) obtain the residuals and verify that $\sum_{i=1}^{15} r_i = 0$,

ii) determine the standardised residuals, and

iii) construct the residual and normal probability plots.

Solution: From Example 2, the best fitted regression model is given by

$$\hat{Y} = 71.5436 + 0.8462\ X_1 + 0.1048\ X_2 + 0.1223\ X_3$$

We now determine the predicted values of Y, the values of residuals (r_i), squares of the residual (r_1^2) and standardised residuals for all observations given in the data and arrange them in Table 5 as follows:

Table 5: Computation of residuals and standardised residuals.

S. No.	Y_i	Predicted Values $(\hat{y}_i)$	Residuals (r_i)	r_i^2	Standardised Residuals $s_i = \frac{r_i}{\hat{\sigma}}$	Ordered Standardised Residuals $s_{(i)}$	Percentiles Cumulative Probabilities (P_i)
1	124	122.7116	1.2884	1.6599	1.0700	– 1.4674	3.3333
2	134	133.8132	0.1868	0.0349	0.1552	– 1.2583	10.0000
3	135	134.9920	0.0080	0.0001	0.0066	– 0.9361	16.6667
4	121	119.5538	1.4462	2.0915	1.2011	– 0.5119	23.3333
5	122	122.6163	– 0.6163	0.3799	– 0.5119	– 0.5067	30.0000
6	119	120.7668	– 1.7668	3.1216	– 1.4674	– 0.3369	36.6667
7	128	126.1508	1.8492	3.4196	1.5358	0.0066	43.3333
8	118	117.9208	0.0792	0.0063	0.0658	0.0182	50.0000
9	120	118.9419	1.0581	1.1196	0.8788	0.0658	56.6667
10	123	123.6101	– 0.6101	0.3722	– 0.5067	0.0857	63.3333
11	129	128.8968	0.1032	0.0107	0.0857	0.1552	70.0000
12	117	118.5150	– 1.5150	2.2952	– 1.2583	0.8788	76.6667
13	131	130.9781	0.0219	0.0005	0.0182	1.0700	83.3333
14	126	127.1271	– 1.1271	1.2704	– 0.9361	1.2011	90.0000
15	134	134.4057	– 0.4057	0.1646	– 0.3369	1.5358	96.6667
Total	1881	1881	0	15.9469	0		

From Table 3 we observe that $\sum_{i=1}^{15} r_i \cong 0$, which is as expected.

We also have n – k–1 = 15 – 3 –1 = 11 and $\sum_{i=1}^{15} r_i^2 = 15.9469$

From equation (35), the variance of the residuals can also be estimated as:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{15} r_i^2}{11} = \frac{15.9469}{11} = 1.4497$$

$$\text{or } \hat{\sigma} = \sqrt{1.4497} = 1.2040$$

To obtain a residual plot, we take the predicted Y values and standardised residuals on the horizontal and the vertical axes, respectively. The resulting residual plot is shown in Fig. 7.4.

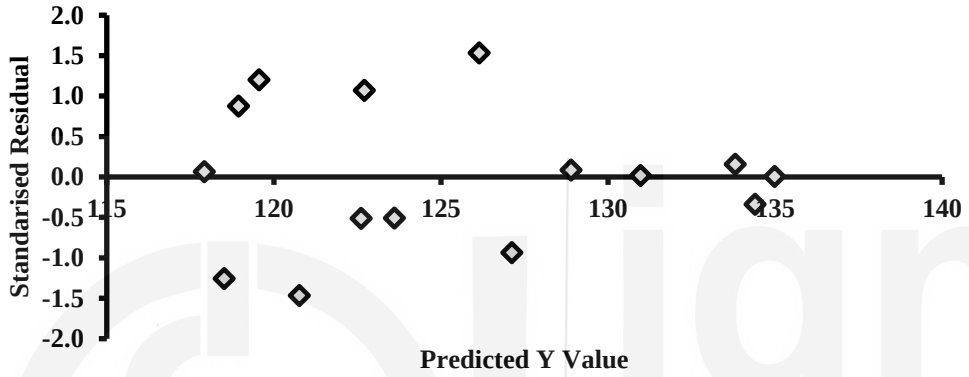

Fig: 7.4: Residual plot.

