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:

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

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:

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

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: $Y_i = b_1 + b_2 X_{2i} + b_3 X_{3i} + e_i$. 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:

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

$$= \sum (Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i})^2 \text{ [since } \hat{Y}_i = b_1 + b_2 X_{2i} + b_3 X_{3i}]$$

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:

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

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

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

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 $\widehat{Y}\imath$ is equal to the mean value of actual Y_i, i.e., $\overline{\hat{Y}}_i = \overline{Y}$.

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) $Cov\left(e_i, \hat{Y}_i\right) = 0$, 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:

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

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

……

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

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

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

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

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:

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

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:

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

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:

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

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

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

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

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

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

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

Now, consider:

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

Squaring both sides and summing up we get

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

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

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

It means TSS = ESS + RSS. Now, consider: $R^2 = \frac{ESS}{TSS}$ 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, $R^2 = \frac{ESS}{TSS} = \frac{b_2 \sum y_i x_{2i} + b_3 \sum y_i x_{3i}}{\sum y_i^2} \qquad \ldots (7.18)$

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., $\boldsymbol{\mathbf{cov}(u_i, u_j) = 0}$; (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) ≠ 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 $Y_i = b_1 + b_2X_{2i} + b_3X_{3i} + e_i$ the R^2 is defined as: $R^2 = \dfrac{ESS}{TSS} = \dfrac{b_2\sum y_i x_{2i} + b_3\sum y_i x_{3i}}{\sum y_i^2}$.

(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: $\overline{R}^2 = 1 - \left(1 - R^2\right)\dfrac{n-1}{n-k}$. 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.

UNIT 8 MULTIPLE LINEAR REGRESSION MODEL: INFERENCES*

Structure

8.0 OBJECTIVES

After going through this unit, you should be able to

- explain the need for the assumption of normality in the case of multiple regression;
- describe the procedure of testing of hypothesis on individual estimators;
- test the overall significance of a regression model;
- test for the equality of two regression coefficients;
- explain the procedure of applying the Chow test;
- make prediction on the basis of multiple regression model;
- interpret the results obtained from the testing of hypothesis, both individual and joint; and
- apply various tests such as likelihood ratio (LR), Wald (W) and Lagrange Multiplier Test (LM).

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

8.1 INTRODUCTION

In the previous unit we discussed about the interpretation and estimation of multiple regression models. We looked at the assumptions that are required for the ordinary least squares (OLS) and maximum likelihood (ML) estimation. In the present Unit we look at the methods of hypothesis testing in multiple regression models.

Recall that in Unit 3 of this course we mentioned the procedure of hypothesis testing. Further, in Unit 5 we explained the procedure of hypothesis testing in the case of two variable regression models. Now let us extend the procedure of hypothesis testing to multiple regression models. There could be two scenarios in multiple regression models so far as hypothesis testing is concerned: (i) testing of individual coefficients, and (ii) joint testing of some of the parameters. We discuss the method of testing for structural stability of regression model by applying the Chow test. Further, we discuss three important tests, viz., Likelihood Ratio test, Wald test, and Lagrange Multiplier test. Finally, we deal with the issue of prediction on the basis of multiple regression equation.

One of the assumptions in hypothesis testing is that the error variable u_i follows normal distribution. Is there a method to test for the normality of a variable? We will discuss this issue also. However, let us begin with an overview of the basic assumptions of multiple regression models.

8.2 ASSUMPTIONS OF MULTIPLE REGRESSION MODELS

In Unit 7 we considered the multiple regression model with two explanatory variables X_2 and X_3. The stochastic error term is u_i.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad \ldots (8.1)$$

8.2.1 Classical Assumptions

There are seven assumptions regarding the multiple regression model. Most of these assumptions are regarding the error term. We discussed about these assumptions in the previous Unit. Let us briefly mention those assumptions again.

a) The regression model is linear in parameters and variables.

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

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

c) There is no serial correlation (or autocorrelation) among the error terms. The error terms are not correlated. It implies that the covariance between the error term associated with i^{th} observation u_i and the error term associated with j^{th} observation, u_j is zero.

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

d) Homoscedasticity: The assumption of homoscedasticity states that the error variance is constant throughout the population. The variance of the error term associated at each observation has the same variance.

$$var(u_i) = \sigma^2$$

e) Exogeneity of explanatory variables: There is no correlation between the explanatory variables and the error term. This assumption is also called exogeneity, because the explanatory variables are assumed to be exogenous (given from outside; X is not determined inside the model). In contrast, Y is determined within the model. When the explanatory variable is correlated with the error term, it is called endogeneity problem. In order to avoid this problem, we assume that the explanatory variables are kept fixed across samples.

f) Independent variables are not linear combination of one another. If there is perfect linear relationship among the independent variables, the explanatory variables move in harmony and it is not possible to estimate the parameters. It is also called multicollinearity problem.

g) The error variable is normally distributed. This assumption is not necessary in OLS method for estimation of parameters. It is required for construction of confidence interval and hypothesis testing. In the maximum likelihood method discussed in the previous Unit, in order to estimate the parameters we assumed that the error term follows normal distribution.

8.2.2 Test for Normality of the Error Term

As pointed out earlier, we look into the assumption of normality of the error term. In order to test for normality of the error term we apply the Jarque-Bera test (often called the JB test). It is an asymptotic or large sample test. We do not know the error terms in a regression model; we know the residuals. Therefore, the JB test is based on the OLS residuals. Recall two concepts from statistics: skewness and kurtosis. A skewed curve (i.e., asymmetric) is different from a normal curve. A leptokurtic or platykurtic curve (i.e., tall or short in height) is different from a normal curve. The JB test utilises the measures of skewness and kurtosis.

We know that for a normal distribution S = 0 and K = 3. A signification deviation from these two values will confirm that the variable is not normally distributed.

Jarque and Bera constructed the J-statistic given by

$$JB = \frac{n}{6}\left[S^2 + \frac{(K-3)^2}{4}\right] \quad \ldots (8.2)$$

where

n = sample size

S = measure of skewness ($\frac{\mu_3}{\sigma^3}$)

K = measure of kurtosis ($\frac{\mu_4}{\mu_2^2}$)

Skewness and kurtosis are measured in terms of the moments of a variable. As you know from BECC 107, Unit 4, the formula for calculating the rth moment of variable X_i is

$$\mu_r = \frac{1}{n}\sum_{i-1}^{n} f_i(X_i - \bar{X})^r \quad \ldots (8.4)$$

Variance is the second moment μ_2.

In equation (8.2) the JB statistic follows chi-square distribution with 2 degrees of freedom, $\sim\chi^2_{(2)}$.

Let us find out the value of the JB statistic if a variable follows normal distribution. For the normal distribution, as mentioned above S = 0 and K = 3. By asubstitutingthese values in equation (8.2) we obtain

$$JB = \frac{n}{6}[0 + 0] = \frac{n}{6} \times 0 = 0 \quad \ldots (8.3)$$

For a variable not normally distributed JB statistics will assume increasingly large values. The null hypothesis is

H_0: The random variable follows normal distribution.

We draw inferences from the JB statistic as follows:

a) If the calculated value of JB statistic is greater than the tabulated value of χ^2 for 2 degrees of freedom, we reject the null hypothesis. We infer that the random variable is not normally distributed.

b) If the calculated value of the JB statistic is less than the tabulated value of χ^2 for 2 degrees of freedom, we do not reject the null hypothesis. We infer that the random variable is normally distributed.

Check Your Progress 1

1) List the assumptions of multiple regression models.

