UNIT 11 MULTIPLE LINEAR REGRESSION

Structure

11.1 INTRODUCTION

In previous units, we have discussed the linear relationship between the dependent variable Y and an independent variable X. The coefficients a and b were unknown and for the given data on Y and X, we have obtained least squares estimates of parameters, i.e., $\hat{a}$ and $\hat{b}$. We have also gone through the inferential study to examine whether there exists a significant linear relationship between Y and X or not. We have discussed the simple linear regression model and estimation of model parameters, and determined standard errors.

In this unit, we discuss the multiple linear regression model along with the estimation of parameters in Secs. 11.2 and 11.3. In multiple linear regression, the basic concept is the same as that of simple regression. However, instead of one independent variable, there are several independent variables, say, X_1, X_2, X_3, …, X_p. For example, the number of units sold by a car manufacturing company per year may not depend on only one independent variable such as price, but also on mileage per unit of fuel, appearance of the car, comfort level, durability and money spent on advertising, etc. Here we may like to identify the important independent variables, which contribute more to the variation in the dependent variable(s). For this purpose, a mathematical relationship between the dependent and independent variables is established and this relation is further used for prediction purposes. We also discuss the inferential study in multiple linear regression in Sec. 11.4.

Since the model may involve several independent variables affecting the dependent variable because of their relationship via regression, it may be of interest to estimate their importance by estimating regression coefficients along with their standard errors. The adequacy of model fit may be examined by overall fit of the model with the help of coefficient of determination (R^2). In this unit, we also discuss a method for calculating R^2 and adjusted R^2 in Sec. 11.5. The regression analysis with dummy variables is also discussed in Sec. 11.6.

In the next unit, we shall discuss how to calculate the extra sum of squares explained by the regressor variables on the response variable. We shall also discuss the methods of selection of important regressor variables which play an important role in selection of the best fitted models.

Objectives

After studying this unit, you should be able to:

- explain the concept of multiple linear regression;
- formulate a multiple linear regression model;
- estimate the regression coefficients and their standard errors;
- calculate the coefficient of determination (R^2) and adjusted R^2; and
- predict the dependent variable for given values of independent variables.

11.2 MULTIPLE LINEAR REGRESSION MODEL

In this section, we generalise the simple regression model considered in Unit 9. We have assumed in Unit 9 that (Y_1, X_1), (Y_2, X_2),…, (Y_n, X_n) are n pairs of values. The equation of the simple linear regression model may be written as

$$Y = a + bX + e$$

where e represents the error term, which arises from the difference of the observed Y and the straight line Y = a + bX. To minimise the term e, we use the method of least squares. From the above equation, we may write a simple regression model as

$$Y_i = a + bX_i + e_i \qquad i = 1, 2, \ldots, n$$

for the sample data of n pairs given in terms of (Y_i, X_i) (i= 1, 2, …n).

In agriculture, the crop yield depends on more than one variable such as fertility of the soil, amount of rainfall, amount of fertilisers, etc. A multiple regression model that might describe this relationship is

$$Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + e$$

where Y denotes the yield, X_1 denotes the fertility of soil, X_2 denotes the rainfall and X_3 denotes the amount of fertilisers used. This is called the **multiple linear regression model** with three independent/regressor variables. The term linear is used because the dependent/response variable Y is a linear function of the unknown parameters B_0, B_1, B_2 and B_3.

In general, the response variable may be related to p regressors or independent variables. Let Y be the dependent variable and $X_1, X_2, ..., X_p$ be p independent variables. Then the multiple regression model can be written as:

$$Y = B_0 + B_1X_1 + B_2X_2 + ... + B_pX_p + e \qquad \ldots (1)$$

The parameters $B_0, B_1, \ldots, B_p$ are called the **regression coefficients**. The parameters B_i (i = 0, 1, 2, …, p) represent the expected change in the response variable Y per unit change in X_i when the remaining regressor variables are treated as constant.

For the sake of simplicity, we shall attach a dummy variable X_0 with the intercept B_0; X_0 takes value 1 for all n observations. Now the model in equation (1) can be written as:

$$Y = B_0X_0 + B_1X_1 + B_2X_2 + ... + B_pX_p + e \qquad \ldots (2)$$

The simple regression model considered in Unit 9 becomes a particular case of this model with $X_0 = 1$, $B_0 = a$, $B_1 = b$ and $B_i = 0$, $(i \geq 2)$. The interpretation of coefficients B_j ($j = 1, 2, \ldots, p$) is that B_j represents the amount of change in Y for a unit change in X_j, keeping the other independent variables X_k $(k \neq j)$ fixed. These coefficients are known as **partial regression coefficients** as the effect of one independent variable is studied on the dependent variable while the other variables are held fixed or constant. We use the term multiple linear regression for this model because two or more than two variables are included in the regression analysis and the parameters $B_0, B_1, \ldots, B_p$ appear in a linear form. Moreover, the effect of these variables can be studied jointly. Here X_i can be any continuous function such as log X, X^2, X^3, X^{-1}, etc. However, it is necessary that the equation is linear. Let us consider a polynomial model

$$Y = B_0 + B_1X + B_2X^2 + ... + B_pX^p + e$$

If we let $X_1 = X$, $X_2 = X^2$, $X_3 = X^3$ and so on, the above model can be written in a linear form as given in equation (2).

As in the case of simple linear regression (Unit 9), here too we make the assumptions that e is normally and independently distributed with mean zero and constant variance σ^2.

11.3 ESTIMATION OF MODEL PARAMETERS

Recall that in Unit 9, we have estimated the parameters a and b of a simple linear regression equation using the method of least squares. In this method, we minimise the total error term, so that the sum of the squares of the differences between the observed values and their expected values is minimum, i.e., the sum of squares of the error terms is minimum. We also use the method of least squares to estimate the regression coefficients given in equation (2).

Let the number of observations be n (> p). Let y_i denote the ith observed value and x_{ij} denote the jth observation of the regressor variable X_i. The data is represented as given in the table below:

Response Variable Y	**Regressor Variables**			
	$\mathbf{X_1}$	$\mathbf{X_2}$	. . . .	$\mathbf{X_p}$
y_1	x_{11}	x_{21}	. . . .	x_{p1}
y_2	x_{12}	x_{22}	. . . .	x_{p2}
y_3	x_{13}	x_{23}	. . . .	x_{p3}
. . .	. . .	. . .	. . . .	. . .
y_n	x_{1n}	x_{2n}	. . . .	x_{pn}

Then the multiple regression model for the ith observation Y_i can be written as:

$$Y_i = B_0 + B_1X_{1i} + B_2X_{2i} + ... + B_pX_{pi} + e_i, \qquad i = 1, 2, ..., n$$

where $X_{1i}, X_{2i}, ..., X_{pi}$ are the corresponding values of p independent variables, B_0 is the intercept, $B_1, B_2, ..., B_p$ are p regression coefficients corresponding to independent variables $X_1, X_2, \ldots, X_p$, respectively.

We now minimise $\sum e_i^2$, the sum of squares of errors in the model given in equation (2):

$$E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\left(Y_i - B_0\, X_{0i} - B_1\, X_{1i} - \ldots - B_p X_{pi}\right)^2 \qquad \ldots (3)$$

with respect to B_0, B_1, ..., B_p to obtain their least squares estimates. For estimating the model parameters $B_0, B_1, B_2, \ldots, B_p$, we differentiate E with respect to $B_0, B_1, B_2, \ldots, B_p$, respectively, and equate the result to zero. If we differentiate E with respect to B_j, we obtain the j^{th} ($j = 0, 1, \ldots, p$) normal equations as follows:

$$\frac{\partial E}{\partial B_j} = -2\sum_{i=1}^{n}\left(Y_i - B_0 X_{0i} - B_1 X_{1i} - \ldots - B_p X_{pi}\right)X_{ji} = 0, \quad j = 0, 1, 2, \ldots, p \ \ldots (4)$$