The residual plot shown in Fig. 7.4 shows an approximate inward opening funnel pattern. It indicates that variance of the errors is not constant but the error variance is approximately decreasing with the response variables.

Then we plot the ordered standardised residuals against the percentiles and obtain a normal probability plot as shown in Fig. 7.5.

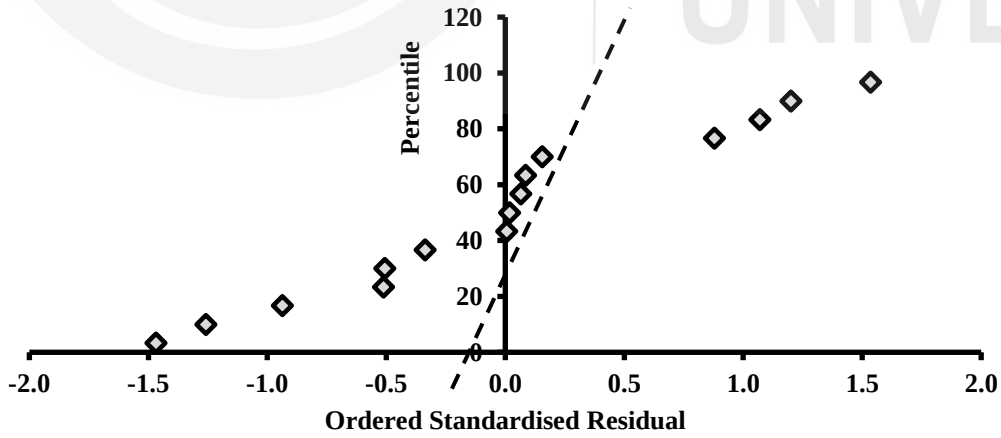

Fig: 7.5: Normal probability plot.

Note that the resulting points do not lie along a straight line. This indicates that there is some problem with the normality assumption. An upward and downward deviation at both ends of the curve indicates deviation from normal curve. Therefore, the plot in Fig. 7.5 cannot be considered as normal.

You may like to solve the following exercises to acquire a better understanding of residual analysis.

E7) For the multiple regression model fitted in **E6** for studying the effect of two regressor variables: duration of daily walk and increase in the serum creatinine on the sugar level of diabetic patients,

i) obtain the residuals and verify that $\sum_{i=1}^{15} r_i = 0$,

ii) determine the standardised residuals,

iii) construct the residual and normal probability plots, and interpret the results.

E8) For the multiple regression model fitted in **E7** to check the relationship of birth weight on gestational age, increase in mother's weight and mother's height,

i) obtain the residuals and standardised residuals,

ii) verify the properties of residuals,

iii) construct the residual and normal probability plots, and interpret the results.

We now summarise what we have discussed in this unit.

7.6 SUMMARY

In this unit, we have discussed the following aspects of multiple linear regression:

1. The relationship between a response variable Y and a set of k regressor variables can be represented by a multiple linear regression model as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

where $\beta_0, \beta_1, \beta_2, ..., \beta_k$ are known as the regression coefficients.

2. In case of two regressor variables, we have obtained the values of regression coefficients using the method of least squares. Using this method, we obtained the following normal equations:

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_{1i} + \beta_2 \sum_{i=1}^{n} x_{2i} = \sum_{i=1}^{n} y_i$$

$$\beta_0 \sum_{i=1}^{n} x_{1i} + \beta_1 \sum_{i=1}^{n} x_{1i}^2 + \beta_2 \sum_{i=1}^{n} x_{1i} x_{2i} = \sum_{i=1}^{n} y_i x_{1i}$$

$$\beta_0 \sum_{i=1}^{n} x_{2i} + \beta_1 \sum_{i=1}^{n} x_{2i} x_{1i} + \beta^2 \sum_{i=1}^{n} x_{2i}^2 = \sum_{i=1}^{n} y_i x_{2i}$$

3. For two regressor variables, the normal equations are simultaneous linear equations and can be solved using standard methods. If the numbers of unknown parameters, i.e., regression coefficients are large, we use the matrix approach for estimating the regression coefficients.

