UNIT 7 MULTIPLE LINEAR REGRESSION MODEL: ESTIMATION*

Structure

7.0 OBJECTIVES

After going through this unit, you will be able to:

- specify the multiple regression model involving more than one explanatory variable;
- estimate the parameters of the multiple regression model by the OLS method stating their properties;
- interpret the results of an estimated multiple regression model;
- indicate the advantage of using matrix notations in multiple regression models;
- explain the maximum likelihood method of estimation showing that the 'maximum likelihood estimate (MLE)' and the OLS estimate are asymptotically similar;
- derive the expression for the coefficient of determination (R^2) for the case of a simple multiple regression model with two explanatory variables; and
- distinguish between R^2 and adjusted R^2 specifying why adjusted R^2 is preferred in practice.

7.1 INTRODUCTION

By now you are familiar with the simple regression model where there is one dependent variable and one independent variable. The dependent variable is explained by the independent variable. Now let us discuss about the multiple regression model. In a multiple regression model, there is one dependent variable

* Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi

and more than one independent variable. The simplest possible multiple regression model is a three-variable regression model, with one dependent variable and two explanatory variables. Such a three-variable multiple regression equation or model is expressed as follows:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \qquad \ldots (7.1)$$

Throughout this unit, we shall be mostly dealing with a multiple regression model as specified in equation (7.1) above. Here, Y is the dependent variable and X_2 and X_3 are independent variables. u_i is the stochastic error term. The interpretation of this error term is the same as in the simple regression model. You may wonder as to why there is no X_1 in equation (7.1). The answer is that X_1 is implicitly taken as 1 for all observations. In the above equation, the parameter β_1 is the intercept term. We can think of Y, X_2 and X3 as some variables from economic theory. We may treat it as a demand function, where Y stands for quantity demanded of a good, and X_2 and X_3 are price of that good and the consumer's income, respectively. As another example, we can think of a production/demand function with two inputs. Here Y is the quantity produced or demanded of a good, and X_2 is the labour input, and X_3 the capital input. You can think of many similar examples.

7.2 ASSUMPTIONS OF MULTIPLE REGRESSION MODEL

Recall that the simple regression model is based on certain assumptions. These assumptions are the benchmark for a regression model. When these assumptions are fulfilled by a regression model, we call it as the classical linear regression model (CLRM). The main assumptions for the classical multiple regression models remain the same as the simple regression model. There is one change. This relates to a new assumption on multicollinearity. Since we are considering more than one independent variable X_i, it is now necessary to assume that the X_i's are not perfectly correlated. Let us recapitulate the assumptions of the CLRM with this new assumption added as follows:

(i) The regression model is linear in parameters. This assumption implies that the dependent variable is a linear function of the parameters, *βs*. The regression model could be non-linear in explanatory variables.

(ii) There is no covariance between u_i and X_i variables. This implies, in a multiple regression model like that in equation (7.1), there is no correlation between the error term and explanatory variables. That is:

$$Cov(u_i, X_{2i}) = Cov(u_i, X_{3i}) = 0 \qquad \ldots (7.2)$$

In order to avoid this problem, we assume that all explanatory variables are non-stochastic in nature. This implies that the values taken by the explanatory variables X are considered fixed in repeated samples.

(iii) The mean of the error terms is zero. In other words, the expected value of the error term conditional upon the explanatory variables X_{2i} and X_{3i} is zero. This means:

$$E(u_i) = 0 \text{ or } E(u_i|X_{2i}, X_{3i}) = 0 \quad \ldots (7.3)$$

(iv) No autocorrelation: This assumption means that there is no serial correlation or autocorrelation between the error terms of the individual observations. This implies that the covariance between the error term associated with the i^{th} observation u_i and that with the j^{th} observation u_j is zero. In notations, this means:

$$\text{cov}(u_i, u_j) = 0 \quad \ldots (7.4)$$

(v) Homoscedasticity: The assumption of homoscedasticity implies that the error variance is constant for all observations. This means:

$$var(u_i^2) = \sigma^2 \quad \ldots (7.5)$$

(vi) No exact collinearity between the X variables. This is the new additional assumption made for multiple regression models. This implies that there is no exact linear relationship between X_2 and X_3 This is referred to as the assumption of no perfect multicollinearity.

(vii) The number of observations n must be greater than the number of parameters to be estimated. In other words, the number of observations n must be greater than the number of explanatory variables k.