Simplifying equation (4), we obtain the least squares normal equations:

$$\begin{array}{l}
nB_0 + B_1\sum_{i=1}^{n} X_{1i} + B_2\sum_{i=1}^{n} X_{2i} + \ldots + B_p\sum_{i=1}^{n} X_{pi} = \sum_{i=1}^{n} Y_i \\
B_0\sum_{i=1}^{n} X_{1i} + B_1\sum_{i=1}^{n} X_{1i}^2 + B_2\sum_{i=1}^{n} X_{1i}X_{2i} + \ldots + B_p\sum_{i=1}^{n} X_{1i}X_{pi} = \sum_{i=1}^{n} X_{1i}Y_i \\
B_0\sum_{i=1}^{n} X_{2i} + B_1\sum_{i=1}^{n} X_{2i}X_{1i} + B_2\sum_{i=1}^{n} X_{2i}^2 + \ldots + B_p\sum_{i=1}^{n} X_{2i}X_{pi} = \sum_{i=1}^{n} X_{2i}Y_i \\
. \qquad\qquad . \qquad\qquad . \qquad\qquad . \qquad\qquad . \qquad\qquad \ldots (5) \\
. \qquad\qquad . \qquad\qquad . \qquad\qquad . \qquad\qquad . \\
. \qquad\qquad . \qquad\qquad . \qquad\qquad . \qquad\qquad . \\
B_0\sum_{i=1}^{n} X_{pi} + B_1\sum_{i=1}^{n} X_{pi}X_{1i} + B_2\sum_{i=1}^{n} X_{pi}X_{2i} + \ldots + B_p\sum_{i=1}^{n} X_{pi}^2 = \sum_{i=1}^{n} X_{pi}Y_i
\end{array}$$

These are p + 1 normal equations and can be solved using the methods of solving simultaneous linear equations.

The solutions of the above normal equations called the **least squares estimates** are $\hat{B}_0, \hat{B}_1, \hat{B}_2, \ldots, \hat{B}_p$, respectively.

For simplicity, we shall rewrite the model in equation (1) by centralising the independent variables $X_1, X_2, \ldots\ldots\ldots.., X_p$, i.e., by taking differences from their means:

$$\begin{aligned}
Y &= B_0 + B_1\bar{X}_1 + \ldots + B_p\bar{X}_p + B_1\left(X_1 - \bar{X}_1\right) + \ldots.. + B_p\left(X_p - \bar{X}_p\right) + e_i \\
&= B_0'\, X_0 + B_1(X_1 - \bar{X}_1) + \ldots\ldots\ldots. + B_p(X_p - \bar{X}_p) + e
\end{aligned}$$

where $B_0' = B_0 + B_1\bar{X}_1 + \ldots + B_p\bar{X}_p$. Here $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_p$ are the means of p independent/ regressor variables. With this, the normal equation becomes

$$\frac{\partial E}{\partial B_j} = -2\sum_{i=1}^{n}\left(Y_i - B_0'\, X_{0i} - B_1\left(X_{1i} - \bar{X}_1\right) - \ldots\ldots - B_p\left(X_{pi} - \bar{X}_p\right)\right))(X_{ji} - \bar{X}_j) = 0, \qquad \ldots (6)$$

Note that $X_{0i} = 1$ for all i. The coefficients $B_1, B_2, \ldots, B_p$ remain the same, but the intercept changes from B_0 to B_0'. Once we have obtained the estimates of B_0', $B_1, B_2, \ldots, B_p$, we can obtain $\hat{B}_0$ from the following equation:

$$\hat{B}_0 = \hat{B}_0^{'} - \hat{B}_1\bar{X}_1 - \ldots\ldots\ldots. - \hat{B}_p\bar{X}_p \qquad \ldots (7)$$

Let us consider an application of these results.

Example 1: A statistical analyst is analysing the vending machine routes in the distribution system. He/she is interested in predicting the amount of time required by the route driver to service the vending machines in an outlet. The company manager responsible for the study has suggested that the two most important variables affecting the delivery time Y (in minutes), are
(i) the number of cases (X_1) and (ii) the distance travelled (in m) by the route driver (X_2). The delivery time data collected by the statistical analyst is given below:

Time (Y)	No. of Cases (X_1)	Distance (X_2)
20	10	50
10	5	20
10	5	30
15	5	10
15	10	10
20	10	30
10	5	10
25	15	40
30	10	80
15	10	20
20	10	10
10	5	40

Check whether there is a linear relationship between Y (Time) and the two independent variables X_1 (number of cases) and X_2 (distance). Calculate the values of the regression coefficients and fit the regression equation.

Solution: To find the values of regression coefficients and fit the regression equation for the given data, we form the following table:

Time (Y)	No. of Cases (X_1)	Distance (X_2)	Y^2	$(X_1)^2$	$(X_2)^2$	X_1Y	X_2Y	X_1X_2
20	10	50	400	100	2500	200	1000	500
10	5	20	100	25	400	50	200	100
10	5	30	100	25	900	50	300	150
15	5	10	225	25	100	75	150	50
15	10	10	225	100	100	150	150	100
20	10	30	400	100	900	200	600	300
10	5	10	100	25	100	50	100	50
25	15	40	625	225	1600	375	1000	600
30	10	80	900	100	6400	300	2400	800
15	10	20	225	100	400	150	300	200
20	10	10	400	100	100	200	200	100
10	5	40	100	25	1600	50	400	200
$\sum Y_i$ =200	$\sum X_{1i}$ =100	$\sum X_{2i}$ =350	$\sum Y_i^2$ =3800	$\sum X_{1i}^2$ =950	$\sum X_{2i}^2$ =15100	$\sum X_{1i}Y_i$ =1850	$\sum X_{2i}Y_i$ =6800	$\sum X_{1i}X_{2i}$ =3150

On putting the values from the above table in the normal equations (5) for p = 2, and noting that $X_0 = 1$, we get

$$n\hat{B}_0 + \hat{B}_1 \sum X_{1i} + \hat{B}_2 \sum X_{2i} = \sum Y_i$$

$$\hat{B}_0 \sum X_{1i} + \hat{B}_1 \sum X_{1i}^2 + \hat{B}_2 \sum X_{1i} X_{2i} = \sum Y_i X_{1i}$$

$$\hat{B}_0 \sum X_{2i} + \hat{B}_1 \sum X_{1i} X_{2i} + \hat{B}_2 \sum X_{2i}^2 = \sum Y_i X_{2i}$$

On putting the values calculated in the table in the above equations, we get

$$12\,\hat{B}_0 + 100\,\hat{B}_1 + 350\,\hat{B}_2 = 200 \qquad \ldots \text{(i)}$$

$$100\,\hat{B}_0 + 950\,\hat{B}_1 + 3150\,\hat{B}_2 = 1850 \qquad \ldots \text{(ii)}$$

$$350\,\hat{B}_0 + 3150\,\hat{B}_1 + 15100\,\hat{B}_2 = 6800 \qquad \ldots \text{(iii)}$$

From equation (i), we have

$$\hat{B}_0 = \frac{\left(200 - 100\,\hat{B}_1 - 350\,\hat{B}_2\right)}{12} \qquad \ldots \text{(iv)}$$

On putting the value of $\hat{B}_0$ in equations (ii) and (iii) and simplifying, we get

$$1400\,\hat{B}_1 + 2800\,\hat{B}_2 = 2200 \qquad \ldots \text{(v)}$$

$$2800\,\hat{B}_1 + 58700\,\hat{B}_2 = 11600 \qquad \ldots \text{(vi)}$$

On solving equations (v) and (vi), we get

$$\hat{B}_1 = 1.3002 \qquad \hat{B}_2 = 0.1356$$

and $$\hat{B}_0 = \frac{\left(200 - 100\,(1.3002) - 350(0.1356)\right)}{12} = 1.8765$$

Hence, the fitted equation is

$$Y = 1.8765 + 1.3002\ X_1 + 0.1356\ X_2$$

So we can conclude that there is a linear relationship between Y (time in seconds) and the two independent variables X_1 (number of cases) and X_2 (distance). As the regression coefficients for both variables are positive, these affect the delivery time. The numerical value of the regression coefficient $\hat{B}_1$ associated with X_1 is higher than the value of $\hat{B}_2$ associated with X_2. It shows that the number of cases affects the delivery time more than the distance travelled.