4. We have defined the multiple linear regression model in matrix form as:

$$\underset{\sim}{Y} = \mathbf{X}\ \underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

where $\underset{\sim}{Y}_{n\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ y_n \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & ... & x_{k1} \\ 1 & x_{12} & x_{22} & ... & x_{k2} \\ 1 & x_{13} & x_{23} & ... & x_{k3} \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & x_{1n} & x_{2n} & ... & x_{kn} \end{bmatrix}$, $\underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_k \end{bmatrix}$ and $\underset{\sim}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ . \\ . \\ \varepsilon_n \end{bmatrix}$

5. We have represented the (k + 1) normal equations in matrix form as:

   $X'X \ \underset{\sim}{\beta} = X' \underset{\sim}{Y}$

6. We have obtained the least squares estimates $(\underset{\sim}{\hat{\beta}})$ of the regression coefficients as:

   $$\underset{\sim}{\hat{\beta}} = (X'X)^{-1} X' \underset{\sim}{Y}$$

   where $\underset{\sim}{\hat{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k)$.

7. We have determined the fitted value of Y for the i^{th} observation as:

   $$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_k x_{ki} \quad ; i = 1, 2, ..., n$$

   which can be written in matrix form as:

   $$\underset{\sim}{\hat{Y}} = X\underset{\sim}{\hat{\beta}}$$

   where $\underset{\sim}{\hat{Y}} = [\hat{y}_1, \hat{y}_2, ..., \hat{y}_n]$ is an $(n \times 1)$ vector of n predicted values of the response variable (Y).

8. We have defined the difference between the observed and the corresponding predicted (fitted) values of the response variable as residuals. We have also defined the i^{th} residual as:

   $R_i = y_i - \hat{y}_i \quad ; i = 1, 2, ..., n$

   In matrix form, we can write this as:

   $$\underset{\sim}{R} = \underset{\sim}{Y} - \underset{\sim}{\hat{Y}}$$

   where $\underset{\sim}{R} = [R_1, R_2, ..., R_n]$

9. We have used residual analysis to check the validity of some basic assumptions and to ensure the adequacy of the regression model. For detecting non-linearity and/or unequal variance of error terms, we have used the residual plot. We have used the normal probability plot for detecting non-normality in the error terms.

10. As the value of σ^2 was unknown, we have estimated it from the given data as:

    $$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} r_i^2}{n-k-1} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-k-1}$$

7.7 SOLUTIONS / ANSWERS

E1) Refer to **Sec. 5.2**.

E2) The scatterplot matrix for the data on the average duration of walk, serum creatinine and random blood sugar level of 15 diabetic patients is shown in Fig. 7.6.

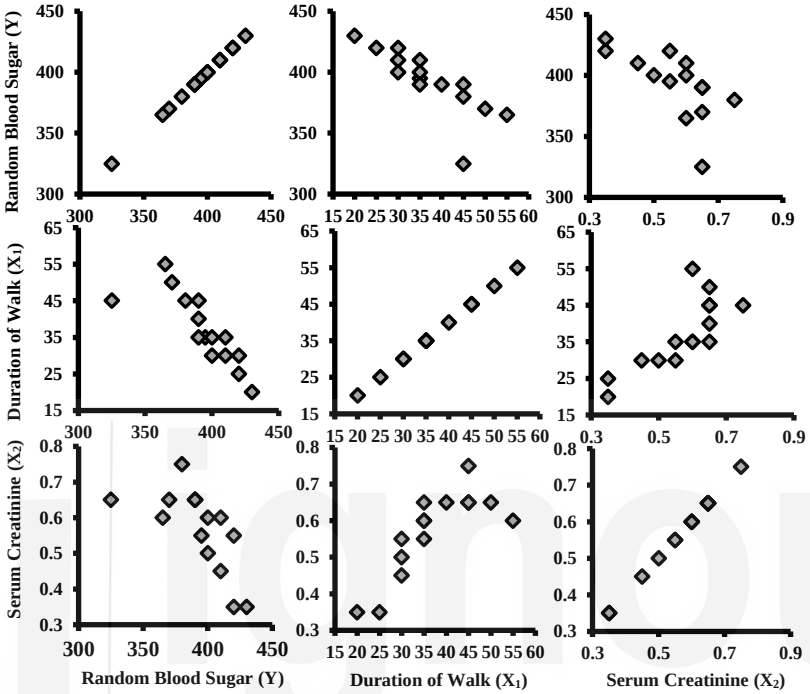

Fig. 7.6

E3) From the data given in **E2**, we have

$$\sum_{i=1}^{15} y_i = 5895, \quad \sum_{i=1}^{15} x_{1i} = 555, \quad \sum_{i=1}^{15} x_{2i} = 8.55,$$