.....

.....

.....

.....

2) State the Jarque-Bera test for normality.

.....

.....

.....

.....

8.3 TESTING OF SINGLE PARAMETER

The population regression function is not known to us. We estimate the parameters on the basis of sample data. Since we do not know the error variance σ^2, we should apply *t*-test instead of *z*-test (based on normal distribution).

Let us consider the population regression line given at equation (8.1).

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

The sample regression line estimated by ordinary least squares (OLS) method is

$$\hat{Y}_i = b_1 + b_2 X_{2i} + b_3 X_{3i} \quad \ldots (8.4)$$

where b_1, b_2 and b_3 are estimators of β_1, β_2 and β_3 respectively. The estimator of error variance σ^2 is given by $\hat{\sigma}^2 = \frac{RSS}{n-k}$.

There are two approaches to hypothesis testing: (i) test of significance approach, and (ii) confidence interval approach. We discuss both the approaches below.

8.3.1 Test of Significance Approach

In this approach we proceed as follows:

(i) Take the point estimate of the parameter that we want test, viz., b_1, or b_2 or b_3.

(ii) Set the null hypothesis. Suppose we expect that variable X_2 has no influence on *Y*. It implies that β_2 should be zero. Thus, null hypothesis is $H_0: \beta_2 = 0$. In this case what should be alternative hypothesis? The alternative hypothesis is $H_A: \beta_2 \neq 0$.

(iii) If $\beta_2 \neq 0$, then $\beta_2$could be either positive or negative. Thus we have to apply two-tail test. Accordingly, the critical value of the t-ratio has to be decided.

(iv) Let us consider another scenario. Suppose we expect that $\beta_3$should be positive. It implies that our null hypothesis is $H_0: \beta_3 > 0$. The alternative hypothesis is $H_A: \beta_3 \leq 0$.

(v) If $\beta_3 > 0$, then β_3 could be either zero or negative. Thus the critical region or rejection region lies on one side of the *t* probability curve. Therefore, we have to apply one-tail test. Accordingly the critical value of t-ratio is to be decided.

(vi) Remember that the null hypothesis depends on economic theory or logic. Therefore, you have to set the null hypothesis according to some logic. If you expect that the explanatory variable should have no effect on the dependent variable, then set the parameter as zero in the null hypothesis.

(vii) Decide on the level of significance. It represents extent of error you want to tolerate. If the level of significance is 5 per cent (α = 0.05),

your decision on the null hypothesis will go be wrong 5 per cent times. If you take 1 per cent level of significance (α = 0.01), then your decision on the null hypothesis will be wrong 1 per cent times (i.e., it will be correct 99 per cent times).

(viii) Compute the t-ratio. Here the standard error is the positive square root of the variance of the estimator. The formula for the variance of the OLS estimators in multiple regression models is given in Unit 7.

$$t = \frac{b_2 - \beta_2}{se(b_2)} \quad \ldots (8.5)$$

(ix) Compare the computed value of the t-ratio with the tabulated value of the t-ratio. Be careful about the two issues while reading the t-table: (i) level of significance, and (ii) degree of freedom. Level of significance we have mentioned above. Degree of freedom is (n–k), as you know from the previous Unit.

(x) If the computed value of t-ratio is greater than the tabulated value of t-ratio, reject the null hypothesis. If computed value of t-ratio is less than the tabulated value of t-ratio, do not reject the null hypothesis and accept the alternative null hypothesis.

8.3.2 Confidence Interval Approach

We have discussed about interval estimation in Unit 3 and Unit 5. Thus, here we bring out the essential points only.

(i) Remember that confidence interval (CI) is created individually for each parameter. There cannot be a single confidence interval for a group of parameters.

(ii) Confidence interval is build on the basis of the logic described above in the test of significance approach.

(iii) Suppose we have the null hypothesis $H_0: \beta_2 = 0$ and the alternative hypothesis is $H_A: \beta_2 \neq 0$. The estimator of β_2 is b_2. We know the standard error of b_2.

(iv) Here also we decide on the level of significance (α). We refer to the t-table and find out the t-ratio for desired level of significance.

(v) The degree of freedom is known to us, i.e., (n–k).

(vi) Since the above is case of two-tailed test, we take $\alpha/2$ on each side of the t probability curve. Therefore, we take the t-ratio corresponding to the probability $\alpha/2$ and the degrees of freedom applicable.

(vii) Remember that confidence interval is created with the help of the estimator and its standard error. We test whether the parameter lies within the confidence interval or not.

(viii) Construct the confidence interval as follows:

$$\left[b_2 - t_{\alpha/2}SE(b_2) \leq \beta_2 \leq b_2 + t_{\alpha/2}SE(b_2)\right]. \quad \ldots (8.6)$$

(ix) The probability of the parameter remaining in the confidence interval is $(1-\alpha)$. If we have taken the confidence interval as 5 per cent, then the probability that β_2 will remain in the confidence interval is 95 per cent.

$$P_r\left[b_2 - t_{\alpha/2}SE(b_2) \leq \beta_2 \leq b_2 + t_{\alpha/2}SE(b_2)\right] = (1-\alpha) \ldots (8.7)$$

(x) If the parameter (in this case, β_2) remains in the confidence interval, do not reject the null hypothesis.

(xi) If the parameter does not remain within the confidence interval, reject the null hypothesis, and accept the alternative null hypothesis.

Check Your Progress 2

1) Describe the steps you would follow in testing the hypothesis that $\beta_2 < 0$.

.....

.....

.....

.....

.....

2) Create a confidence interval for the population parameter of the partial slope coefficient.

.....

.....

.....

.....

.....

8.4 TEST OF OVERALL SIGNIFICANCE

The overall test of significance of a multiple regression model is carried out by applying *F*-test. We have discussed about the *F*-test in Unit 5 of this course in the context of two variable models. For testing of the overall significance of a multiple regression model we proceed as follows:

(i) Set the null hypothesis. The null hypothesis for testing the overall significance of a multiple regression model is given as follows:

$$H_0: \beta_2 = \beta_3 = \ldots \beta_k = 0 \quad \ldots (8.8)$$

(ii) Set the corresponding alternative hypothesis.

$$H_A: \beta_2 = \ldots = \beta_k \neq 0 \quad \ldots (8.9)$$

(iii) Decide on the level of significance. It has the same connotation as in the case of *t*-test described above.

(iv) For multiple regression model the *F*-statistic is given by

$$F = \frac{ESS/(k-1)}{RSS(n-k)} \quad \ldots (8.10)$$

(v) Find out the degrees of freedom. The *F*-statistic mentioned in equation (8.10) follows *F* distribution with degrees of freedom (k–1, n–k).

(vi) Find out the computed value of *F* on the basis of equation (8.10). Compare it with the tabulated value of *F* (given at the end of the book). Read the tabulated *F* value for desired level of significance and applicable degrees of freedom.

(vii) If the computed value of *F* is greater than the tabulated value, then reject the null hypothesis.

(viii) If the computed value is less than the tabulated value, do not reject the null hypothesis.

8.5 TESTOF EQUALITY BETWEEN TWO PARAMETERS

We can compare between the parameters of a multiple regression model. Particularly, we can test whether two parameters are equal in a regression model. For this purpose we apply the same procedure as we have learnt in the course BECC 107.

Let us take the following regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \quad \ldots (8.11)$$

Recall that we do not know the variance of the parameters. Thus, for comparison of the parameters we apply the *t*-test. Secondly, we do not the parameters. Therefore, we take their OLS estimators for comparison purposes.