11.3.1 Use of Matrix Notation

When p is greater than 2, it is more convenient to write the normal equations in matrix form. The regression equations in matrix notation can be written as:

$$Y = X\,B + e \qquad \ldots (8)$$

where

$$Y_{n\times 1}=\begin{pmatrix} y_1\\ y_2\\ .\\ .\\ y_n\end{pmatrix}, X_{n\times(p+1)}=\begin{pmatrix} 1 & X_{11} & . & X_{p1}\\ 1 & X_{12} & . & X_{p2}\\ . & . & . & .\\ . & . & . & .\\ 1 & X_{n1} & . & X_{pn}\end{pmatrix}, B_{(p+1)\times 1}=\begin{pmatrix} B_0\\ B_1\\ .\\ .\\ B_p\end{pmatrix} \text{ and } e_{n\times 1}=\begin{pmatrix} e_1\\ e_2\\ .\\ .\\ e_n\end{pmatrix}$$

In general, Y is an n×1 vector of the observed values of the response variable Y, X is a (p+1)×n matrix of the values of regressor variables, B is a (p+1) ×1 vector of regression coefficients and e is an n×1 vector of random errors.

In matrix notation, the (p+1) normal equations can be written as follows:

$$X'X\hat{B} = X'Y \qquad \text{… (9a)}$$

Equation (9a) represents the normal least squares equations. For the sake of simplicity, we may write them as

$$\begin{pmatrix} n & \sum x_{1i} & . & \sum x_{pi}\\ \sum x_{1i} & \sum x_{1i}^2 & . & \sum x_{1i}x_{pi}\\ . & . & . & .\\ . & . & . & .\\ \sum x_{pi} & \sum x_{pi}x_{1i} & . & \sum x_{pi}^2\end{pmatrix}\begin{pmatrix} B_0\\ B_1\\ .\\ .\\ B_p\end{pmatrix}=\begin{pmatrix} \sum y_i\\ \sum x_{1i}y_i\\ .\\ .\\ \sum x_{pi}y_i\end{pmatrix} \qquad \text{… (9b)}$$

To solve the normal least squares equations given in equation (9a), we multiply both sides by the inverse of $X'X$. Thus, the estimates of the regression coefficients are given by

$$\hat{B} = \left(X'X\right)^{-1}X'\,Y \qquad \text{… (10)}$$

On putting the values of the estimates in equation (2), we get the fitted regression model corresponding to the observations of the regressor variables X_1, X_2, …, X_p as

$$\hat{Y} \;= \hat{B}_0 + \hat{B}_1X_1 + \hat{B}_2X_2 + ... + \hat{B}_pX_p + e \qquad \text{… (11)}$$

The matrix representation of the fitted values corresponding to the observed values are similar to the equation (9a) and are given as

$$\hat{Y} = X\,\hat{B} = X\left(X'\,X\right)^{-1}X'\,Y \qquad \text{… (12)}$$

The difference between the observed value y_i and the corresponding estimated value $\hat{y}_i$ is called the i^{th} residual r_i, i.e.,

$$r_i = y_i - \hat{y}_i \qquad i = 1, 2, 3, ..., n.$$

The residuals may be written in matrix notation as

$$r = Y - \hat{Y} \quad = \quad Y - X\,\hat{B} \qquad \text{… (13)}$$

Here, we shall use the following notation:

$$Y_{n\times 1}=\begin{pmatrix} Y_1\\ .\\ .\\ Y_n\end{pmatrix},$$

$$X_{n\times(p+1)}=\begin{pmatrix} X_{01} & X_{11}-\overline{X}_1 & . & X_{p1}-\overline{X}_p\\ . & . & . & ..\\ . & . & .. & .\\ X_{0n} & X_{n1}-\overline{X}_1 & . & X_{pn}-\overline{X}_p\end{pmatrix}.$$

Note that X_{0i}'s are all unity and other variables are centralised (deviations from mean). Then (k+1) normal equations can be written as

$$X¢X\hat{B} = X¢Y$$

where $\hat{B}' = (\hat{B}_0', \hat{B}_1, \text{-----}, \hat{B}_n)$

In case $X¢X$ is non-singular, i.e. ($X¢X$) is of rank (p+1), then least squares estimates of B, denoted by $\hat{B}$, can be written as

$$\hat{B} = (X¢X)^{-1}X¢Y$$

11.3.2 Properties of Least Squares Estimates

We now describe the statistical properties of least squares estimates. When X_j, (j = 0, 1, ..., p) are linearly related, $(X'X)$ is not invertible. In this case we cannot obtain unique estimates of B. We shall not consider this case any more.

It is to be noted that $\hat{B}$ is an unbiased estimate of B because

$$\begin{aligned} E(\hat{B}) &= (X'X)^{-1}X'E(Y) \\ &= (X'X)^{-1}X'E(XB+e) = (X'X)^{-1}(X'X)B \\ &= B \text{ since } E(e) = 0 \text{ and } (X'X)^{-1}(X'X) = I \end{aligned}$$

This shows that $\hat{B}$ is unbiased.

The variance of Y (which is actually the variance-covariance matrix as Y is a vector) is given as

$$V(Y) = s^2 I_n$$

where I_n is an identity matrix of order n. The variance-covariance matrix of $\hat{B}$ is given by

$$\begin{aligned} V(\hat{B}) &= (X'X)^{-1}X'\,V(Y)X(X'X)^{-1} \\ &= (X'X)^{-1}X'\sigma^2 I_n X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \qquad \dots (14) \end{aligned}$$

where

$$X'X = \begin{bmatrix} n & \sum X_{1i} & \sum X_{2i} & . \; . \; . & \sum X_{pi} \\ \sum X_{1i} & \sum X_{1i}^2 & \sum X_{1i}X_{2i} & & \sum X_{1i}X_{pi} \\ \sum X_{2i} & \sum X_{1i}X_{2i} & \sum X_{2i}^2 & & \sum X_{2i}X_{pi} \\ . & . & . & & . \\ . & . & . & & . \\ \sum X_{pi} & \sum X_{1i}X_{pi} & \sum X_{2i}X_{pi} & & \sum X_{pi}^2 \end{bmatrix}$$

and $s_{jk} = \sum_{i=1}^{n} X_{ji}\,X_{ki}$.

Here $V(\hat{B}) = \sigma^2 (X'X)^{-1}$ is a (p+1)×(p+1) matrix and its diagonal elements give the variances of coefficients and off diagonal elements give the covariances. If we use the notation

$$V(\hat{B}) = s^2 (X'X)^{-1} = (s_{jk}), \qquad j,k = 0,1, ..., p$$

we can write

$$V(\hat{B}_j) = s_{jj}, \quad \text{and} \quad Cov(\hat{B}_j, \hat{B}_k) = s_{jk} \qquad \dots (15)$$

The standard error of $\hat{B}_j$ is given by

$$S.E.\ (\hat{B}_j) = \sqrt{s_{jj}} \qquad \dots (16)$$

The residual sum of squares SS_{Res} is obtained by substituting the least squares estimates of $B_0, B_1, \ldots, B_p$ in equation (3):

$$SS_{Res} = \sum_{i=1}^{n}\left(Y_i - \hat{B}_0 X_{oi} - \hat{B}_1 X_{1i} - \ldots - \hat{B}_p X_{pi}\right)^2$$

This is the sum of squares not accounted for by the regression model. In matrix notation, this can be written as

$$SS_{Res} = Y'Y - Y'X\hat{B}$$

$$= \sum Y_i^2 - \hat{B}_0(Y'X_0) - \hat{B}_1(Y'X_1) - \hat{B}_2(Y'X_2)\ldots - \hat{B}_p(Y'X_p) \quad \ldots(17)$$

Note that $X_1, X_2, \ldots, X_p$ are deviations from respective means. As we have fitted (p +1) parameters, the degree of freedom of residual sum of squares is (n – p –1). An unbiased estimate of σ^2 is obtained by dividing the residual sum of squares, i.e., SS_{Res}, by its degree of freedom (n – p –1). Thus

$$\hat{\sigma}^2 = SS_{Res}/(n - p - 1) \quad \ldots (18)$$

If we are interested in predicting the mean value of Y for a given set of independent variables $X_1, \ldots, X_p$, then we use the fitted model. The predicted mean value of Y for given $X_{10}, \ldots, X_{p0}$ is given by

$$\hat{Y}_0 = \hat{B}_0 + \hat{B}_1 X_{10} + \ldots\ldots.. + \hat{B}_p X_{p0}$$

Let us explain the matrix method with the help of an example.

Example 2: Using the data of Example 1, find the estimate of regression coefficients and SS_{Res} by using the matrix method. Also predict the expected time Y at $X_1 = 7$, $X_2 = 20$.