$$\sum_{i=1}^{15} x_1^2 = 21825, \ \sum_{i=1}^{15} x_{2i}^2 = 5.0575$$

$$\sum_{i=1}^{15} x_{1i}x_{2i} = 328.5, \ \sum_{i=1}^{15} y_i x_{1i} = 215275 \text{ and } \sum_{i=1}^{15} y_i x_{2i} = 3332.5$$

$$\bar{y} = \frac{5895}{15} = 393, \quad \bar{x}_1 = \frac{555}{15} = 37 \text{ and } \bar{x}_2 = \frac{8.55}{15} = 0.57$$

On substituting the values of $\sum y_i, \sum x_{1i}, \sum x_{2i}, \sum yx_{1i}, \sum yx_{2i}$, $\sum x_{1i}^2$ and $\sum x_{2i}^2$ in the normal equations (7) to (9), we obtain

$$15\beta_0 + 555\beta_1 + 8.55\beta_2 = 5895$$

$$555\beta_0 + 21825\beta_1 + 328.5\beta_2 = 215275$$

$$8.55\beta_0 + 328.5\beta_1 + 5.0575\beta_2 = 3332.5$$

On solving the normal equations, we obtain

$$\hat{\beta}_0 = 477.3269, \ \hat{\beta}_1 = -2.0795 \text{ and } \hat{\beta}_2 = -12.9545$$

Thus, the fitted model is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\hat{Y} = 477.3269 - 2.0795\ X_1 - 12.9545\ X_2$$

For $X_1 = 30$ and $X_2 = 0.5$, the predicted value of blood sugar is computed as:

$$\hat{y} = 477.3269 - 2.0795 \times 30 - 12.9545 \times 0.5$$

$$= 408.4636$$

E4) From the given data, we have

$$\sum_{i=1}^{12} y_i = 36, \quad \sum_{i=1}^{12} x_{1i} = 456, \quad \sum_{i=1}^{12} x_{2i} = 120,$$

$$\sum_{i=1}^{12} x_1^2 = 17488.16, \ \sum_{i=1}^{12} x_{2i}^2 = 1290.5$$

$$\sum_{i=1}^{12} x_{1i} x_{2i} = 4651.05, \ \sum_{i=1}^{12} y_i x_{1i} = 1390.81 \text{ and } \sum_{i=1}^{12} y_i x_{2i} = 374.05$$

$$\bar{y} = \frac{36}{12} = 3, \quad \bar{x}_1 = \frac{456}{12} = 38 \text{ and } \bar{x}_2 = \frac{120}{12} = 10$$

On substituting the values of $\sum y_i, \sum x_{1i}, \sum x_{2i}, \sum y x_{1i}, \sum y x_{2i}$, $\sum x_{1i}^2$ and $\sum x_{2i}^2$ in the normal equations (7) to (9), we obtain

$$12\beta_0 + 456\beta_1 + 120\beta_2 = 36$$

$$456\beta_0 + 17488.16\beta_1 + 4651.05\beta_2 = 1390.81$$

$$120\beta_0 + 4651.05\ \beta_1 + 1290.5\beta_2 = 374.05$$

On solving the normal equations, we obtain

$$\beta_0 = -2.0877, \ \beta_1 = 0.1265 \text{ and } \beta_2 = 0.0279$$

Thus, the fitted model is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\hat{Y} = -2.0877 + 0.1265\ X_1 + 0.0279\ X_2$$

E5) We obtain $\hat{\beta}_1$ and $\hat{\beta}_2$ using the matrix approach.