(viii) No specification bias: It is assumed that the model is correctly specified. The assumption of no specification bias implies that there are no errors involved while specifying the model. This means that both the errors of including an irrelevant variable and not including a relevant variable are taken care of while specifying the regression model.

(ix) There is no measurement error, i.e., X's and Y are correctly measured.

7.2.1 Interpretation of the Model

In the multiple regression model as in equation (7.1), the intercept β_1 measures the expected value of the dependent variable Y, when the values of explanatory variables X_2 and X_3 are zero. The other two parameters, β_2 and β_3, are the partial regression coefficients. Let us know more about these coefficients. The regression coefficients β_2 and β_3 are also known as the partial slope coefficients. β_2 measures the change in the mean value of Y [i.e., E(Y)] per unit change in X_2, holding the value of X_3 constant. This means: $\beta_2 = \frac{\Delta E(Y)}{\Delta X_2}$. It gives the 'direct' or the 'net' effect of a unit change in X_2 on the mean value of Y holding the effect of X_3 constant. Likewise, β_3 measures the change in the mean value of Y, per unit change in X_3, holding the value of X_2 constant. Like β_2, β_3 is given by: $\beta_3 = \frac{\Delta E(Y)}{\Delta X_3}$. Thus, the slope coefficients of multiple regression measures the impact of one explanatory variable on the dependent variable keeping the effect of the other variables fixed.

7.3 ESTIMATION OF MULTIPLE REGRESSION MODEL

The multiple regression equation is estimated to describe the Population Regression Function (PRF): $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$. This function consists of two components. The first is the deterministic component given by E ($Y_i \mid X_{2i}, X_{3i}$). This is also referred to as the Population Regression Line. The second component is the random component given by u_i. The PRF is estimated by using the sample. The estimated function (i.e., the sample regression function) is indicated by: $Y_i = b_1 + b_2X_{2i} + b_3X_{3i} + e_i$. Recall that $Y_i = \hat{Y}_i + e_i$ where $\hat{Y}_i$ is the estimated value of Y_i given by $E(Y_i|X_{2i}, X_{3i})$ and e_i is the residual term. In the sample regression function, b_1 is the estimator of population intercept β_1 and b_2 and b_3 are the estimators of population partial slope coefficient β_2 and β_3 respectively. The residual e_i is the estimator of population error term u_i. We know that the sample regression line is obtained in the OLS method by minimizing the residual sum of squares as follows:

$$Min \sum e_i^2 = \sum\left(Y_i - \hat{Y}_i\right)^2$$

$$= \sum(Y_i - b_1 - b_2X_{2i} - b_3X_{3i})^2 \text{ [since } \hat{Y}_i = b_1 + b_2X_{2i} + b_3X_{3i}]$$

We now consider the three first order conditions, i.e., $\frac{\partial \sum e_i^2}{\partial b_1} = 0$, $\frac{\partial \sum e_i^2}{\partial b_2} = 0$ and $\frac{\partial \sum e_i^2}{\partial b_3} = 0$. From these three partial derivatives, we obtain the estimators as:

(i) $b_1 = \bar{Y} - b_2\bar{X}_2 - b_3\bar{X}_3$

(ii) $b_2 = \frac{(\sum y_i.x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i}x_{3i})}{(\sum X_{2i}^2)(\sum X_{3i}^2) - (\sum X_{2i}X_{3i})^2}$

(iii) $b_3 = \frac{(\sum y_i.x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i}x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2}$

The corresponding variances and standard errors of the parameters are given as:

$$V(b_1) = \left[\frac{1}{n} + \frac{\bar{X}_2^2\sum x_{3i}^2 + \bar{X}_3^2\sum x_{2i}^2 - 2\bar{X}_2\bar{X}_3\sum x_{2i}x}{\sum x_{2i}^2\sum x_{3i}^2 - (\sum x_{2i}x_{3i})^2}\right]\sigma^2$$

$$SE(b_1) = +\sqrt{V(b_1)}$$

$$V(b_2) = \frac{\sum x_{3i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2} \times \sigma^2$$

$$SE(b_2) = +\sqrt{V(b_2)}$$

$$V(b_3) = \frac{\sum x_{2i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})} \times \sigma^2$$

$V(b_3) = \frac{\sigma^2}{\sum x_{3i}^2(1-r_{23}^2)}$

$SE(b_3) = +\sqrt{V(b_3)}$

You should note further that:

(i) $COV(b_2, b_3) = \frac{-r_{23}\sigma^2}{(1-r_{23}^2)\sqrt{x_{2i}^2}\sqrt{x_{3i}^2}}$

and the estimates of error variance and the partial correlation coefficients are given by:

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k} = \frac{RSS}{n-k} \qquad \dots (7.6)$$

For a regression model with 3 explanatory variables (such as equation (7.1)) we have $\hat{\sigma}^2 = \frac{RSS}{n-3}$.

$$r_{23} = \frac{(\sum x_{2i}x_{3i})^2}{\sum x_{2i}^2 \sum x_{3i}^2} \qquad \dots (7.7)$$

Note that in the above expressions, lower case letters represent deviations from the mean. We know that, since we are considering the 'classical' linear multiple regression model, the OLS estimators of the intercept and the partial slope coefficients satisfy the following properties:

a) The regression line passes through the means, $\bar{Y}, \bar{X}_2$ and $\bar{X}_3$. In a *k*-variable linear regression model, there is one regressand Y_i and $(k - 1)$ regressors since one of the coefficients is the intercept term β_1. Hence, the estimate of this intercept term is obtained as: $b_1 = \bar{Y} - b_2\bar{X}_2 - b_3\bar{X}_3$.

b) The mean value of the estimated $\widehat{Yi}$ is equal to the mean value of actual Y_i, i.e., $\bar{\hat{Y}}_i = \bar{Y}$.

c) $\frac{1}{n}\sum e_i = \bar{e}_i = 0$.

d) $Cov(e_i, X_{2i}) = Cov(e_i, X_{3i}) = 0$. That is, the residual e_i is uncorrelated with X_{2i} and X_{3i}. In other words: $(\sum e_i X_{2i}) = (\sum e_i X_{3i}) = 0$.

e) $Cov(e_i, \hat{Y}_i) = 0$, i.e., residual e_i is uncorrelated with $\hat{Y}_i$ and $\sum e_i \hat{Y}_i = 0$.

f) As r_{23}, the correlation coefficient between X_2 and X_3, increases towards 1, the variances of b_2 and b_3 increases for given values of $\sigma^2, \sum x_{2i}^2$ or $\sum x_{3i}^2$.

g) In view of f) above, given the values of r_{23} and $\sum x_{2i}^2$ or $\sum x_{3i}^2$ the variances of OLS estimators are directly proportional to σ^2.

h) Given the assumptions of CLRM, OLS estimators of partial regression coefficients are not only linear and unbiased but also have minimum variances in the class of all unbiased estimators, i.e., they are BLUE. In other words, they satisfy the Gauss-Markov theorem.

7.4 MAXIMUM LIKELIHOOD METHOD OF ESTIMATION

The method of 'maximum likelihood estimation' estimates the parameters of a probability distribution function (pdf). This is done by maximizing the likelihood function of the pdf. Hence, the estimators that maximize the likelihood function are called the 'maximum likelihood estimators'. To understand this concept better, let us derive the maximum likelihood estimators $(\tilde{\beta})$. We have used the notation $\tilde{\beta}$ to distinguish the ML estimators from the OLS estimators $(\hat{\beta})$. Let us assume that the pdf follows normal distribution. Thus, $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$. Taking log of the likelihood function of this pdf on its both sides, we get:

$$lnL = -\frac{n}{2}ln\sigma^2 - \frac{n}{2}ln(2\pi) - \frac{1}{2}\sum\frac{(Y_i - \beta_1 - \beta_2 X_{2i} - \beta_k X_{ki})^2}{\sigma^2}$$

Differentiating the above function partially with respect to $\beta_1, \beta_2, ..., \beta_k$ and σ^2 we obtain the following $(k + 1)$ equations:

$$\frac{\partial \ln L}{\partial \beta_1} = \frac{1}{\sigma^2}\sum(Y_i - \beta_1 - \beta_2 X_{2i} - \ldots - \beta_k X_{ki})(-1) \quad (1)$$

$$\frac{\partial \ln L}{\partial \beta_2} = \frac{1}{\sigma^2}\sum(Y_i - \beta_1 - \beta_2 X_{2i} - \ldots - \beta_k X_{ki})(-X_{2i}) \quad (2)$$