Solution: Using the matrix notation we have from the data:

$$Y^{¢}_{1'12} = [20, 10, 10, 15, 15, 20, 10, 25, 30, 15, 20, 10]$$

$$X'X = \begin{pmatrix} 12 & 100 & 350 \\ 100 & 950 & 3150 \\ 350 & 3150 & 15100 \end{pmatrix}, \qquad X'Y = \begin{pmatrix} 200 \\ 1850 \\ 6800 \end{pmatrix}$$

$$(X'X)^{-1} = \begin{pmatrix} 0.7139 & -0.0658 & -0.0028 \\ -0.0658 & 0.0095 & -0.0005 \\ -0.0028 & -0.0005 & 0.0002 \end{pmatrix}$$

and

$$\begin{pmatrix} \hat{B}_0' \\ \hat{B}_1 \\ \hat{B}_2 \end{pmatrix} = (X'X)^{-1} X'Y$$

$$= \begin{pmatrix} 0.7139 & -0.0658 & -0.0028 \\ -0.0658 & 0.0095 & -0.0005 \\ -0.0028 & -0.0005 & 0.0002 \end{pmatrix} \times \begin{pmatrix} 200 \\ 1850 \\ 6800 \end{pmatrix} = \begin{pmatrix} 1.8765 \\ 1.3002 \\ 0.1356 \end{pmatrix}$$

Hence the fitted equation is

$$Y = 1.8765 + 1.3002\ X_1 + 0.1356\ X_2$$

Now, we calculate the value of residual sum of squares to obtain an estimate of $\hat{\sigma}^2$ as follows:

$$SS_{Res} = YY' - Y'X\hat{B}_0$$

$$= 3800 – 200 \times (1.8765) – 1850 \times (1.3002) – 6800 \times (0.1356)$$

$$= 3800 – 375.3 – 2405.37 – 922.08 = 97.25$$

Therefore, on putting the value of SS_{Res} in equation (18), we get

$$\hat{\sigma}^2 = 97.25/(12-3) = 10.8017$$

Using the above results and putting the values $X_1 = 7$ and $X_2 = 20$ in the fitted equation for multiple regression, we get

$$\hat{Y} = 1.8765 + 1.3002\ X_1 + 0.1356\ X_2$$

$$\hat{Y} = 1.8765 + 1.3002 \times 7 + 0.1356 \times 20 = 13.6899$$

As far as the interpretation of coefficients is concerned, there is an increase of 1.3002 seconds in time for one unit increase in X_1. Similarly, for one unit increase in X_2 there is an increase of 0.1356 seconds in time.

You may like to pause here and solve the following exercises to check your understanding.

E1) In a study of 10 firms, the dependent variable was the total delivery time (Y) and the independent variables were the distance covered (X_1) and the packaging time (X_2). The delivery time data collected by the statistical analyst is given below:

Time (Y)	**Distance (X_1)**	**Packaging Time (X_2)**
18	61	30
14	95	25
17	72	30
14	84	25
13	98	10
24	53	35
13	68	15
22	54	40
12	89	30
19	73	20
$\sum Y_i = 166$	$\sum X_{1i} = 747$	$\sum X_{2i} = 260$

Estimate the parameters B_0, B_1, and B_2 by solving normal equations and find the estimated multiple linear regression equation.

E2) Use the matrix method to estimate parameters from the data given in E1).

11.4 TEST OF SIGNIFICANCE IN MULTIPLE REGRESSION

So far you have learnt how to estimate the parameters and fit the multiple regression model. You may now like to test the adequacy of the fitted model and examine whether the independent variables contribute significantly in explaining the variability in Y or not. For this purpose, we use the test of significance of equality of variances of the regressor variables.

If there is a linear relationship between the response variable Y and any of the independent variables X_1, X_2, …, X_p, we use the **test of significance of regression**. The test of significance of regression is a test to determine the linear relationship between the response variable and regressor variables and is often used to examine the adequacy of the model.

In order to test whether the contribution of independent variables $X_1,\ldots,X_p$ is significant or not, we test whether B_1, B_2, …, B_p are all zero in the model or at least one of them is not zero. This hypothesis can be written as:

$$H_0 : B_1 = B_2 = ... = B_p = 0$$

H_1: At least one of the regression coefficients is not zero

It can be tested by considering the following F-ratio:

$$F = \frac{SS_{Reg}/p}{SS_{Res}/(n-p-1)} \qquad \ldots (19)$$

In this test, the total sum of squares SS_T is partitioned into a sum of squares due to the contribution of regressor variables (SS_{Reg}) and a residual sum of squares (SS_{Res}). From equation (17), the residual sum of squares (SS_{Res}) is:

$$SS_{Res} = \sum Y_i^2 - \hat{B}_0(Y'X_0) - \hat{B}_1(Y'X_1) - ..... - \hat{B}_p(Y'X_p)$$

$$\text{or} \quad SS_{Res} = Y'Y - Y'X\hat{B} \qquad \ldots (20)$$

If B_1, B_2, …, B_k are all zeros, i.e., independent variables do not contribute to the variability in Y, then the total sum of squares, denoted by SS_T, is given as:

$$SS_T = \sum Y_i^2 - n\bar{Y}^2 = Y'Y - \frac{\left(\sum Y_i\right)^2}{n} \qquad \ldots (21)$$

This is the total variability present in Y around the mean $\overline{Y}$. We can rewrite equation (20) as

$$SS_{Res} = \sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n} - \left\{\sum BjY'Xj - \frac{\left(\sum Y_i\right)^2}{n}\right\}$$

that is, $\quad SS_{Res} = SS_T - SS_{Reg}$

Hence, the difference of $SS_T - SS_{Res}$ gives the contribution of independent variables X_1, X_2, …, X_p, in explaining the variability in Y, i.e.,

$$SS_T - SS_{Res} = \hat{B}_0(Y'X_0) + \hat{B}_1(Y'X_1) + ... + \hat{B}_p(Y'X_p) - n\overline{Y}^2$$

$$\Rightarrow SS_{Reg} = \left\{ Y'X\hat{B} - \frac{\left(\sum Y_i\right)^2}{n} \right\}$$

$$\text{or } SS_{Reg} = \sum \hat{B}_j Y'X_j - n\overline{Y}^2 \qquad \text{… (22)}$$

We now summarise these results in the following ANOVA Table:

ANOVA TABLE

Sources of Variation	**Degree of Freedom (d.f.)**	**Sum of Squares (S.S.)**	**Mean Sum of Squares**	**Variance Ratio**
Independent Variables (X_1, X_2, …, X_p)	p	$SS_{Reg} = \sum_{j=0}^{p} \hat{B}_j Y'X_j - n\,\overline{Y}^2$	$\frac{SS_{Reg}}{p}$	$F = \frac{SS_{Reg}/p}{SS_{Res}/(n-p-1)}$
Residuals (SS_{Res})	n–p–1	$SS_{Res} = Y'Y - \sum_{j=0}^{p} \hat{B}_j Y'X_j$	$\frac{SS_{Res}}{(n-p-1)}$	
Total	**n–1**	$Y'Y - n\overline{Y}^2$		

Under the null hypothesis, i.e., when $B_1 = B_2 = \ldots = B_p = 0$, F is distributed as Fisher's F-distribution with p and (n–p–1) degree of freedom, i.e.,

$$F \sim F_{p,(n-p-1)} \qquad \text{… (23)}$$

If the calculated F is less than the tabulated $F_{p,\,(n-p-1)}$ at α level of significance, then we conclude that the contribution of X_1, X_2, ..., X_p to the variability in Y is not significant. Thus, they have no contribution in prediction. It may be of further interest to examine whether any one coefficient (say B_j) corresponding to the independent variable X_j is different from zero, after accounting for other variables X_k (all $k \neq j$). This can be tested by considering the statistic t:

$$t = \frac{\hat{B}_j}{S.E.(\hat{B}_j)} \qquad \text{… (24)}$$

where S.E.($\hat{B}_j$) uses the estimated value of $\hat{\sigma}^2$ given in equation (18). Under the null hypothesis, i.e., $B_j = 0$, the proposed statistic t follows the Student's t-distribution with (n–p–1) d.f. Thus, if

$$|t| \le t_{\alpha/2,\,n-p-1} \qquad \text{… (25)}$$

we accept H_0. Otherwise, we reject it. If B_j is significantly different from zero, it contributes significantly to the variability in Y after taking into account the contribution of other variables. If B_j is not significantly different from zero, its contribution is not significant after accounting for other variables in the model.