From the solution of **E3**, we obtain $X'X$ and $X'Y$ using equations (24) and (25), respectively, as follows:

$$X'X = \begin{bmatrix} 15 & 555 & 8.55 \\ 555 & 21825 & 328.5 \\ 8.55 & 328.5 & 5.0575 \end{bmatrix}, \text{ and } X'Y = \begin{bmatrix} 5895 \\ 215275 \\ 3332.5 \end{bmatrix}$$

To compute $(X'X)^{-1}$, we need to determine $|X'X|$ and $\text{adj}\,(X'X)$:

$|X'X| = 1346.0625$

$$\text{Adj}(X'X) = \begin{bmatrix} 2467.6875 & 1.7625 & -4286.25 \\ 1.7625 & 2.76 & -182.25 \\ -4286.25 & -182.25 & 19350 \end{bmatrix}$$

We then compute the inverse matrix of $X'X$:

$$(X'X)^{-1} = \begin{bmatrix} 1.8333 & 0.0013 & -3.1843 \\ 0.0013 & 0.0021 & -0.1354 \\ -3.1843 & -0.1354 & 14.3753 \end{bmatrix}$$

From equation (28), $\underset{\sim}{\hat{\beta}} = (X'X)^{-1} X'\underset{\sim}{Y}$ and we get

$$\underset{\sim}{\hat{\beta}} = \begin{bmatrix} 1.8333 & 0.0013 & -3.1843 \\ 0.0013 & 0.0021 & -0.1354 \\ -3.1843 & -0.1354 & 14.3753 \end{bmatrix} \times \begin{bmatrix} 5895 \\ 215275 \\ 3332.5 \end{bmatrix}$$

$$= \begin{bmatrix} 477.3269 \\ -2.0795 \\ -12.9545 \end{bmatrix}$$

Thus, $\hat{\beta}_0 = 477.3269$, $\hat{\beta}_1 = -2.0795$ and $\hat{\beta}_2 = -12.9545$

The value of $\hat{\beta}_1 = -2.07953754$ indicates the expected increase or decrease of approximately 2.08 mg/dL in blood sugar corresponding to per minute decrease or increase in average walking time (X_1) when serum creatinine (X_2) is considered as constant. Similarly, one mg/dL increase or decrease in serum creatinine will result in $(\hat{\beta}_2 = -12.95445048 \cong) -12.95$, decrease or increase in sugar level when X_1 is held constant for the given data.

Thus, the fitted model $(\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2)$ is

$$\hat{Y} = 477.3269 - 2.0795\ X_1 - 12.9545\ X_2$$

You can see that the solutions of **E3** and **E6** are the same.

E6) The matrices $X'X$ and $X'Y$ using equations (24) and (25), respectively, can be written as:

$$X'X = \begin{bmatrix} 15 & 456 & 120 & 1908 \\ 456 & 17488.16 & 4651.05 & 72406.5 \\ 120 & 4651.05 & 1290.5 & 18983.5 \\ 1908 & 72406.5 & 18983.5 & 303850 \end{bmatrix}, \text{and}$$

$$X'\underset{\sim}{Y} = \begin{bmatrix} 36 \\ 1390.81 \\ 374.05 \\ 5710.3 \end{bmatrix}$$

Thus, $\underset{\sim}{\hat{\beta}} = (X'X)^{-1} X'\underset{\sim}{Y}$

$$\underset{\sim}{\hat{\beta}} = \begin{bmatrix} 15 & 456 & 120 & 1908 \\ 456 & 17488.16 & 4651.05 & 72406.5 \\ 120 & 4651.05 & 1290.5 & 18983.5 \\ 1908 & 72406.5 & 18983.5 & 303850 \end{bmatrix}^{-1} \begin{bmatrix} 36 \\ 1390.81 \\ 374.05 \\ 5710.3 \end{bmatrix}$$

As explained in **Example 2**, the determinant of $(X'X)$, i.e., $|X'X|$ and the adjoint matrix of $X'X$, i.e., $\text{adj}(X'X)$ can be determined as follows:

$$|X'X| = 27927122.82$$

$$\text{Adj}(X'X) = \begin{bmatrix} 2498585187.965 & -11457546.900 & -5041379.220 & -12644385.870 \\ -11457546.900 & 407361.000 & -409357.800 & -449.100 \\ -5041379.220 & -409357.800 & 804602.760 & 78936.780 \\ -12644385.870 & -449.100 & 78936.780 & 74452.530 \end{bmatrix}$$