……………………………………………………………………………

$$\frac{\partial \ln L}{\partial \beta_k} = \frac{1}{\sigma^2}\sum(Y_i - \beta_1 - \beta_2 X_{2i} - \ldots - \beta_k X_{ki})(-X_{ki}) \quad (k)$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^2}\sum(Y_i - \beta_1 - \beta_2 X_{2i} - \ldots - \beta_k X_{ki})^2 \quad (k+1)$$

Setting these equations to zero (i.e., applying the first-order condition for optimization), and re-arranging terms, and denoting by $\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_k$ and $\tilde{\sigma}^2$ as the 'maximum likelihood estimates (MLEs)', we get:

$$\sum Y_i = n\tilde{\beta}_1 + \tilde{\beta}_2\sum X_{2i} + \cdots + \tilde{\beta}_k\sum X_{ki}$$

$$\sum Y_i X_{2i} = \tilde{\beta}_1 + \sum X_{2i} + \tilde{\beta}_2\sum X_{2i}^2 + \ldots + \tilde{\beta}_k\sum X_{2i}X_{ki}$$

……………………………………………………………

$$\sum Y_i X_{ki} = \tilde{\beta}_1\sum X_{ki} + \tilde{\beta}_2\sum X_{2i}X_{ki} + \ldots + \tilde{\beta}_k\sum X_{ki}^2$$

The above equations are precisely the normal equations of the OLS method of estimation. Therefore, the MLEs of the $\tilde{\beta}'s$ are the same as the OLS estimates of the $\tilde{\beta}'s$. Thus, substituting the MLEs (or the OLS estimators) into the $(K+1)^{st}$ equation above, and simplifying, we obtain the MLEs of σ^2 as

$$\tilde{\sigma}^2 = \frac{1}{n}\sum\left(Y_i - \tilde{\beta}_1 - \tilde{\beta}_2 X_{2i} - \ldots - \tilde{\beta}_k X_{ki}\right)^2$$

$$= \frac{1}{n}\sum \hat{u}_i^2$$

You may note that this estimator differs from the OLS estimator $\hat{\sigma}^2 = \sum u_i^2/(n-k)$. Since the latter is an unbiased estimator of σ^2, the MLE of $\tilde{\sigma}^2$ is a biased estimator. However, you should note that, asymptotically, $\tilde{\sigma}^2$ is also unbiased. This means, asymptotically, the estimates of MLE and OLS are similar. Further, the MLE estimator is biased but it is consistent.

For multiple regression models, the above algebraic expressions become unwieldy. Hence, we can take recourse to matrix algebra (on which you have studied in your earlier course BECC 104) to depict the multiple regression model. For this, let:

$$X_0 = \begin{bmatrix} 1 \\ X_{02} \\ X_{03} \\ \cdot \\ \cdot \\ \cdot \\ X_{0k} \end{bmatrix} \quad \ldots (7.8)$$

be the vector of values of the X variables for which we wish to predict $\hat{Y}_0$ the mean prediction of Y. Now the estimated multiple regression equation in the scalar form is:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \ldots + \hat{\beta}_k X_{ki} + u_i \quad \ldots (7.9)$$

In matrix notation (7.9) can be written compactly as:

$$\hat{Y}_i = x'_i \hat{\beta} \quad \ldots (7.10)$$

where $x'_i = [1 \quad X_{2i} \quad X_{3i} \ldots \quad X_{ki}]$ and

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \cdot \\ \cdot \\ \cdot \\ \hat{\beta}_k \end{bmatrix}$$

Equation (7.9) or (7.10) is the mean prediction of Y_i corresponding to given x'_i. Hence, if x'_i is as given in (7.8), (7.10) becomes

$$\left(\hat{Y}_I | x'_0\right) = x'_0 \hat{\beta} \quad \dots (7.11)$$

where, the values of x_0 are specified. Note that (7. 11) gives an unbiased prediction of $E(Y_i|x'_0)$, since $E\left(x'_0\hat{\beta}\right) = x'_0\hat{\beta}$. The estimate of the variance of $\left(\hat{Y}_0|x'_0\right)$ is given by:

$$Var\left(\hat{Y}_0|x'_0\right) = \sigma^2 x'_0 (X'X)^{-1} x_0 \quad \dots (7.12)$$

where σ^2 is the variance of u_i, x'_0 are the given values of the X variables for which we wish to predict the future values, and $(X'X)$ is the matrix. In practice, we replace σ^2 by its unbiased estimator $\hat{\sigma}^2$.