Our null hypothesis and alternative hypothesis are as follows:

$$H_0: \beta_3 = \beta_4 \quad \text{or} \quad (\beta_3 - \beta_4) = 0 \quad \ldots (8.12)$$

$$H_1: \beta_3 \neq \beta_4 \quad \text{or} \quad (\beta_3 - \beta_4) \neq 0 \quad \ldots (8.13)$$

For testing of the above hypothesis, the t -statistic is given as follows:

$$t = \frac{(b_3 - b_4) - (\beta_3 - \beta_4)}{SE(b_3 - b_4)} \quad \ldots (8.14)$$

The above follows *t*-distribution with (n – k) degrees of freedom.

Since $\beta_3 = \beta_4$ under the null hypothesis, we can re-arrange equation (8.14) as follows:

$$t = \frac{b_3 - b_4}{\sqrt{V(b_3) + V(b_4) - 2\text{cov}(b_3, b_4)}} \quad \ldots (8.15)$$

The computed value of t-statistic is obtained by equation (8.15). We compare the computed value of *t*-ratio with the tabulated value of *t*-ratio. We read the *t*-table for desired level of significance and applicable degrees of freedom.

If the computed value of *t*-ratio is greater than the tabulated value, then we reject the null hypothesis. If the computed value of *t*-ratio is less than the tabulated value, then we do not reject the null hypothesis and accept the alternative hypothesis.

We need to interpret our results. If we reject the null hypothesis we conclude that the partial slope coefficients β_3 and β_4 are statistically significantly different. If we do not reject the null hypothesis, we conclude that there is no statistically significant difference between the slope coefficients β_3 and β_4.

Check Your Progress 3

1) Mention the steps of carrying out a test of the overall significance a multiple regression model.

.....

.....

.....

.....

.....

2) State how the equality between two parameters can be tested.

.....

.....

.....

.....

.....

8.6 TEST OF LINEAR RESTRICTIONS ON PARAMETERS

Many times we come across situations where we have to test for linear restrictions on parameters. For example, let us consider the Cobb-Douglas production function.

$$Y_i = \beta_1 X_{2i}^{\beta_2} X_{3i}^{\beta_3} e^{u_i} \quad \ldots (8.16)$$

where Y_i is output, X_{2i} is capital and X_{3i} is labour. The parameters are β_2 and β_3. The stochastic error term is u_i. The subscript 'i' indicates the i^{th} observation. The Cobb-Douglas production function exhibits constant returns to scale if the parameters fulfil the following condition:

$$\beta_2 + \beta_3 = 1 \quad \ldots (8.17)$$

As we have discussed in Unit 6, by taking natural log, the Cobb-Douglas production function can be expressed in linear form as

$$ln\, Y_i = ln\, \beta_1 + \beta_2 lnX_{2i} + \beta_3 lnX_{3i} + u_i \quad \ldots (8.18)$$

Suppose have collected data on a sample of firms; our sample size is *n*. The production function is Cobb-Douglas as given above. We want to test whether the production function exhibits constant returns to scale. For this purpose we need to apply the F-test. We can follow two approaches as discussed below.

8.6.1 The t-Test Approach

We will discuss two procedures for testing the hypothesis.

(a) For the In this case our null hypothesis and alternative hypothesis are as follows:

$$H_0: \beta_2 + \beta_3 = 1 \quad \ldots (8.19)$$

$$H_A: \beta_2 + \beta_3 \neq 1 \quad \ldots (8.20)$$

For testing of the above hypothesis, the *t* -statistic is given as follows:

$$t = \frac{(b_2+b_3)-(\beta_2+\beta_3)}{SE(b_2+b_3)} \quad \ldots (8.21)$$

The above follows *t*-distribution with (n – k) degrees of freedom.

We can re-arrange equation (8.21) as follows:

$$\mathrm{t} = \frac{b_2+b_3-1}{\sqrt{V(b_2)+V(b_3)+2\mathrm{cov}(b_2,b_3)}} \quad \ldots (8.22)$$

The computed value of t-statistic is obtained by equation (8.22). We compare the computed value of *t*-ratio with the tabulated value of *t*-ratio. We read the *t*-table for desired level of significance and applicable degrees of freedom.

If the computed value of *t*-ratio is greater than the tabulated value, then we reject the null hypothesis. If the computed value of *t*-ratio is less than the tabulated value, then we do not reject the null hypothesis and accept the alternative hypothesis.

We need to interpret our results. If we reject the null hypothesis we conclude that the firms do not exhibit constant returns to scale. If we do not reject the null hypothesis, we conclude that the firms exhibit constant returns to scale.

(b) Let us look again at the null hypothesis given at (8.19).

$$H_0: \beta_2 + \beta_3 = 1$$

If the above restriction holds, then we should have

$$\beta_2 = (1 - \beta_3)$$

Let us substitute the above relationship in the Cobb-Douglas production function

$$ln\, Y_i = ln\, \beta_1 + (1 - \beta_3) lnX_{2i} + \beta_3 lnX_{3i} + u_i \quad \ldots (8.23)$$

We can re-arrange terms in equation (8.23) to obtain

$$ln\, Y_i - lnX_{2i} = ln\, \beta_1 - \beta_3 lnX_{2i} + \beta_3 lnX_{3i} + u_i$$

Or,

$$ln(Y_i/X_{2i}) = \beta_0 + \beta_3 ln(X_{3i}/X_{2i}) + u_i \quad \ldots (8.24)$$

Note that the dependent variable in the above regression model is output-labour ratio and the explanatory variable is capital-labour ratio. We can estimate the regression model given at equation (8.24) and find the OLS estimator of β_3.

If $\beta_3 = 1$, then the Cobb-Douglas production will exhibit constant returns to scale.

Therefore, we set the null hypothesis and alternative hypothesis as

$H_0: \beta_3 = 1$ and $H_A: \beta_3 \neq 1$

We apply t-test for individual parameters as mentioned in sub-section 8.3.1. If the null hypothesis is rejected we conclude that the firms do not exhibit constant returns to scale.

8.6.2 Restricted Least Squares

The t-test approach mentioned above may not be suitable in all cases. There may be situations where we have more than two parameters to be tested. In such circumstances we apply the *F*-test. This approach is called the restricted least squares.

Let us consider the multiple regression model given at equation (8.11).

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$$

Suppose we have to test the hypothesis that X_3 and X_4 do not influence the dependent variable Y. In such a case, the parameters β_3 and β_4 should be zero.

Recall that if we increase the number of explanatory variables in a regression model, there is an increase R^2. Recall further that $R^2 = \frac{ESS}{TSS}$. Thus, if two of the explanatory variables in equation (8.11) are dropped (i.e., their coefficients are zero), there will be a decrease in the value R^2. If the variables that X_3 and X_4 are relevant, there will be a significant decline the value of R^2. On the other hand, if the variables X_3 and X_4 are not relevant for the regression model, then the decline in the value of R^2 will be insignificant. We use this property of the regression model to test hypotheses on a group of parameters. Therefore, while applying F-test in restricted least squares we estimated the regression model twice: (i) the unrestricted model, and (ii) the restricted model.

We proceed as follows:

(i) Suppose there are *k* explanatory variables in the regression model.

(ii) Out of these *k* explanatory variables, suppose the first *m* explanatory variables are not relevant.

(iii) Thus our null hypothesis will be as follows:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_m = 0 \quad \ldots (8.25)$$

(iv) The corresponding alternative hypothesis will be that the βs are not zero.