Example 4: Using the data of Example 1 and the results of Example 2, construct the ANOVA table, apply a relevant test of hypothesis and interpret the results.

Solution: As per the data given in Example 1 and the results of Example 2, we have

$$SS_{Res} = 97.2155 \text{ and } \sum_{j=1}^{p} \hat{B}_j Y'X_j = 3702.7845$$

Using these values, we construct the ANOVA table as follows:

ANOVA TABLE

Sources of Variation	Degree of Freedom (d.f.)	Sum of Squares (S.S.)	Mean Sum of Squares	Variance Ratio
Independent Variables (X_1, X_2)	2	$SS_{Reg} = \sum_{j=0}^{p} \hat{B}_j Y'X_j - n\overline{Y}^2$ = 369.4512	$\frac{SS_{Reg}}{2}$ = 184.7256	$F = \frac{SS_{Reg}/p}{SS_{Res}/(n - p - 1)}$ =17.1015
Residuals (SS_{Res})	9	$SS_{Res} = Y'Y - \sum_{j=0}^{p} \hat{B}_j Y'X_j$ = 97.2155	$\frac{SS_{Res}}{(n-p-1)}$ =10.8017	
Total	**11**	$Y'Y - n\overline{Y}^2$ **= 466.6667**		

We have obtained the Variance Ratio F = 17.1012, whereas the tabulated value of $F_{2,\,9}$ at $\alpha = 0.05$ is 4.26. Hence, we reject H_0 and conclude that X_1 and X_2 contribute significantly to the variability. It may be of further interest to examine whether the coefficient B_j corresponding to independent variable X_j is different from zero, after accounting for other variables X_k (all $k \neq j$). This can be tested by considering the statistic t:

$$t = \frac{\hat{B}_j}{S.E.(\hat{B}_j)}$$

From the result of Example 2, we also have

$$\hat{B}_1 = 1.3002 \text{ and } \hat{B}_2 = 0.1356$$

The Variance-Covariance matrix is

$$V(\hat{B}) = \hat{\sigma}^2 (X'X)^{-1}$$

$$= 10.8017 \begin{pmatrix} 0.7139 & -0.0658 & -0.0028 \\ -0.0658 & 0.0095 & -0.0005 \\ -0.0028 & -0.0005 & 0.0002 \end{pmatrix}$$

Thus $$V(\hat{B}) = \begin{pmatrix} 7.7112 & -0.7105 & -0.0305 \\ -0.7105 & 0.1024 & -0.0049 \\ -0.0305 & -0.0049 & 0.0024 \end{pmatrix}$$

Using equation (15), we obtain

$$V(\hat{B}_0) = 7.7112, \quad V(\hat{B}_1) = 0.1024 \quad \text{and } V(\hat{B}_2) = 0.0024$$

and therefore,

$$S.E.(\hat{B}_0) = \sqrt{7.7112} = 2.7769$$

$$\text{S.E.}\left(\hat{B}_1\right) = \sqrt{0.1024} = 0.3199$$

$$\text{S.E.}\left(\hat{B}_2\right) = \sqrt{0.0024} = 0.0494$$

Therefore, the statistic t is given as:

$$t_0 = \frac{\hat{B}_0}{\text{S.E.}(\hat{B}_0)} = \frac{1.8765}{2.7769} = 0.6758$$

$$t_1 = \frac{\hat{B}_1}{\text{S.E.}(\hat{B}_1)} = \frac{1.3002}{0.3199} = 4.0642$$

$$t_2 = \frac{\hat{B}_2}{\text{S.E.}(\hat{B}_2)} = \frac{0.1356}{0.0494} = 2.7444$$

But the tabulated value of t-statistic for $\alpha = 0.05$ is

$$t_{0.025,22} = 2.262$$

Hence, both variables contribute significantly to the variability in Y.

You may now like to solve the following exercise.

E3) Make the ANOVA table, calculate standard errors of estimates and test their significance using the data in E1. Interpret the results.

11.5 COEFFICIENT OF DETERMINATION (R^2) AND ADJUSTED R^2

We define the coefficient of determination, R^2, in the same way as for simple regression. It gives a measure of adequacy of model fit.

We define R^2 as follows:

R^2 = Variability accounted by independent variables/Total variability around the mean

$$= \frac{\sum_{j=0}^{p} \hat{B}_j Y' X_j - n\bar{Y}^2}{Y'Y - n\bar{Y}^2} \qquad \ldots (26)$$

Its value always lies between 0 and 1. When the fit is good, $R^2 \sim 1$. Otherwise, $R^2 \sim 0$.

The value of R^2 always increases with p. The increase may be negligible, but R^2 never decreases. When we compare two models with different values of p, the model with larger p is preferable if R^2 corresponding to it is significantly larger than R^2 with smaller p. A model with smaller p with large R^2 is always preferable as it is a simple model. Hence, you should choose a model with small p if its R^2 is not much smaller than R^2 for a model with a larger p.

For this, we define an adjusted R^2, viz., R^2_{Adj}, which penalises R^2 when p increases but R^2 does not increase significantly. We know that

$$R^2 = \frac{SS_{Reg}}{SS_T} \qquad \ldots(27)$$

$$1 - R^2 = \frac{SS_T - SS_{Reg}}{SS_T} = \frac{SS_{Res}}{SS_T}$$

Then we define R^2_{Adj} as

$$1 - R^2_{Adj} = \frac{SS_{Res}/(n-p-1)}{SS_T/(n-1)} = \frac{(n-1)(1-R^2)}{(n-p-1)} \qquad \ldots(28)$$

Here, we have divided the numerator and denominator by their degree of freedom.

$SS_{Res}/(n-p-1)$ may decrease with increase in p even when there is no appreciable decrease in R^2. Hence,

$$R^2_{Adj} = 1 - \frac{(n-1)(1-R^2)}{(n-p-1)} \qquad \ldots(29)$$

Therefore, we should stop including the terms in the model if R^2_{Adj} starts decreasing. We prefer a model with larger R^2_{Adj} and smaller p than a model with smaller R^2_{Adj} but larger p.

Example 5: Using the data of Example 1 and the results of Examples 2 and 3, calculate R^2, R^2_{Adj} and interpret the results.

Solution: Using the data of Example 1 and the results of Examples 2 and 3, and on putting the values in equation (27), we get

$$R^2 = \frac{SS_{Reg}}{SS_T} = \frac{369.4512}{466.6667} = 0.7917$$

Therefore, the adjusted R^2 is obtained as follows:

$$R^2_{Adj} = 1 - \frac{(12-1)(1-0.7917)}{(12-2-1)} = 0.7454$$

From the coefficient of determination, R^2, we see that 79% variability in Y is due to X. This is quite a good fit. Adjusted R^2 is 0.7454, which is quite large. Hence we conclude that both X_1 and X_2 contribute adequately to the model fit.

You may now like to calculate R^2 and adjusted R^2 yourself. Try the following exercise.

E4) Calculate R^2 and adjusted R^2 and comment on the goodness of fit of the model, for the data given in E1.

11.6 REGRESSION WITH DUMMY VARIABLES

In previous sections, we have dealt with multiple linear regression when the independent / regressor variables are quantitative. The quantitative variables such as height, distance, temperature, time, income, pressure, etc. have a well defined scale of measurement. However, sometimes independent variables include qualitative variables such as sex (male/female), regions (north, south, east, west, etc.), religion such as Hindu, Muslim, Christian, etc. Such variables called **categorical variables** cannot be measured and hence no quantitative number can be assigned to them. We define dummy variables to account for the effect that the qualitative variables may have on the response variable. Dummy variables are also known as **indicator variable**s. Suppose, k is the number of levels a categorical variable takes. Then we define (k –1) dummy variables. For example, if we have two categories of male or female in the data, i.e., k = 2 and we define one dummy variable.

Suppose that a statistical analyst is analysing the vending machine's efficiency in the distribution of a product. She/he is interested in relating the time required to service the consumer with the distance travelled by the product in the vending machine for machines of two types, A and B.