Check Your Progress 1 [answer the questions in 50-100 words within the given space]

1) Specify the simplest form of a multiple regression model with examples. Why is it the simplest?

....
....
....
....
....

2) Enumerate the assumptions made for the CLRM in broad terms. What is the additional assumption made for the multiple regression model?

....
....
....
....
....

3) How are the estimated parameters of a multiple regression model interpreted?

....
....
....
....
....

4) Specify the satisfaction of the property which makes the OLS estimators obey the Gauss Markov theorem?

....
....
....
....
....

7.5 COEFFICIENT OF DETERMINATION: R^2

In multiple regression, a measure of goodness of fit is given by R^2. This is also called as the 'coefficient of determination'. It is the ratio of the 'explained sum of squares' to the 'total sum of squares'. In other words, it is the proportion of total variation in the dependent variable explained by the independent (or the explanatory) variables included in the model. To derive R^2, we consider the sample regression function or equation as follows:

$$Y_i = b_1 + b_2X_{2i} + b_3X_{3i} + e_i \quad \dots (7.13)$$

where $b_1 = \bar{Y} - b_2\bar{X}_2 - b_3\bar{X}_3$. Substituting b_1 in (7.13), and by considering X_{2i} and X_{3i} in their means, we get:

$$Y_i = \bar{Y} - b_2\bar{X}_2 - b_3\bar{X}_3 + b_2X_{2i} + b_3X_{3i} + e_i$$

Therefore, $Y_i - \bar{Y} = b_2(X_{2i} - \bar{X}_2) + b_3(X_{3i} - \bar{X}_3) + e_i$

Rewriting the above in lower case, i.e., by considering in deviation from mean, we get:

$$y_i = b_2x_{2i} + b_3x_{3i} + e_i \quad \dots (7.14)$$

We have $\hat{Y}_i - \bar{Y} = \hat{Y}_i$ where:

$$\hat{Y}_i = b_1 + b_2X_{2i} + b_3X_{3i}$$

$$\bar{Y} = b_1 + b_2\bar{X}_2 + b_3\bar{X}_3$$

$$\therefore\ \hat{Y}_i - \bar{Y} = (b_1 + b_2X_{2i} + b_3X_{3i}) - (b_1 + b_2\bar{X}_2 + b_3\bar{X}_3)$$

$$\hat{Y}_i - \bar{Y} = b_2(X_{2i} - \bar{X}_2) + b_3(X_{3i} - \bar{X}_3)$$

$$\hat{Y}_i - \bar{Y} = b_2x_{2i} + b_3x_{3i} \quad \dots\dots\dots (7.15)$$

Now, consider:

$$y_i = \hat{y}_i + e_i$$

Squaring both sides and summing up we get

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2 + 2\sum \hat{y}_i e_i$$

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2 + 0 \text{ since } Cov(\hat{y}_i, e_0) = 0$$

$$\therefore \sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2 \qquad \ldots (7.16)$$

It means TSS = ESS + RSS. Now, consider: $R^2 = \dfrac{ESS}{TSS}$ where ESS = $\sum \hat{y}_i^2$.

Since $e_i = y_i - \hat{y}_i$ with $\hat{y}_i = b_2 x_{2i} + b_3 x_{3i}$ we have: $e_i = y_i - (b_2 x_{2i} + b_3 x_{3i})$

Now, $\sum e_i^2 = \sum (e_i e_i)$

$= \sum [e_i (y_i - b_2 x_{2i} - b_3 x_{3i})]$

$= \sum e_i y_i - b_2 \sum e_i x_{2i} - b_3 \sum e_i x_{3i}$

$\therefore \sum e_i^2 = \sum e_i y_i$ [since $\sum e_i x_{2i} = \sum e_i x_{3i} = 0$].