We compute the inverse matrix of $X'X$ as

$$(X'X)^{-1} = \begin{bmatrix} 89.4680 & -0.4103 & -0.1805 & -0.4528 \\ -0.4103 & 0.0146 & -0.0147 & 0.00001 \\ -0.1805 & -0.0147 & 0.0288 & 0.0028 \\ -0.4528 & 0.00001 & 0.0028 & 0.0027 \end{bmatrix}$$

From equation (28), we have

$$\underset{\sim}{\hat{\beta}} = \begin{bmatrix} 89.4680 & -0.4103 & -0.1805 & -0.4528 \\ -0.4103 & 0.0146 & -0.0147 & 0.00001 \\ -0.1805 & -0.0147 & 0.0288 & 0.0028 \\ -0.4528 & 0.00001 & 0.0028 & 0.0027 \end{bmatrix} \times \begin{bmatrix} 36 \\ 1390.81 \\ 374.05 \\ 5710.3 \end{bmatrix}$$

$$= \begin{bmatrix} -2.6916 \\ 0.1266 \\ 0.0317 \\ 0.00355590 \end{bmatrix}$$

Thus, $\beta_0 = -2.69159728$, $\beta_1 = 0.1266$, $\beta_2 = 0.0317$ and $\beta_3 = 0.0036$

Thus, the fitted model $(\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2)$ is

$$\hat{Y} = -2.6916 + 0.1266\ X_1 + 0.0317\ X_2 + 0.0036\ X_3$$

E7) From the solution of exercise **E3** of Unit 7, the best fitted regression model is given by

$$\hat{Y} = 477.3269 - 2.0795\ X_1 - 12.9545\ X_2$$

We calculate the predicted values $(\hat{y}_i)$, residuals (r_i), squares of the residual $\left(r_1^2\right)$ and standardised residuals (s_i), ordered standardised residuals and percentiles cumulative probabilities for all given observations in the data as follows:

S. No.	Y_i	Predicted Values $(\hat{y}_i)$	Residuals (r_i)	r_i^2	Standardised Residuals $s_i = \frac{r_i}{\hat{\sigma}}$	Ordered Standardised Residuals $s_{(i)}$	Percentiles Cumulative Probabilities (P_i)
1	430	431.2021	– 1.2021	1.4451	– 0.0717	– 3.0005	3.3333
2	420	420.8044	– 0.8044	0.6471	– 0.0480	– 0.5046	10.0000
3	410	409.1113	0.8887	0.7898	0.0530	– 0.3650	16.6667
4	400	408.4636	– 8.4636	71.6321	– 0.5046	– 0.1442	23.3333
5	390	375.3273	14.6727	215.2868	0.8748	– 0.0717	30.0000
6	395	397.4182	– 2.4182	5.8475	– 0.1442	– 0.0480	36.6667
7	420	407.8159	12.1841	148.4535	0.7264	0.0530	43.3333
8	410	396.7704	13.2296	175.0212	0.7888	0.1925	50.0000
9	400	396.7704	3.2296	10.4300	0.1925	0.2549	56.6667
10	390	385.7250	4.2750	18.2754	0.2549	0.3023	63.3333
11	380	374.0319	5.9681	35.6182	0.3558	0.3558	70.0000
12	370	364.9297	5.0703	25.7084	0.3023	0.5855	76.6667
13	390	396.1227	– 6.1227	37.4877	– 0.3650	0.7264	83.3333
14	365	355.1797	9.8203	96.4385	0.5855	0.7888	90.0000
15	325	375.3273	– 50.3273	2532.8415	– 3.0005	0.8748	96.6667
Total	5895	5895	0	3375.9228	0		

From Column 3 of the above table, it is clear that $\sum_{i=1}^{15} r_i \cong 0$. This verifies the property of the residual. We have n – k–1 = 15 − 3 = 12 and

$$\sum_{I=1}^{15} r_i^2 = 3375.9228$$

From equation (35), the variance of the residuals is estimated as:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{15} r_i^2}{12} = \frac{3375.9228}{12} = 281.3269$$

or $\hat{\sigma} = \sqrt{281.3269} = 16.7728$

To obtain a residual plot, we consider the predicted Y values and standardised residuals on the horizontal and the vertical axes, respectively as shown in Fig. 7.7.