(v) Estimate the unrestricted regression model given at (8.11). Obtain the residual sum of squares (RRS) on the basis of the estimated regression equation. Denote it as RSS_{UR}.

(vi) Estimate the restricted regression model by excluding the explanatory variables for which the parameters are zero. Obtain the residual sum of squares (RRS) from this restricted model. Denote it as RSS_R.

(vii) Our F-statistic is

$$F = \frac{RSS_R - RSS_{UR}/m}{RS_{UR}/(n-k)} \quad \ldots (8.26)$$

The *F*-statistic at (8.26) follows *F*-distribution with degrees of freedom (*m*, *n–k*).

(ix) Find out the computed value of *F* on the basis of equation (8.10). Compare it with the tabulated value of *F* (given at the end of the book). Read the tabulated *F* value for desired level of significance and applicable degrees of freedom.

(x) If the computed value of *F* is greater than the tabulated value, then reject the null hypothesis.

(xi) If the computed value is less than the tabulated value, do not reject the null hypothesis.

As mentioned earlier, the residual sum of squares (RSS) and the coefficient of determination (R^2) are related. Therefore, it is possible to carry out the *F*-test on the basis of R^2 also. If we have the coefficient of determination for the unrestricted model (R^2_{UR}) and the coefficient of determination for the restricted model (R^2_R), then we can test the joint hypothesis about the set of parameters.

The *F*-statistic will be

$$F = \frac{R^2_{UR} - R^2_R/m}{(1-R^2_{UR})/(n-k)} \quad \ldots (8.27)$$

which follows *F*-distribution with degrees of freedom (*m*, *n–k*).

The conclusion to be drawn and interpretation of results will be the same as described in points (x) and (xi) above.

8.7 STRUCTURAL STABILITY OF A MODEL: CHOW TEST

Many times we come across situations where there is a change in the pattern of data. The dependent and independent variables may not remain the same throughout the sample. For example, saving behaviour of poor and rich households may be different. The production of an industry may be different after a policy change. In such situations it may not be appropriate to run a single regression for the entire dataset. There is a need to check for structural stability of the econometric model.

There are various procedures to bring in structural breaks in a regression model. We will discuss about the dummy variable cases in unit 9. In this Unit we discuss a very simple and specific case.

Suppose we have data on n observations. We suspect that the first n_1 observations are different from the remaining n_2 observations (we have $n_1 + n_2 = n$). In this case run the following three regression equations:

$Y_t = \lambda_1 + \lambda_2 X_t + u_t$ (number of observations: n_1) … (8.28)

$Y_t = r_1 + r_2 X_t + v_t$ (number of observations: n_2) … (8.29)

$Y_t = \alpha_1 + \alpha_2 X_t + w_t$ (number of observations: $n = n_1 + n_2$) … (8.30)

If both the sub-samples are the same, then we should have $\lambda_1 = r_1 = \alpha_1$

and $\lambda_2 = r_2 = \alpha_2$. If both the sub-samples are different then there will be a structural break in the sample. It implies the parameters of equations (8.28) and (8.29) are different. In order to test for the structural stability of the regression model we apply Chow test.

We process as follows:

(i) Run the regression model (8.28). Obtain residual sum of squares RSS_1.

(ii) Run regression model (8.29). Obtain residual sum of squares RSS_2.

(iii) Run regression model (8.30). Obtain residual sum of squares RSS_3.

(iv) In regression model (8.30) we are forcing the model to have the same parameters in both the sub-samples. Therefore, let us call the residual sum of squares obtained from this model RSS_R.

(v) Since regression models given at (8.28) and (8.29) are independent, let us call this the unrestricted model. Therefore, $RSS_{UR} = RSS_1 + RSS_2$

(vi) Suppose both the sub-samples are the same. In that case there should not be any difference between RSS_{UR} and RSS_R. Our null hypothesis in that case is H_0: There is not structural change (or, there is parameter stability).

(vii) Test the above by the following test statistic:

$$F = \frac{RSS_R - RSS_{UR})/K}{RSS_{UR}/n_1+n_2-2k} \quad \ldots (8.31)$$

It follows F-distribution with degrees of freedom $k, (n_1 + n_2 - 2k)$, where k is the number of explanatory variables in the regression model.

(viii) Check the F-distribution table given at the end of the book for desired level of significance and applicable degrees of freedom.

(ix) Draw the inference on the basis of computed value of the F-statistic obtained at step(vii).

(x) If the computed value of *F* is greater than the tabulated value, then reject the null hypothesis.

(xi) If the computed value is less than the tabulated value, do not reject the null hypothesis.

The Chow test helps us in testing for parameter stability. Note that there are three limitations of the Chow test.

(i) We assume that the error variance σ^2 is constant throughout the sample. There is no difference in the error variance between the sub-samples.

(ii) The point of structural break is not known to us. We assume that point of structural change.

(iii) We cannot apply Chow test if there are more than one structural break.

8.8 PREDICTION

In Unit 5 we explained how prediction is made on the basis of simple regression model. We extend the same procedure to multiple regression models. As in the case of simple regression models, there are two types of prediction in multiple regression models.

If we predict an individual value of the dependent variable corresponding to particular values of the explanatory variables, we obtain the 'individual prediction'. When we predict the expected value of *Y* corresponding to particular values of the explanatory variables, it is called 'mean prediction'. The expected of *Y* in both the cases (individual prediction and mean prediction) is the same. The difference between mean and individual predictions lies in their variances.

8.8.1 Mean Prediction

Let

$$\mathrm{X}_0 = \begin{bmatrix} 1 \\ X_{02} \\ X_{03} \\ \vdots \\ X_{0k} \end{bmatrix} \quad (8.32)$$

be the vector of values of the X variables for which we wish to predict $\hat{Y}_0$.

The estimated multiple regression equation, in scalar form, is

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots \hat{\beta}_k X_{ki} + u_i \quad \text{... (8.33)}$$

which in matrix notation can be written compactly as

$$\hat{Y}_i = X_i' \hat{\beta} \quad \text{... (8.34)}$$

where

$$X_i' = [1 \ X_{2i} \ X_{3i} \ \ldots X_{ki}] \quad \text{... (8.35)}$$

and

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} \quad \text{... (8.36)}$$

Equation (8.34) is the mean predication of Y_i corresponding to given X_i'.

If X_i' is as given in (8.35), then (8.34) becomes

$$\left(\hat{Y}_i | X_0'\right) = X_0' \hat{\beta} \quad \text{... (8.37)}$$

where the values of x_0 are fixed. You should note that (8.36) gives an unbiased prediction of $E\left(\hat{Y}_i | X_0'\right)$, since $E\left(X_0' \hat{\beta}\right) = X_0' \hat{\beta}$.

Variance of Mean Prediction

The formula to estimate the variance of $\left(\hat{Y}_0 | X_0'\right)$ is as follows:

$$\text{var}\left(\hat{Y}_0 | X_0'\right) = \sigma^2 X_0' (X'X)^{-1} X_0 \quad \text{... (8.38)}$$

where σ^2 is the variance of u_i

X_0' are the X variables for which we wish to predict, and

since we do not know the error variance (σ^2), we replace it by its unbiased estimator $\hat{\sigma}^2$.

8.8.2 Individual Prediction

As mentioned earlier, expected value of individual prediction is the same as that of individual prediction, i.e., $\hat{Y}_i$. The variance of the individual prediction is

$$\text{var}(Y_0 | X_0) = \sigma^2 [1 + X'_0 (X'X)^{-1} X_0] \quad \text{... (8.39)}$$

where $\text{var}(Y_0 | X_0)$ stands for $E\left[Y_0 - \hat{Y}_0 \mid X\right]^2$. In practice we replace σ^2 by its unbiased estimator $\hat{\sigma}^2$.