The second regressor variable, machine type is qualitative, and has two levels: Type A and Type B. It allows us to code the types of machines used. Therefore, we define a dummy variable X_2 which takes on the values 0 and 1 to identify the types of machines as follows:

$$X_2 = \begin{cases} 1, & \text{if distribution is done by machine A} \\ 0, & \text{if distribution is done by machine B} \end{cases}$$

The variable X_2 is called an indicator variable because it is used to indicate the presence or absence of Machine A or B.

For such situations, we have a multiple linear regression model given by

$$Y = B_0 + B_1X_1 + B_2X_2 + e \qquad \ldots (30)$$

To determine the regression coefficients in this model, we first consider machine type A for which X_2 takes value 0. Then the regression model is given by:

$$Y = B_0 + B_1X_1 + B_2(0) + e$$

$$\Rightarrow \quad Y = B_0 + B_1X_1 + e \qquad \ldots (31)$$

The relationship between the response variable Y and regressor variable X_1, i.e., distance travelled by the product in the machine is a straight line with intercept B_0 and slope B_1.

For machine of type B, we have $X_2 = 1$. Then the regression model becomes

$$Y = B_0 + B_1X_1 + B_2(1) + e$$

$$\Rightarrow \quad Y = (B_0 + B_2) + B_1X_1 + e \qquad \ldots (32)$$

which shows that the relationship between Y and X_1 is also a straight line with slope B_1 but intercept $(B_0 + B_2)$.

Note that these models are linear with the same slope B_1 but different intercepts. Hence, these two models describe two parallel regression lines, i.e., two lines with a common slope and different intercepts. The vertical distance between these two lines is the difference in the intercepts, i.e., B_2. The two parallel regression lines formed by the above models given in equations (31) and (32) are shown Fig. 11.1.

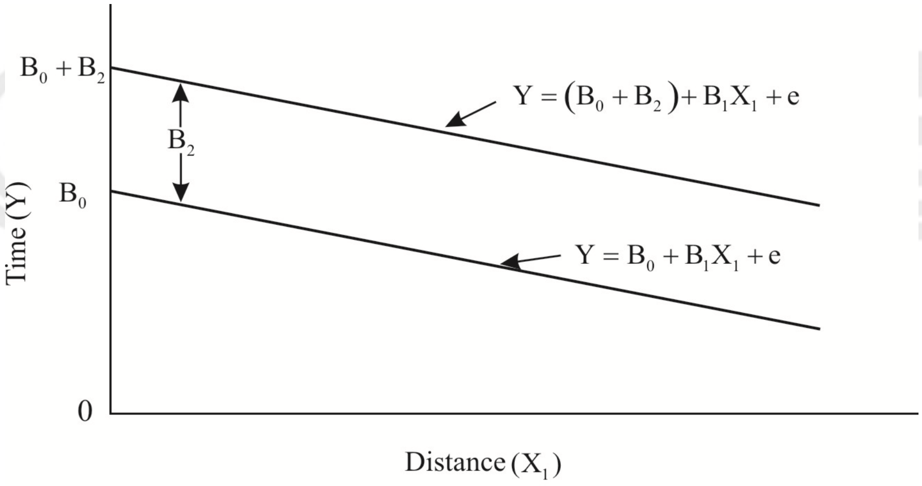

Fig. 11.1

For three Machine types A, B and C, two dummy variables X_2 and X_3 are used. The model becomes

$$Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + e \qquad \dots (33)$$

The levels of dummy variable would be:

X_2=0	X_3=0	For Machine Type A
X_2=1	X_3=0	For Machine Type B
X_2=0	X_3=1	For Machine Type C

In general, a categorical variable with k categories is denoted by (k – 1) dummy variables.

Let us try to understand regression analysis using dummy variables with the help of an example.

Example 6: A statistical analyst is analysing the performance of washing machines in the distribution system. He/she is interested in predicting the amount of time required by the driver to service washing machines of two types: i) Type A and ii) Type B. The data on the required time collected by the statistical analyst is given below:

Time (Y)	Distance (X_1)	Machine Type (X_2)
20	50	1
10	20	1
10	30	1
15	10	0
15	10	0
20	30	0
10	10	0
25	40	0
30	80	1
15	20	0
20	10	0
10	40	1

Check whether there is a linear relationship between Y (time) and the two independent variables X_1(distance) and X_2(type). Calculate the values of the coefficients and fit the regression equation.

Solution: Since two types of washing machines A and B have been used, k = 2. Here we have to define one dummy variable X_2, which takes two values:

$X_2 = 0$ if the observation is from machine A

$=1$ if the observation is from machine B

We form the following table from the given data to fit the regression equation:

Time (Y)	**Distance (X_1)**	**Machine Type (X_2)**	Y^2	$(X_1)^2$	$(X_2)^2$	X_1Y	X_2Y	X_1X_2
20	50	1	400	2500	1	1000	20	50
10	20	1	100	400	1	200	10	20
10	30	1	100	900	1	300	10	30
15	10	0	225	100	0	150	0	0
15	10	0	225	100	0	150	0	0
20	30	0	400	900	0	600	0	0
10	10	0	100	100	0	100	0	0
25	40	0	625	1600	0	1000	0	0
30	80	1	900	6400	1	2400	30	80
15	20	0	225	400	0	300	0	0
20	10	0	400	100	0	200	0	0
10	40	1	100	1600	1	400	10	40
$\sum Y_i$ =200	$\sum X_{1i}$ =350	$\sum X_{2i}$ = 05	$\sum Y_i^2 =$ 3800	$\sum X_{1i}^2$ = 15100	$\sum X_{2i}^2$ =5	$\sum X_{1i} Y_i$ = 6800	$\sum X_{2i} Y_i$ = 80	$\sum X_{1i} X_{2i}$ =220

The normal equations (5) for p = 2 and $X_{0i} = 1$ are:

$$n\,\hat{B}_0^{'} + \hat{B}_1 \sum X_{1i} + \hat{B}_2 \sum X_{2i} = \sum Y_i$$

$$\hat{B}_0 \sum X_{1i} + \hat{B}_1 \sum X_{1i}^2 + \hat{B}_2 \sum X_{1i} X_{2i} = \sum Y_i X_{1i}$$

$$\hat{B}_0 \sum X_{2i} + \hat{B}_1 \sum X_{1i} X_{2i} + \hat{B}_2 \sum X_{2i}^2 = \sum Y_i X_{2i}$$

On putting the values of the sums calculated in the above table, we get

$$12\ \hat{B}_0 + 350\ \hat{B}_1 + 05\hat{B}_2 = 200 \qquad \dots \text{(i)}$$

$$350\ \hat{B}_0 + 15100\,\hat{B}_1 + 220\,\hat{B}_2 = 6800 \qquad \dots \text{(ii)}$$

$$05\ \hat{B}_0 + 220\,\hat{B}_1 + 05\,\hat{B}_2 = 80 \qquad \dots \text{(iii)}$$

From equation (iii), we have

$$\hat{B}_0 = 16 - 44\ \hat{B}_1 - \hat{B}_2 \qquad \dots \text{(iv)}$$

On putting the value of $\hat{B}_0$ in equations (i) and (ii) and simplifying, we get

$$-178\hat{B}_1 - 7\hat{B}_2 = 8 \qquad \text{… (v)}$$

$$-300\hat{B}_1 - 130\hat{B}_2 = 1200 \qquad \text{… (vi)}$$

On solving equations (v) and (vi), we get

$$\hat{B}_1 = 0.3498, \quad \hat{B}_2 = \text{-}\ 10.038$$

and $\quad \hat{B}_0 = 16 - 44\ \hat{B}_1 - \hat{B}_2 = 10.646$

Hence, the fitted regression equation is

$$Y = 10.646 + 0.3498\ X_1 - 10.038\ X_2 \qquad \text{… (vii)}$$

We conclude that there is a linear relationship between Y (time in seconds) and the two independent variables X_1 (distance) and X_2 (type of machine). Since the regression coefficient for the variable X_2 is negative, it affects the delivery time. The numerical value of the regression coefficient associated with X_2 is higher than that of the other regressor variable. It shows that distance travelled (in m) affects the delivery time less than the type of machines.

To determine the regression coefficients in this model for each type of machine, we first consider machine A for which X_2 takes value 0. We put the values of regression coefficients in equation (28). Then the regression model becomes

$$Y = 10.646 + 0.3498 X_1 \qquad \text{… (viii)}$$

For machine B, we put the value of the regression coefficient and X_2= 1. Then the regression model becomes

$$Y = (0.606) + 0.3498 X_1 \qquad \text{… (ix)}$$

Note that as discussed in Sec 11.5, these estimated regression lines have the same slope, i.e., 0.3498, but have different intercepts, i.e., 10.646 and 0.606.