$\sum e_i^2 = \sum y_i e_i = \sum y_i (y_i - b_2 x_{2i} - b_3 x_{3i})$

$$\Rightarrow \sum e_i^2 = \sum y_i^2 - b_2 \sum y_i x_{2i} - b_3 \sum y_i x_{3i} \qquad \ldots (7.17)$$

Using (7.17) in (7.16) we get:

$\sum y_i^2 = \sum \hat{y}_i^2 + \sum y_i^2 - b_2 \sum y_i x_{2i} - b_3 \sum y_i x_{3i}$

$\Rightarrow \sum \hat{y}_i^2 = b_2 \sum y_i x_{2i} + b_3 \sum y_i x_{3i}$ = ESS

Therefore, $R^2 = \dfrac{ESS}{TSS} = \dfrac{b_2 \sum y_i x_{2i} + b_3 \sum y_i x_{3i}}{\sum y_i^2}$ … (7.18)

The relationship between R^2 and variance of a partial regression coefficient (b_i) in a *k*-variable multiple regression model is given by:

$$V(b_i) = \frac{\sigma^2}{\sum x_y^2} - \left(\frac{1}{1 - R_i^2}\right)$$

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum y_i^2}$$

$$= 1 - \frac{(n-k)\hat{\sigma}^2}{(n-1)Sy^2}$$

$$\because \sum e_i^2 \text{ or } \hat{\sigma}^2 = \frac{\sum e_i^2}{n-k} \Rightarrow \sum e_i^2 = (n-k)\hat{\sigma}^2$$

$$Sy^2 = \frac{\sum y_i^2}{n-1} \Rightarrow \sum y_i^2 = (n-1)Sy^2$$

7.6 ADJUSTED-R^2

In comparing two regression models with the same dependent variable but differing number of X variables, one should be careful in choosing the model with highest R^2. In order to understand why this is important, consider:

$$R^2 = \frac{ESS}{TSS} = \frac{1-RSS}{TSS} = \frac{1-\sum e_i^2}{\sum y_i^2}$$

Note that as the number of explanatory variables increase, the numerator ESS keeps on increasing. In other words, R^2 increases as k, the number of independent variables increase. The above expression for R^2 implies that R^2 does not give any weightage to the number of independent variables in the model. Due to this reason, for comparison of two regressions with differing number of explanatory variables, we should not use R^2. We now need an alternative coefficient of determination which takes into account the number of parameters estimated, i.e., k. For this, we consider the following measure called the adjusted R^2 defined as follows.

$$\bar{R}^2 = 1 - \frac{RSS/n-k}{TSS/n-1}$$

$$= 1 - \frac{\sum e_i^2/(n-k)}{\sum y_i^2/(n-1)}$$

where k is the number of parameters in the model including the intercept term. The above is same as saying:

$$\bar{R}^2 = \frac{1-\hat{\sigma}^2}{S_y^2}$$

where $\hat{\sigma}^2$ is the residual variance which is an unbiased estimator of true σ^2. S_y^2 is the sample variance of Y. Now, a relationship between $\bar{R}^2$ and R^2 is given by

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k} \quad \ldots (7.19)$$

Now, for deciding on whether R^2 or $\bar{R}^2$ should be used, we must note the following:

(i) If $k > 1, \bar{R}^2 < R^2$. This implies that as the no. of explanatory variables X increases, the adjusted R^2 increases less than the usual R^2

(ii) $\bar{R}^2$ can be negative but R^2 is necessarily non-negative. This is because, in (7.18):

If $R^2 = 1, \bar{R}^2 = 1$.

If $R^2 = 0$, $\bar{R}^2 = \frac{1-k}{n-k}$. Hence, if $k > 1$ then $\bar{R}^2 < 0$.

Thus, adjusted R^2 can be negative. In such cases, it is conventional to take the value of $\overline{R}^2$ as zero. Thus, a conclusive opinion on which of the two is superior to indicate the goodness of fit of a regression model is not possible. However, in practice, in multiple regression models, adjusted R^2 is used to decide for the goodness of fit of the model for the reason that it takes into account the number of regressors and thereby the number of parameters estimated.

Check Your Progress 2 [answer the questions in 50-100 words within the given space]

1) Distinguish between the OLS estimate and the MLE.

....
....
....
....
....

2) How is R^2 defined? Indicate with suitable expressions.

....
....
....
....
....
....

3) State the importance of adjusted-R^2 as compared to R^2.

....
....
....
....
....

4) How are R^2 and adjusted-R^2 related? What is the difference between the two?

....
....
....
....
....

5) How is the situation of adjusted-R^2 being negative dealt with in practice?

....
....
....
....
....