Check Your Progress 4

1) Consider a Cobb-Douglas production. Write down the steps of testing the hypothesis that it exhibits constant returns to scale.

.....

.....

.....

.....

2) Write down the steps of carrying out Chow test.

.....

.....

.....

.....

3) Point out why individual prediction has higher variance than mean prediction.

.....

.....

.....

.....

8.9 LET US SUM UP

This unit described the assumptions of classical multiple regression that fortifies normality of error term also tested by Jarque-Bera Test (J-Test for Normality). The testing of hypothesis about individual coefficients is distinguished from the overall significance test in the unit. The unit also describes the testing of equality of two regression coefficients. Later the structural stability is tested using Chow test. The multiple regression is also used for prediction of dependent variables for given values of independent variables. Both individual and joint hypothesis testing is described in the unit. Various tests such as likelihood ratio (LR), Wald (W) and Lagrange Multiplier Test (LM) are explained in the unit

8.10 ANSWERS/ HINTS TO CHECK YOUR PORGRESS EXERCISES

Check Your Progress 1

1) Refer to Sub-Section 8.2.1 and answer.

2) The Jarque-Bera test statistic is given at equation (8.2). Describe how thetest is carried out.

Check Your Progress 2

1) Refer to Sub-Section 8.3.1 and answer. Decide on the null and alternative hypotheses. Describe the staps you would follow.

2) Refer to Sub-Section 8.3.2 and answer.

Check Your Progress 3

1) It can be tested by F-test. See Section 8.4 for details.

2) Refer to Sub-Section 8.5 and answer.

Check Your Progress 4

1) We have explained in Sub-Section 8.6.1. Refer to it.

2) Refer to Sub-Section 8.7 and answer.

3) Refer to Sub-Section 8.8 and answer. It has the same logic as in the case of two variable models discussed in Section 5.7 of Unit 5.

UNIT 9 EXTENSION OF REGRESSION MODELS: DUMMY VARIABLE CASES*

Structure

9.0 OBJECTIVES

After reading this unit, you will be able to:

- define a qualitative or dummy variable;
- discuss the ANOVA model with a single dummy as exogenous variable;
- specify an ANCOVA model with one quantitative and one dummy variable;
- interpret the results of dummy variable regression models;
- differentiate between 'differential intercept coefficient' and 'differential slope coefficient;
- describe the concepts of 'concurrent, dissimilar and parallel' regression models that you encounter while considering 'differential slope dummies'; and
- explain how more than two dummies and interactive dummies can be formulated into a regression model.

9.1 INTRODUCTION

In real life situations, some variables are qualitative. Examples are gender, choices, nationality, etc. Such variables may be dichotomous or binary, i.e., with responses limited to two such as in 'yes' or 'no' situations. Or they may have more than two categorical responses. We need methods to include such variables in the regression model. In this unit, we consider some such cases. We limit this unit to consider regressions in which the dependent variable is quantified. You may note in passing that when the dependent variable itself is a dummy variable, we have to deal with them by models such as Probit or Logit. In such models, the

* Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi and Prof. B S Prakash, Indira Gandhi National Open University, New Delhi

OLS method of estimation does not apply. In this unit, we will not consider such cases. You will study about them in the course 'BECE 142: Applied Econometrics'.

In this unit, we consider only such cases in which the independent variable is a dummy variable. Qualitative variables are not straightaway quantified. By treating them as dummy variables we can make them quantified (or categorical). For instance, consider variables such as male or female, employed or unemployed, etc. These are quantifiable in the sense that by treating them as 1 if 'female, and 0 if 'male'. Similar examples could be 1 if yes and 0 if no; 1 if employed and 0 if unemployed, etc. In the above, we have converted a qualitative response into quantitative form. Thus, the qualitative variable is now quantified. Such regressions could be a simple regression, i.e., there is only one independent variable which is qualitative and treated as dummy variable. Or there could be two independent variables, one of which can be treated as dummy and the other is its covariant, i.e., there is a close relationship with the variable treated as dummy. For instance, pre-tax income of persons can be classified above a threshold level and treated as dummy variable, i.e., above or below the threshold level income with response taken as 1 or 0. Now, the post-tax income, which is a co-variant of pre-tax income, can be considered by its actual quantified value. There could be similar extension of situations where you have to consider multiple dummies and cases where you have to consider interactive dummies. The nature of such regressions, particularly for their inference or interpretational interest, is what we consider in the present unit.

9.2 THE CASE OF SINGLE DUMMY: ANOVA MODEL

We first consider a simple regression model with only one independent variable. Further, this independent variable is a dummy variable such as:

$$Y_i = \beta_1 + \beta_2 D_i + u_i \quad \ldots (9.1)$$

Here, we take Y as the annual expenditure on food and D_i as gender taking the values 0 if the person is male and 1 if female. The D_i's are thus fixed and hence non-stochastic. Now, if we assume that $u_i \sim N(0, \sigma^2)$, the OLS method can be applied to estimate the parameters in (9.1). If we do this, the mean food expenditure for males and females are respectively given by:

$$E(Yi \mid Di = 0) = \beta_1 + \beta_2(0) = \beta_1 \quad \ldots (9.2)$$

$$E(Y_i \mid D_i = 1) = \beta_1 + \beta_2 \quad \ldots (9.3)$$

Here, β_1 gives the average or mean food expenditure of males. It is the category for which the dummy variable is given the value 0. The slope coefficient β_2 tells us by how much the mean food expenditure of females differ from that of the mean food expenditure of males. Hence, $\beta_1 + \beta_2$ gives the mean food expenditure for females. In view of this, it is not correct to call β_2 as the slope coefficient since there is no continuous regression line here. Hence, β_2 is the 'differential

intercept coefficient'. It tells us by how much the value of intercept term differs between the two categories. A question that arises now is, what would have happened if we had interchanged the assignment of '0' between the two categories of males and females (i.e., if we had assigned the value '0' to females). You may note that, so long as we have only two categories as in the present instance, i.e., it is a case of simple regression with only one independent variable taken as a dummy variable D_i with the category of responses dichotomous or binary, it basically does not matter which category gets the value of 1 and which gets the value 0. However, some minor difference would be there. Let us see what this is.

The category to which we assign the value 0 is called as the base category. It is also called by alternative names such as reference or benchmark or the comparison category. In such an assignment, the intercept value represents the mean value of the category that gets the value 0 (which is males in our case above). What equation (9.3) tells us is, depending on such an assignment, the mean value of expenditure on food for females is to be obtained by adding the 'slope coefficient to the intercept value'. If the assignment of dummy is made the other way, i.e., females 0 and males 1, we see a change in the numerical value of the intercept term and its t value. Barring this, the R^2 value, the absolute value of the estimated dummy variable coefficient and its standard error, will remain the same. Let us see this with the help of an example for better understanding.

Consider the data on 'expenditure on food' and income for males and females as in Table 9.1. The data are averages based on the actual number of people (who are in thousands) in different age groups. We first construct Table 9.2 from the data in Table 9.1 as below.

Table 9.1: Data on Income and Food Expenditure by Gender

(Figures in $)

Age	Food Expenditure (female)	Income (female)	Food Expenditure (male)	Income (male)
< 25	1983	11557	2230	11589
25-34	2987	29387	3757	33328
35-44	2993	31463	3821	36151
45-54	3156	29554	3291	35448
55-64	2706	25137	3429	32988
> 65	2217	14952	2533	20437

Source: Table 6-1, Chapter 6, Gujarati.