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

E5) Using the data given in the following table, find the regression coefficients and obtain the estimated regression equations for the model given in equations (27), (28) and (29) :

Time (hour) Y	**Distance (feet) X_1**	**Machine Type**	**X_2**
18	61	A	0
14	95	A	0
17	72	A	0
14	84	A	0
13	98	A	0
24	53	B	1
13	68	B	1
22	54	B	1
12	89	B	1
19	73	B	1

Check whether there is a linear relationship between Y (time) and the two independent variables X_1 (distance) and X_2 (machine type). Calculate the values of the coefficients and fit the regression equation.

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

11.7 SUMMARY

1. The basic concept of multiple linear regression is the same as that of simple regression. However, instead of one independent variable, there are several independent variables, say, $X_1, X_2, X_3, \ldots, X_p$.

2. A **multiple regression model** is given by

$$Y = B_0 + B_1X_1 + B_2X_2 + \ldots + B_pX_p + e$$

where Y is the dependent variable and $X_1, X_2, \ldots, X_p$ are p independent variables. This is called the multiple linear regression model with p independent/regressor variables. The term linear is used because the dependent/response variable Y is a linear function of the unknown parameters $B_0, B_1, B_2, \ldots, B_p$.

3. The simple regression model considered in Unit 9 becomes a particular case of this model with $X_0 = 1$, $B_0 = a$, $B_1 = b$ and $B_i = 0$ $(i \geq 2)$. The interpretation of coefficients B_j $(j = 1, 2, \ldots, p)$ is that B_j represents the amount of change in Y for a unit change in X_j, keeping the other independent variables X_ks $(k \neq j)$ fixed. These coefficients are known as **partial regression coefficients** as the effect of one independent variable is studied on the dependent variable while the other variables are held fixed or constant. We use the term multiple linear regression for this model because several variables are included in the regression and the parameters $B_0, B_1, \ldots, B_p$ appear in a linear form.

4. We estimate the parameters of a multiple linear regression equation using the method of least squares. In this method, we minimise the total error term, so that the sum of the squares of the differences between the observed values Y_i and its expected values is minimum, i.e., the sum of squares of the error terms is minimum.

   When p is greater than 2, it is more convenient to write the normal equations in matrix form. The regression equations in matrix notation can be written as Y = X B + e, where Y is an n × 1 vector of the observed values of the response variable Y, X is a (p + 1) × n matrix of the values of regressor variables, B is a (p + 1) × 1 vector of regression coefficients and e is an n × 1 vector of random errors. In matrix notation, the (p + 1) normal equations can be written as

$$X'X\hat{B} = X'Y$$

5. The variance-covariance matrix of $\hat{B}$ is given by

$$V(\hat{B}) = \sigma^2 (X'X)^{-1}$$

where $V(\hat{B}) = \sigma^2 (X'X)^{-1}$ is a (p + 1) × (p + 1) matrix and its diagonal elements give the variances of coefficients and off diagonal elements give the covariances. If we use the notation

$$V(\hat{B}) = \sigma^2 (X'X)^{-1} = (\sigma_{jk}), \qquad j, k = 0, 1, \ldots, p.$$

we can write

$$V(\hat{B}_j) = \sigma_{jj}, \quad \text{and} \quad Cov(\hat{B}_j, \hat{B}_k) = \sigma_{jk}$$

The standard error of $\hat{B}_j$ is given by

$$S.E.\left(\hat{B}_j\right) = \sqrt{\sigma_{jj}}$$

6. If there is a linear relationship between the response variable Y and any of the independent variables X_1, X_2, …, X_p, we use the **test of significance of regression**. The test of significance of regression is a test to determine the linear relationship between the response variable and regressor variables and is often used to examine the adequacy of the model.

7. The **coefficient of determination**, R^2 and adjusted R^2 are measures of goodness of fit of the multiple regression model. The value of R^2 always increases with p. The increase may be negligible, but R^2 never decreases. When we compare two models with different values of p, the model with larger p is preferable if R^2 corresponding to it is significantly larger than R^2 with smaller p. A model with smaller p with large R^2 is always preferable as it is a simple model. Hence, one should choose a model with small p if its R^2 is not much smaller than R^2 for a model with a larger p.

8. We define dummy variables to account for the effect that qualitative variables may have on the response variable. Dummy variables are also known as **categorical** as **indicator variables**. Suppose, k represents the number of levels a categorical variable takes, then we define (k – 1) dummy variables. For example, if we have two categories, male or female, in the data, k = 2 and we define one dummy variable.

11.8 SOLUTIONS/ANSWERS

E1) We do the following calculations for the given data:

Time (Y)	**Distance (X_1)**	**Vending Time (X_2)**	**Y^2**	**$(X_1)^2$**	**$(X_2)^2$**	**X_1Y**	**X_2Y**	**X_1X_2**
18	61	30	324	3721	900	1098	540	1830
14	95	25	196	9025	625	1330	350	2375
17	72	30	289	5184	900	1224	510	2160
14	84	25	196	7056	625	1176	350	2100
13	98	10	169	9604	100	1274	130	980
24	53	35	576	2809	1225	1272	840	1855
13	68	15	169	4624	225	884	195	1020
22	54	40	484	2916	1600	1188	880	2160
12	89	30	144	7921	900	1068	360	2670
19	73	20	361	5329	400	1387	380	1460
$\sum Y_i$ = 166	$\sum X_{1i}$ = 747	$\sum X_{2i}$ = 260	$\sum Y_i^2$ = 2908	$\sum X_{1i}^2$ = 58189	$\sum X_{2i}^2$ = 7500	$\sum X_{1i}Y_i$ = 11901	$\sum X_{2i}Y_i$ = 4535	$\sum X_{1i}X_{2i}$ = 18610

From the above table, putting the values of $\sum Y_i$, $\sum X_{1i}$, $\sum X_{2i}$, $\sum X_{1i}^2$, $\sum X_{2i}^2$, $\sum X_{2i} Y_i$, $\sum X_{1i} Y_i$ and $\sum X_{1i} X_{2i}$ in normal equations, we get

$$10\,\hat{B}_0 + 747\,\hat{B}_1 + 260\,\hat{B}_2 = 166 \qquad \text{… (i)}$$

$$747\,\hat{B}_0 + 58189\,\hat{B}_1 + 18610\,\hat{B}_2 = 11901 \qquad \text{… (ii)}$$

$$260\,\hat{B}_0 + 18610\,\hat{B}_1 + 7500\,\hat{B}_2 = 4535 \qquad \text{… (iii)}$$

Solving these equations, we get

$$\hat{B}_0 = 26.7569,\ \ \hat{B}_1 = -0.1729 \ \text{ and } \ \hat{B}_2 = 0.1062$$

The fitted regression equation is:

$$Y = 26.7569 - 0.1729\,X_1 + 0.1062\,X_2$$

E2) Using the matrix notation, we have from the data:

$$Y'_{1\times10} = [18, 14, 17, 14, 13, 24, 13, 22, 12, 19]$$

$$X'X = \begin{pmatrix} 10 & 747 & 260 \\ 747 & 58189 & 18610 \\ 260 & 18610 & 7500 \end{pmatrix}, \quad X'Y = \begin{pmatrix} 166 \\ 11901 \\ 4535 \end{pmatrix}$$

$$(X'X)^{-1} = \begin{pmatrix} 8.1316 & -0.0690 & -0.1108 \\ -0.0690 & 0.0007 & 0.0007 \\ -0.1108 & 0.0007 & 0.0022 \end{pmatrix}$$

and $$\begin{pmatrix} \hat{B}_0 \\ \hat{B}_1 \\ \hat{B}_2 \end{pmatrix} = (X'X)^{-1} X'Y = \begin{pmatrix} 26.7569 \\ -0.1729 \\ 0.1062 \end{pmatrix}$$