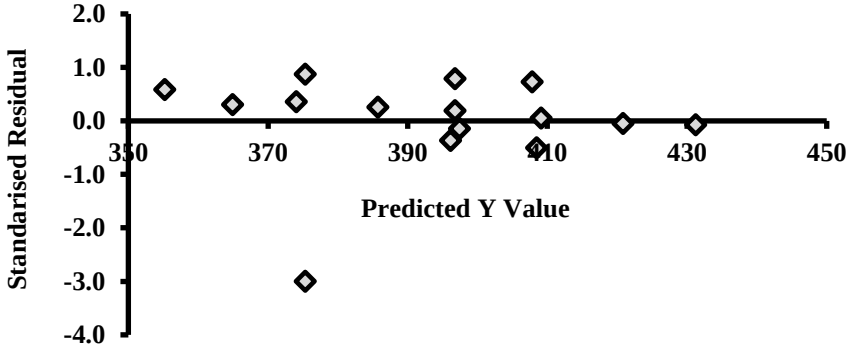

Fig. 7.7

The standardised residuals shown in Fig. 7.7 appear to be in slightly decreasing pattern. Hence, the assumption of linear regression does not

seem to be valid and also the value of 15th standardised residual is greater than 3. It is considered as an outlier.

Next, we plot the ordered standardised residuals against the percentiles and obtain a normal probability plot as shown in Fig. 7.8.

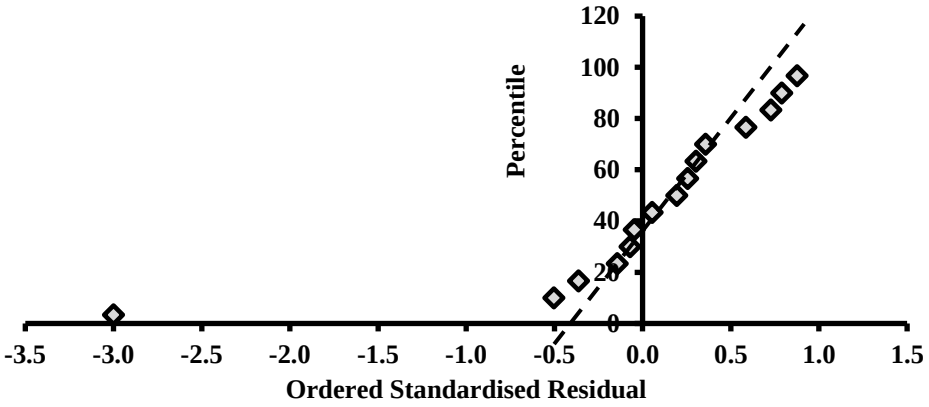

Fig. 7.8

Note that the resulting points are not lying approximately on a straight line as shown in Fig. 7.8. Notice that some points of the distribution are deviating slightly from the straight line and the 15th point lies very far from the central points. It indicates that the distribution of error terms is not normally distributed.

E8) From the solution of exercise **E3**, the best fitted regression model is given by

$$\hat{Y} = -2.6916 + 0.1266\ X_1 + 0.0317\ X_2 + 0.0036\ X_3$$

We calculate the predicted values $(\hat{Y}_i)$, residuals (r_i), squares of the residual $\left(r_i^2\right)$ and standardised residuals (s_i), ordered standardised residuals and percentiles cumulative probabilities for all given observations in the data as follows:

S. No.	Y_i	Predicted Values $(\hat{y}_i)$	Residuals (r_i)	r_i^2	Standardised Residuals $s_i = \frac{r_i}{\hat{\sigma}}$	Ordered Standardised Residuals $s_{(i)}$	Percentiles cumulative probabilities (P_i)
1	2.4	2.3737	0.0263	0.0007	0.0962	– 1.8919	4.1667
2	2.3	2.4771	– 0.1771	0.0314	– 0.6487	– 0.6487	12.5000
3	2.0	1.8884	0.1116	0.0124	0.4086	– 0.4205	20.8333
4	2.9	2.9651	– 0.0651	0.0042	– 0.2386	– 0.2501	29.1667
5	2.7	2.7683	– 0.0683	0.0047	– 0.2501	– 0.2386	37.5000
6	3.2	3.7165	– 0.5165	0.2668	– 1.8919	0.0962	45.8333
7	3.4	3.3579	0.0421	0.0018	0.1542	0.1535	54.1667
8	2.8	2.9148	– 0.1148	0.0132	– 0.4205	0.1542	62.5000
9	3.2	3.1581	0.0419	0.0018	0.1535	0.1812	70.8333
10	3.7	3.4914	0.2086	0.0435	0.7638	0.4086	79.1667
11	4.0	3.5379	0.4621	0.2135	1.6923	0.7638	87.5000
12	3.4	3.3505	0.0495	0.0024	0.1812	1.6923	95.8333
Total	36	36	0	0.5964	0		