Table 9.2: Food Expenditure in Relation to Income and Gender

Observation	Food Expenditure ($)	Income ($)	Gender
1	1983	11557	1
2	2987	29387	1
3	2993	31463	1
4	3156	29554	1
5	2706	25137	1
6	2217	14952	1
7	2230	11589	0
8	3757	33328	0
9	3821	36151	0
10	3291	35448	0
11	3429	32988	0
12	2533	20437	0

Source: Table 6-2, Chapter 6, Gujarati.

Results of food expenditure regressed on the gender dummy variable (without taking into account the income variable at this stage) presents the following results.

$$\widehat{Y}_t = 3176.833 - 503.1667\, D_i$$

$$\text{se} = (233.0446) \quad (329.5749)$$

$$t = (13.6318) \quad (-1.5267) \quad R^2 = 0.1890$$

The results show that the mean expenditure of males is 3177 $ and that of females is (3177 – 503 = 2674 $). The estimated D_i is not statistically significant (since its t value is only –1.53). This means that the difference in the food expenditure between gender is not statistically significant. Recall that we have assigned the value '0' to males. Hence, the intercept value represents the mean value for males. In this assignment, to get the mean value of food expenditure of females, we add the value of the coefficient of the dummy variable to the intercept value. Now, let us re-assign the value '0' to females and '1' to males. The regression results that we get are the following:

$$\widehat{Y}_i = 2673.667 + 503.1667\ D_i$$

se = (233.0446) (329.5749)

t = (11.4227) (–1.5267) $R^2 = 0.1890$

Thus, we notice that the mean food consumption expenditures of the two genders have remained the same. The R^2 value is also the same. The absolute value of the dummy variable coefficient and their standard errors are also the same. The only change is in the numerical value of the intercept term and its t value.

Another question that we may get is: since we have two categories, male and female, can we assign two dummies to them? This means we consider the model as:

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_i + u_i \quad \ldots (9.4)$$

where Y is expenditure on food, $D_2 = 1$ for female and 0 for male and $D_3 = 1$ for male and 0 for female. Essentially, we are trying to see whether we can assign two dummies for male and female separately? The answer is 'no'. To know the reason for this, consider the data for a sample of two females and three males, for which the data matrix is as in Table 9.3. We see that $D_2 = 1 - D_3$ or $D_3 = 1 - D_2$. This is a situation of perfect collinearity. Hence, we must always use only one dummy variable if a qualitative variable has two categories, such as the gender here.

Table 9.3: Data Matrix for the Equation

Gender	Intercept	D_2	D_3
Male Y_1	1	0	1
Male Y_2	1	0	1
Female Y_3	1	1	0
Male Y_4	1	0	1
Female Y_5	1	1	0

A more general rule is: if a model has the common intercept β_1, and the qualitative variable has m categories, then we must introduce only $(m - 1)$ dummy variables. If we do not do this, we get into a problem of estimation called as the 'dummy variable trap'. Finally, note that when we have a simple regression model with only one dummy variable as considered here, the model considered is also called as the ANOVA model. This is because there is no second variable from which we are seeking to know the impact or variability on the dependent variable. When we have this, we get what we call as an ANCOVA model. We take up such a case in the next section.

9.3 ANALYSIS OF COVARIANCE (ANCOVA) MODEL

In economic analysis, it is common to have among explanatory variables some of which are qualitative and some others quantitative. Such models are called as Analysis-of-Covariance (ANCOVA) models. Here, we shall consider a model that has both a quantitative and a dummy variable among the regressors. In general, regression models containing a combination of quantitative and qualitative variables are called ANCOVA models. Here, the quantitative variables are called covariates or control variables. ANCOVA models are an extension of the ANOVA models. They provide a method of statistically controlling the effects of covariates (i.e., a quantitative explanatory variable) in a model that includes both the type of variables with the qualitative variable treated as a dummy variable. The quantitative variable considered is usually a covariate in the sense that it bears close association with the main variable. Because of this, exclusion of covariates from a model results in model specification error. In the example considered above, we regressed 'food expenditure' on only gender dummy [$Y_i = \beta_1 + \beta_2 D_i + u_i$]. Now, let us consider another variable, 'income after taxes', i.e., disposable income (a covariate of food expenditure) as an explanatory variable (X_i). The model now is

$$Y_i = \beta_1 + \beta_2 D + \beta_3 X_i + u_i \qquad \ldots (9.5)$$

where Y = expenditure on food ($), X = after tax income ($), D = 1 for female and = 0 for male. Let us now consider, for better appreciation, the result for the regression in equation (9.5) obtained from the data in Table 9.2 as follows:

$$\hat{Y}_c = 1506.244 - 228.9868 D_i + 0.0589 X_i$$
t = (8.0115) (–2.1388) (9.6417)
R^2 = 0.9284

The dummy variable coefficient is statistically significant. Therefore, we reject the null hypothesis that there is no difference in the average value of expenditure on food for male and female. In other words, we conclude that gender has a significant impact on consumption or food expenditure. Note that this difference in consumption expenditure is inferred holding the effect of after-tax income constant. Likewise, holding the gender differences constant, the after tax income coefficient is significant. The slope coefficient for 'after tax income' indicates that the mean food expenditure [i.e., the marginal propensity to consume (MPC)] increases by 6 cents for every additional dollar of increase in the disposable income. Note that since we have taken '0' for males, the intercept term relates to the MPC for males. For female MPC, we have to add the intercept value to the coefficient of gender dummy (i.e., 1506.2 – 228.9 = 1277.3). Thus, the equations for the MPC of females and males can be respectively written as:

Mean food expenditure for females: $\hat{Y}_i = 1277.2574 + 0.0589 X_i$

Mean food expenditure for males: $\hat{Y}_i = 1506.2440 + 0.0589 X_i$

Since the MPC or the slope is same for both the gender, the two regressions are parallel as in Fig. 9.1 below.

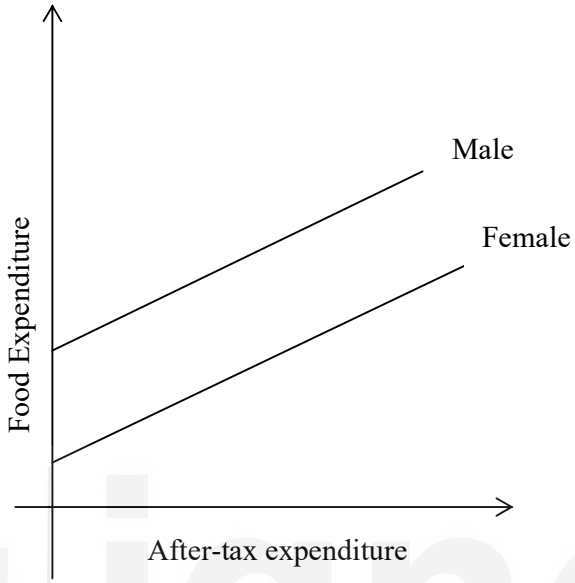


Fig. 9.1 Mean Food Expenditure for Male and Female

The model signifies the role and the impact of both the type of variables (quantitative and qualitative) in explaining a dependent variable. Specifically, in the example considered, the after tax expenditure is seen to affect the food expenditure of both males and females.

Check Your Progress 1 [answer questions in about 50-100 words]

1) Define a qualitative variable.

.....

.....

.....

.....

.....