Hence, the fitted equation is

$$Y = 26.7569 - 0.1729X_1 + 0.1062X_2$$

We now calculate the value of residual sum of squares to obtain an estimate of $\hat{\sigma}^2$ as follows:

$$SS_{Res} = Y'Y - Y'X\hat{B}$$

$$= 2908 - 166\times(26.7569) - 11901\times(-0.1729) - 4535\times(0.1062)$$

$$= 2908 - 4441.6454 + 2057.6829 - 481.617 = 42.4205$$

Therefore, on putting the value of SS_{Res} in equation (18), we get

$$\hat{\sigma}^2 = 42.4205/(10-3) = 6.06$$

E3) Using the data of E1 and the results of E2, we get

$$\hat{B}_0 = 26.7569,\ \ \hat{B}_1 = -0.1729 \ \text{ and } \ \hat{B}_2 = 0.1062$$

As per the data given in Example 1 and the result of Example 2, we have

$$SS_{Res} = 42.4205 \text{ and } \sum_{j=1}^{p} \hat{B}_j Y'X_j = 2865.5795$$

Using these values we construct the ANOVA table as follows:

ANOVA TABLE

Sources of Variation	Degree of Freedom (d.f.)	Sum of Squares (S.S.)	Mean Sum of Squares	Variance Ratio
Independent Variables (X_1, X_2)	2	$SS_{Reg} = \sum_{j=1}^{p} \hat{B}_j Y'X_j - n\bar{Y}^2$ = 109.97956	$\frac{SS_{Reg}}{2} =$ 54.9897	$F = \frac{SS_{Reg}/p}{SS_{Res}/(n-p-1)}$ = 9.074
Residuals (SS_{Res})	7	$SS_{Res} = Y'Y - \sum_{j=1}^{p} \hat{B}_j Y'X_j$ = 42.4205	$\frac{SS_{Res}}{(n-p-1)}$ = 6.06	
Total	**9**	$Y'Y - n\bar{Y}^2 = 152.4$		

The calculated value of Variance Ratio F = 9.074, whereas the tabulated value of $F_{2,22}$ at $\alpha = 0.05$ is 3.44. Hence, we reject H_0 and conclude that X_1 and X_2 contribute significantly in explaining the variability.

It may be of further interest to examine whether the coefficient B_j, corresponding to independent variable X_j, is different from zero, after accounting for other variables X_k (all $k \neq j$). This can be tested by considering statistic t:

$$t = \frac{\hat{B}_j}{S.E.(\hat{B}_j)}$$

From the result of Example 2, we have

$$\hat{B}_1 = -0.1729 \text{ and } \hat{B}_2 = 0.1062$$

The Variance-Covariance matrix is

$$V\left(\hat{B}\right) = \hat{\sigma}^2 \left(X'X\right)^{-1}$$

Thus $$V(\hat{B}) = \begin{pmatrix} 49.7393 & -0.4218 & -0.6777 \\ -0.4218 & 0.0041 & 0.0045 \\ -0.6777 & 0.0045 & 0.0132 \end{pmatrix}$$

Using equation (15), we obtain

$$V(\hat{B}_0) = 49.7393, \quad V(\hat{B}_1) = 0.0041 \quad \text{and } V(\hat{B}_2) = 0.0132$$

and therefore,

$$\text{S.E.}\left(\hat{B}_0\right) = \sqrt{49.7393} = 7.0526$$

$$\text{S.E.}\left(\hat{B}_1\right) = \sqrt{0.0041} = 0.064$$

$$\text{S.E.}\left(\hat{B}_2\right) = \sqrt{0.0132} = 0.11489$$

Therefore, the statistic t is given as:

$$t_0 = \frac{\hat{B}_0}{\text{S.E.}(\hat{B}_0)} = \frac{26.7569}{7.0526} = 3.6521$$

$$t_1 = \frac{\hat{B}_1}{\text{S.E.}(\hat{B}_1)} = \frac{-0.1729}{0.064} = -2.7014$$

$$t_2 = \frac{\hat{B}_2}{\text{S.E.}(\hat{B}_2)} = \frac{0.1062}{0.11489} = 0.924$$

But the tabulated value of t-statistic for $\alpha = 0.05$ is

$$t_{0.025,7} = 2.37$$

Hence, variable X_1 contributes significantly in explaining the variability in Y but the variable X_2 does not.

As far as the interpretation of coefficients is concerned, there is an increase of 0.1062 seconds in time for one unit change in cases (X_1). Similarly, for one unit increase in X_2, there is a 0.1729 seconds decrease in time.

E4) Using the data of E1) and the results of E2) and E3) , we get

R^2 = Sum of Squares due to X_1, X_2 /Total Sum of Squares

= 109.58/152.4 = 0.719

$$R^2_{Adj} = 1 - \frac{(n - 1)(1 - R^2)}{(n - p - 1)}$$

$$= 1 - \frac{9´ \left(1 - 0.719\right)}{7} = 0.6387$$

R^2 indicates that only 72% of variability in Y is explained by X_1 and X_2.

E5) Two types of washing machines A and B have been used. Hence, k = 2. Here we have to define one dummy variable X_2, which takes two values:

X_2= 0 if the observation is from machine type 'A'

=1 if the observation is from machine type 'B'

From the given data, we form the following table to find and fit the regression equation:

Time **Y**	**Distance** **(X_1)**	**(X_2)**	**Y^2**	**$(X_1)^2$**	**$(X_2)^2$**	**X_1Y**	**X_2Y**	**X_1X_2**
18	61	0	324	3721	0	1098	0	0
14	95	0	196	9025	0	1330	0	0
17	72	0	289	5184	0	1224	0	0
14	84	0	196	7056	0	1176	0	0
13	98	0	169	9604	0	1274	0	0
24	53	1	576	2809	1	1272	24	53
13	68	1	169	4624	1	884	13	68
22	54	1	484	2916	1	1188	22	54
12	89	1	144	7921	1	1068	12	89
19	73	1	361	5329	1	1387	19	73
$\sum Y_i$ =166	$\sum X_{1i}$ =747	$\sum X_{2i}$ = 05	$\sum Y_i^2 =$ 2908	$\sum X_{1i}^2$ = 58189	$\sum X_{2i}^2$ =05	$\sum X_{1i}Y_i$ = 11901	$\sum X_{2i}Y_i$ = 90	$\sum X_{1i}X_{2i}$ =337

From the above table, putting the values in the normal equations (5) for p = 2 and noting that $X_0 = 1$, we get

$$10\ \hat{B}_0 + 747\ \hat{B}_1 + 05\hat{B}_2 = 166 \qquad \ldots \text{(i)}$$

$$747\ \hat{B}_0 + 58189\hat{B}_1 + 337\hat{B}_2 = 11901 \qquad \ldots \text{(ii)}$$

$$05\ \hat{B}_0 + 337\hat{B}_1 + 05\hat{B}_2 = 90 \qquad \ldots \text{(iii)}$$

From equation (iii), we have

$$\hat{B}_0 = \frac{90 - 337\ \hat{B}_1 - 05\hat{B}_2}{05} \qquad \ldots \text{(iv)}$$

On putting the value of $\hat{B}_0$ in equations (i) and (ii) and simplifying, we get

$$365\hat{B}_1 - 25\hat{B}_2 = -70 \qquad \ldots \text{(v)}$$

$$39206\hat{B}_1 - 2050\hat{B}_2 = -7135 \qquad \ldots \text{(vi)}$$

On solving equations (v) and (vi), we get

$$\hat{B}_1 = -0.214, \quad \hat{B}_2 = -\ 0.3244$$

and $$\hat{B}_0 = \frac{90 - 337\ (-0.214) - 05(-0.3244)}{05} = 32.748$$

Hence the fitted equation for the model given in equation (27) is

$$Y = 32.748 - 0.214X_1 - 0.3244\ X_2 \qquad \ldots \text{(vii)}$$

Now we can conclude that there is a linear relationship between Y (time) and the two independent variables X_1 (distance) and X_2 (type of machines). As the regression coefficient for the variable X_1 is negative, it affects the delivery time.

To determine whether the regression coefficients in this model are correct, we first consider machine A for which X_2 takes value 0. We put the values of regression coefficients in equation (28). Then the regression model becomes

$$Y = 32.748 - 0.214 X_1 \qquad \ldots \text{(viii)}$$

For machine B, we put the value of regression coefficients and $X_2 = 1$. Then the regression model becomes

$$Y = 31.126 - 0.214 X_1 \qquad \ldots \text{(ix)}$$

Note that as discussed in Sec 11.5, these estimated regression lines have the same slope, i.e., – 0.214, but different intercepts, i.e., 32.748 and 31.126.