From Column 3 of the above table, it is clear that $\sum_{i=1}^{12} r_i \cong 0$. This verifies the property of the residual.

We have n – k–1 = 12 – 3 –1= 8 and $\sum_{i=1}^{12} r_i^2 = 0.5964$

From equation (35), the variance of the residuals is estimated as:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{12} r_i^2}{8} = \frac{0.5964}{8} = 0.0746$$

$$\text{or } \hat{\sigma} = \sqrt{0.0746} = 0.2730$$

To obtain a residual plot, we consider the predicted Y values and standardised residuals on the horizontal and the vertical axes, respectively, shown in Fig. 7.9.

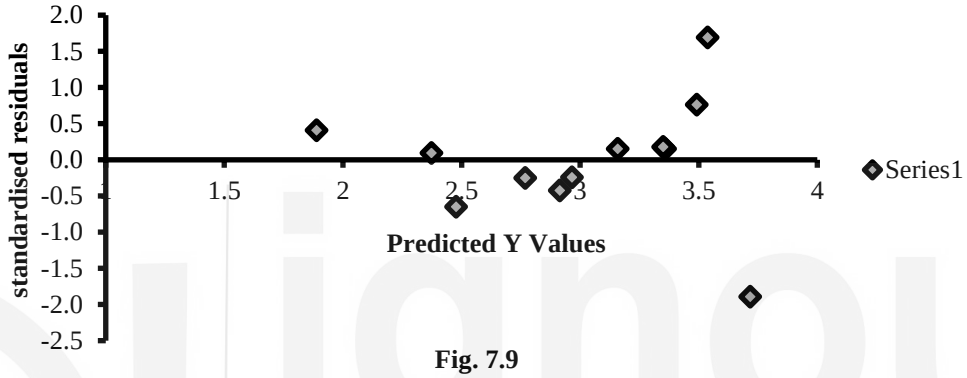

Fig. 7.9

The standardised residuals shown in Fig. 7.9 appear to have a curved pattern. Hence, the assumption of linear regression does not seem to be valid.

Next, we plot the ordered standardised residuals against the percentiles and obtain a normal probability plot shown in Fig. 7.10.

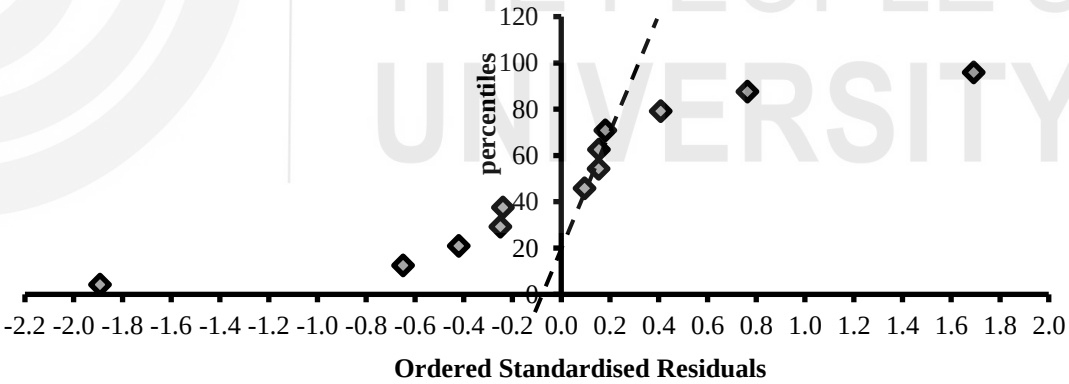

Fig. 7.10

Note from Fig. 7.10. that the resulting points do not lie approximately on a straight line. Notice that some points of the distribution deviate slightly from the straight line. This indicates that the distribution of error terms is not normally distributed.