2) Specify a regression model with a single dummy variable. Mention its features from the point of view of interpretation of estimated coefficients.

.....

.....

.....

.....

.....

3) What happens if the base value is reassigned for the dummy variable, say gender, in a simple regression model as in equation (9.1)?

.....

.....

.....

.....

.....

4) What is meant by 'dummy variable trap'? How do we avoid it?

.....

.....

.....

.....

.....

5) Distinguish between an ANOVA model and an ANCOVA.

.....

.....

.....

.....

.....

6) What is an advantage of ANCOVA model? What is a consequence of omitting the inclusion of a covariant in an ANOVA model?

.....

.....

.....

.....

.....

7) Specify the general form of an ANCOVA model with one qualitative and one quantitative variable. What does the slope oefficient for the quantitative variable considerd indicate in general?

.....

.....

.....

.....

.....

9.4 COMPARISON BETWEEN TWO REGRESSION MODELS

In the example considered above, i.e., for both the ANOVA and the ANCOVA models, we saw that the slope coefficients were same but the intercepts were different. This raises the question on whether the slopes too could be different? How do we formulate the model if our interest is to test for the difference in the slope coefficients too? In order to capture this, we introduce a 'slope drifter'. For the example of consumption expenditure for male or female considered above, let us now proceed to compare the difference in the consumption expenditure by gender by specifying the model with dummies as follows:

$$Y_i = \beta_1 + \beta_2 D_i + \beta_3 X_i + \beta_4 (D_i X_i) + u_i \quad \ldots (9.6)$$

Note that the additional variable added is D_iX_i which is in multiplicative or interactive form. In (9.6), we have taken $D_i = 0$ for males and $D_i = 1$ for females. Now, the 'mean food expenditure' for males is given by:

$$E(Y_i \mid D_i = 0, X_i) = \beta_1 + \beta_3 X_i \quad \ldots (9.7)$$

{since $D_i = 0$}

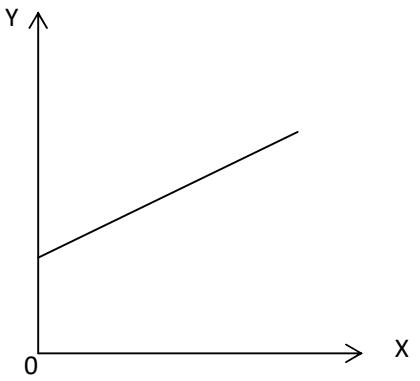
The 'mean food expenditure' for females is given by:

$$E(Y_i \mid D_i = 1, X_i) = \beta_1 + \beta_2 D_i + (\beta_3 + \beta_4 D_i) X_i$$

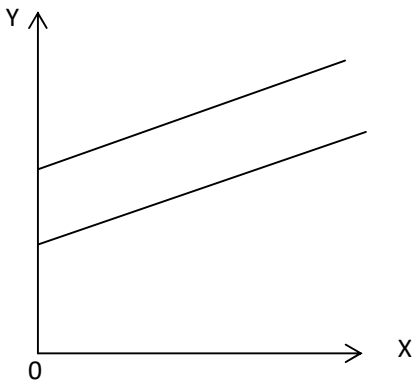
$$= (\beta_1 + \beta_2) + (\beta_3 + \beta_4) X_i \quad \ldots (9.8)$$

{since $D_i = 1$}

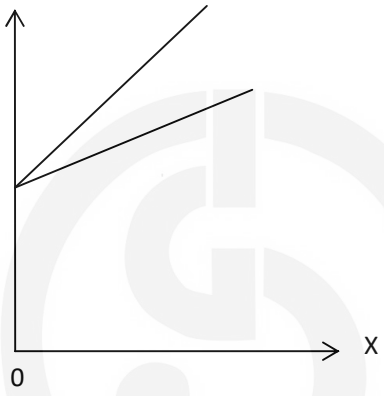
In equation (9.8), $(\beta_1 + \beta_2)$ gives the mean value of Y for the category that receives the dummy value of 1 when X is zero. And, $(\beta_3 + \beta_4)$ gives the slope co-efficient of the income variable for the category that receives the dummy value of 1. Note that the introduction of the dummy variable in the 'additive form' enables us to distinguish between the intercept terms of the two groups. Likewise, the introduction of the dummy variable in the interactive (or multiplicative) form (i.e., D_iX_i) enables us to differentiate between the slope coefficients (or terms) of the two groups. Depending on the statistical significance of the differential intercept coefficient, β_2, and the differential slope coefficient, β_4, we can infer whether the female and male food expenditure functions differ in their intercept values, or their slope values, or both. There can be four possibilities as shown in Fig. 9.2. Fig. 9.2 (a) shows that there is no difference in intercept or the slope coefficient of the two food expenditure regressions. Such regression equations are called 'Coincident Regressions'.



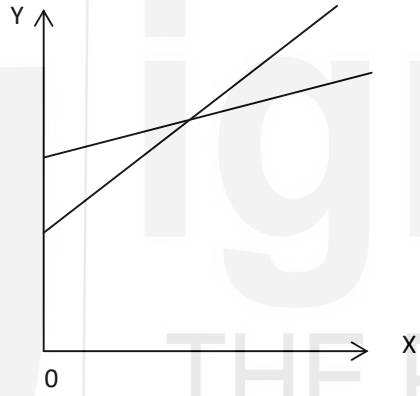
(a) Coincident Regressions



(b) Parallel Regressions



(c) Concurrent Regressions



(d) Dissimilar Regressions

Fig 9.2 Comparison of Regression Equations

Fig. 9.2 (b) shows that the two slope coefficients are the same but intercepts are different. Such regressions are referred to as 'Parallel Regressions'. Fig. 9.2 (c) shows that the two regressions have the same intercepts but

different slopes. Such regressions are referred as 'Concurrent Regressions'. Fig. 9.2 (d) shows that the two intercepts and the two slope coefficients are both different. Such regressions are called 'Dissimilar Regressions'.

9.5 MULTIPLE DUMMIES AND INTERRACTIVE DUMMIES

We often might require to consider more than one dummy variables. Besides, there could be cases where we might be interested in seeing for the impact of dummy variable interactions. Let us consider a case as given below.

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 X_i + u_i \quad \ldots (9.7)$$

where Y is income, X is education measured in number of years of schooling, D_2 is gender (0 if male, 1 if female), D_3 is if in reserved segment or group (e.g. SC/ST/OBC) taking the value 0 if 'not in reserved segment', i.e., in general segment and 1 if 'in reserved segment'. Here, gender (D_2) and reservation (D_3) are qualitative variables and X is quantitative variable. In this formulation (for example, equation 9.7) we have made an implicit assumption that the differential effect of gender is constant across the two segments of reservation. We have likewise assumed that the differential effect of reservation is constant across the two genders. This means if the average income is higher for males than for females, it is so whether the person is in the general segment or in the reservation segment. Likewise, it is assumed here that if the average income is different between the two reservation segments, it is so irrespective of gender. However, in many cases, such assumptions may not be tenable. This means, there could be interaction between gender and reservation dummies. In other words, their effect on average income may not be simply additive as in (9.7) but could be multiplicative. If we wish to consider for this interactive effect, we must specify the model as follows:

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 (D_{2i}\, D_{3i}) + \beta_5 X_i + u_i \quad \ldots (9.8)$$

In equation (9.8), the dummy variable $D_{2i}D_{3i}$ is called as 'interactive or interaction dummy'. It represents the joint or simultaneous effect of two qualitative variables. Taking expectation on both sides of equation (9.8), i.e., by considering the average effect on income across gender and reservation, we get:

$$E(Yi \mid D_{2i}=1, D_{3i}=1, X_i) = \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 X_i \quad \ldots (9.9)$$

Equation (9.9) is the average income function for female reserved category workers where β_2 is the differential effect of being female, β_3 is the differential effect of being in the reserved segment and β_4 is the interactive effect of being both a female and in reserved segment. Depending on the statistical significance of various dummies, we need to make relevant inferences. The specification can easily be generalized for more than one quantitative variable and more than two qualitative variables.