7.7 LET US SUM UP

This unit has described the multiple regression model and its inferences. Recapitulating the assumptions of the multiple classical regression model, the unit indicates how an additional assumption on multicollinearity is necessary in multiple regression models. The interpretation of parameters, i.e., the intercept and the partial slope coefficient are explained. The unit has first discussed the estimation of parameters of the multiple regression model by the OLS (ordinary least squares) method. An alternative method, namely the method of maximum likelihood estimation (MLE) is introduced in the unit next. It is shown that asymptotically the OLS and the MLE coincide. The concept of 'coefficient of determination' or goodness of fit has been described. Finally, the need and the use of adjusted R^2 has been explained.

7.8 ANSWERS/ HINTS TO CHECK YOUR PORGRESS EXERCISES

Check Your Progress 1

1) A multiple regression model is one in which there is more than one independent or the explanatory variable. Hence, the simplest multiple regression model is one with one dependent variable and two independent variables. Such a model is specified as: $\boldsymbol{Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i}$. Examples can be a production function in which the dependent variable is the output and the independent variables are two inputs viz. labour and capital. In microeconomics, it could be a relationship between consumption of a commodity as the dependent variable and price and income as the two independent variables.

2) (i) The model is linear in parameters; (ii) u_i and X_i are not correlated, i.e., $\boldsymbol{cov(u_i, X_{2i}) = Cov(u_i, X_{3i}) = 0}$; (iii) the conditional expectation of the error term is zero, i.e., $\boldsymbol{E(u_i|X_{2i}, X_{3i}) = 0}$; (iv) error terms are not correlated or there is no auto correlation, i.e., $\boldsymbol{\text{cov}(u_i, u_j) = 0}$; (v) there is homoscedasticity or the error variance do not differ, i.e., $\boldsymbol{var(u_i^2) = \sigma^2}$; (vi) no multicollinearity or perfect collinearity, i.e., $Corr(X_i, X_j) \neq 1$; (vii) number of observations (n) is greater than the number of parameters estimated (k); (viii) there is no specification bias, i.e., neither a relevant variable is omitted nor an irrelevant variable is included in the model; and (ix) there is no measurement error in X's and Y. Assumption no (vi) above is the additional assumption required in multiple regression models.

3) The intercept β_1 measures the expected value of the dependent variable Y, given the values of explanatory variables X_2 and X_3. β_2 measures the change in the mean value of Y [i.e., E(Y)] per unit change in X_2, holding the value of X_3 constant. This means: $\beta_2 = \frac{\Delta E(Y)}{\Delta X_2}$. Likewise, β_3 is defined.

4) Under the assumptions of CLRM, the OLS estimators of partial regression coefficients are not only linear and unbiased but also have minimum variances in the class of all unbiased estimators, i.e., they are BLUE (best liner unbiased estimate). It is this property that makes the OLS estimates satisfy the Gauss-Markov theorem.

1) Check Your Progress 2The OLS estimators are obtained by minimizing the residual sum of squares, i.e., Min $\sum e_i^2 = \sum\left(Y_i - \hat{Y}_i\right)^2$. The MLEs are obtained by maximising the 'likelihood function' of the corresponding pdf. There is thus a basic difference in the approach of the two methods. However, once the first order conditions are applied and simplified, the equations that we obtain in the MLE approach is same as the normal equations that we get in the OLS method. Hence, the estimates for the parameters obtained by solving those equations are the same. However, there is an essential difference relating to the unbiased estimate of σ^2. The denominator of the expression for this unbiased estimate in the OLS method is '*n*-*k*' whereas in the ML method it is '*n*'. This important difference makes the estimate of σ^2 in the ML approach biased for small samples. For large samples, it is unbiased. Hence, the estimates of ML and OLS are similar and asymptotically, the OLS and the MLEs coincide.

2) For a 2 independent variables multiple regression model, whose sample regression function is given as $Y_i = b_1 + b_2 X_{2i} + b_3 X_{3i} + e_i$ the R^2 is defined as: $R^2 = \frac{ESS}{TSS} = \frac{b_2 \sum y_i x_{2i} + b_3 \sum y_i x_{3i}}{\sum y_i^2}$.

(iii) For comparing two multiple regressions with differing number of explanatory variables, relying on R^2 could be misleading. This is because R^2 does not take into account the number of explanatory variables.

(iv) They are related as: $\bar{R}^2 = 1 - \left(1 - R^2\right)\frac{n-1}{n-k}$. An important difference is that while R^2 cannot be negative, adjusted R^2 can be negative.

5) When this is negative, conventionally it is taken as zero.