Check Your Progress 2 [answer questions within the given space in about 90-100 words]

1) What is meant by a 'slope drifter'? When is it introduced and for what use? Specify a general model with such a 'slope drifter' and comment on the additional variable introduced.

.....

.....

.....

.....

.....

2) Differentiate between the four type of regressions that we might get when considering a model of the type in equation (9.6) with two slope drifters β_2 and β_4 as therein.

.....

.....

.....

.....

.....

3) List the four types of regression models, with dummy variables to accommodate different cases or situations, as we have considered in this unit. Specify their difference by their name and features.

.....

.....

.....

.....

.....

9.6 LET US SUM UP

This unit makes a distinction between qualitative and quantitative variables. It has considered three types of models in which the focus is kept on inclusion of qualitative variables in the regression models. The first of such models is considered is a simple regression model. In this, we have considered only one dummy variable, as an independent variable, on the RHS of the regression equation. This equation is of the form: $Y_i = \beta_1 + \beta_2 D_i + u_i$. Analysis in this form is called as ANOVA. Quite often, we would be committing a specification bias if we consider the regression model in this form. This happens because the variable Y_i will be clearly related to a variable X_i which is a quantitative variable. To accommodate this, we considered the second type of model in which we included a co-variant (X_i) into the regression equation: $Y_i = \beta_1 + \beta_2 D + \beta_3 X_i + u_i$. Analysis in this form is called as ANCOVA. In both these type of models, our focus was only on observing the significance of difference in the intercepts. But in practice, we do encounter a number of situations in which not only the intercept, but the slope too could vary between categories. To allow for this kind of situation, we considered a third type of model in which we accommodated for the interactive effect of the 'dummy variable with the quantitative variable', i.e., D_iX_i. The regression model considered for this kind of an analysis is of the form: $Y_i = \beta_1 + \beta_2 D_i + \beta_3 X_i + \beta_4 (D_i X_i) + u_i$. In this situation, we noted that we could come across four possibilities viz. coincidental, parallel, concurrent and dissimilar regressions. We have finally considered the case where a regression model may have to be formulated to accommodate more than one qualitative

variable and a case where we might be interested in examining for the interactive effect of the two qualitative variables. For this, we considered models such as $Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 (D_{2i} D_{3i}) + \beta_5 X_i + u_i$.

9.7 ANSWERS/ HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) A qualitative variable is one which has a categorical response such as yes/no or employed/unemployed or male/female. If the response is limited to two, as in these cases, it is called as a dichotomous variable. The responses can be more than two. But they may be classified as 1, 2, 3, …….. Such responses are unambiguous or categorical. Hence, a qualitative variable is also called as dummy variable or categorical variable.

2) The model in this case can be $Y_i = \beta_1 + \beta_2 D_i + u_i$. We are considering the dependent variable Y_i as quantitative variable. The D_i's are thus fixed and hence non-stochastic. D_i is taken a dichotomous, i.e., it takes the values 0 and 1. In such cases, the factor or entity which is assigned the value 0, is called as the base category. The estimated value of the mean of Y_i, given $D_i = 0$, is given by β_1. Here, β_2 is not strictly the slope coefficient but is the 'differential intercept coefficient'. The estimated value of the mean of Y_i, given $D_i = 1$, is given by $\beta_1 + \beta_2$.

3) The mean value of Y_i for the two gender classes, the R^2 value, the absolute value of the estimated dummy variable coefficient and the standard errors will be the same. The numerical value of the intercept term and its t value will change.

4) The number of responses to the dummy variable is called as 'categories' of response. If the dummy variable refers to gender of the respondent, there are two categories of response viz. male and female. If we assign two separate dummies in such cases, we encounter a situation of perfect collinearity. Hence, we will not get unique estimates or one of the two parameters is not estimable. This situation is called as 'dummy variable trap'. To avoid this situation, the general rule is if we have m categories, we limit the number of dummies to '$m - 1$'. The models should also have a common intercept β_1.

5) If the regression model considered has only one independent variable in general, and that variable is a dummy variable as considered here in particular, then the variation or the sources of variability that is sought to be identified for the dependent variable is limited to that one variable. In such cases, the regression model considered is called as an ANOVA model. If the independent variables considered are two, with one considered as dummy variable, and the other variable considered is related to the dummy variable, then such models are called as ANCOVA model.

In other words, regression models in which some independent variables are qualitative and some others are quantitative, are called as ANCOVA models.

6) The advantage is that ANCOVA models provide a method of statistically controlling the effects of covariates. The consequence of excluding a covariant from being included in the model is that the model suffers from 'specification error'. The consequence of committing specification errors are that the ideal assumptions required for the OLS estimators to be efficient are violated. Consequently, they lose out on their efficiency properties.

7) The general form of the model is like: $Y_i = \beta_1 + \beta_2 D + \beta_3 X_i + u_i$. The slope coefficient indicates the rate of increase (or decrease) in the 'marginal propensity to consume (MPC)'. This is when the dependent variable *Y* relates to a consumption variable like expenditure on food and the quantitative independent variable is like disposable income as considered here.

Check Your Progress 2

1) In regression models with one intercept and one slope coefficients, our interest might be to test to know whether: (i) the intercept terms are statistically different and (ii) the slope coefficients are statistically different? For investigating the second question, we need to introduce what is called as a 'slope drifter'. The model specified with such a drifter would be like: $Y_i = \beta_1 + \beta_2 D_i + \beta_3 X_i + \beta_4 (D_i X_i) + u_i$.The additional variable introduced here is *DiXi*. It is a multiplicative variable in the interactive form. Here β_2 and β_4 are the two slope drifters which helps us infer for the statistical difference in the intercept values and the slope values respectively.

2) We get a 'coincident regression' when there is no difference both in intercept as well as the slope. We get a 'parallel regression' when the two intercept terms are different but the two slope coefficients are the same. We get a 'concurrent regression' when the two regressions have the same intercept but different slopes. We get two 'dissimilar regressions' when both the intercept terms and the slope coefficients are different.

3) (i) $Y_i = \beta_1 + \beta_2 D_i + u_i$. (ii) $Y_i = \beta_1 + \beta_2 D + \beta_3 X_i + u_i$. (iii) $Y_i = \beta_1 + \beta_2 D_i + \beta_3 X_i + \beta_4 (D_i X_i) + u_i$. (iv) $Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 (D_{2i}\, D_{3i}) + \beta_5 X_i + u_i$. The first is the ANOVA model in which we have considered only one single dummy variable as the independent variable. The second is the ANCOVA model in which we have considered one qualitative dummy variable and another quantitative exogenous variable related to the dummy variable, the omission of which would lead to a 'specification bias'. The third involves an interactive variable $(D_i X_i)$ in which we try to see whether both the slopes and the intercept coefficients differ. In this, there is a possibility of getting four different type of regressions viz. coincident, parallel, concurrent and dissimilar regressions. The fourth situation considered involves a interactive dummy variable like: $Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 (D_{2i}\, D_{3i}) + \beta_5 X_i + u_i$.