UNIT 13 MODEL SELECTION CRITERIA*

Structure

13.0 OBJECTIVES

After going through this unit, you will be able to

- appreciate the importance of correct specification of an econometric model;
- identify the important issues in specification of econometric models;
- find out the consequences of including an irrelevant variable;
- find out the consequences of excluding a relevant variable; and
- find out the impact of measurement errors in dependent and independent variables.

13.1 INTRODUCTION

In the previous Units of the course we have discussed about various econometric tools. We began with the classical two variable regression model. Later on, we extended it to the classical multiple regression model. The steps of carrying out the ordinary least squares (OLS) method were discussed in details. Recall that the

*Dr. Sahba Fatima, Independent Researcher, Lucknow.

classical regression model is based on certain assumptions. When these assumptions are met, the OLS estimators are the best linear unbiased estimators (BLUE). When these assumptions are violated the OLS estimators are not BLUE – they lose some of their desirable properties. Therefore, when some of the classical assumptions are not fulfilled, we have to adopt some other estimation method.

Thus far our objective has been to explain how various estimation methods are applied. Now let us look into certain other important issues regarding specification of econometric models.

13.2 ISSUES IN SPECIFICATION OF ECONOMETRIC MODEL

A model refers to a simplified version of reality. It allows us to explain, analyse and predict economic behavior. An economic model can be for a microeconomic agent such as household or firm. In macroeconomics, it represents the behavior of the economy as a whole. In economic models we identify relevant economic variables (such as income, output, expenditure, investment, saving, exports, etc.) and establish relationship among them. The relationships among these variables may be expressed through diagrams or mathematical equations. There could be economic models without mathematical expressions, but such models may not be precise.

Recall from Unit 1 of this course that there are eight steps to be followed in an econometric study. The first three steps are as follows:

(i) Construction of a statement of theory or hypothesis

(ii) Specification of mathematical model of the theory

(iii) Specification of econometric model

Based on economic theory or logic we construct the hypothesis. We specify the hypothesis in mathematical terms. Further, we add a stochastic error term (u_i) to transform it into an econometric model. We decide on the estimation method (such as OLS, GLS, maximum likelihood, etc.) subsequently.

13.2.1 Model Specification

While building an econometric model we first consider the logic or theory behind the model. The empirical or methodological considerations come later. The accuracy of the estimated parameters and the inferences drawn from the model depend upon the correct specification of the model.

An econometric model comprises a dependant variable, independent variable(s) and the error term. The dependant variable should be logically explained by the independent variables. Next is the functional form of the regression model, which should be specified correctly.

Let me illustrate the point through an example. In the case of a firm, we assume that there are two factors of production, viz., capital and labour. We club all types of labour into a homogeneous category – we do not distinguish between a manager and a worker in the field! Thus you should remember that we ignore the details and concentrate on the major issues in a model. Secondly, we assume that the production function takes a particular form, say Cobb-Douglas. But, remember that it is just an assumption! The production function in reality could be of some other form. Thus we have to logically explain the functional form (regression equation) of the model.

Regression analysis derives its robustness from the assumption that the econometric model under study is correctly specified. In Unit 4 of this course we specified the assumptions such that the econometric model must bring efficient estimates of the parameters in the model. Ordinary Least Squares (OLS) method is based on the assumption that regression model is correctly specified. Correct specification has three important elements:

a) all the necessary independent variables are included in the model,

b) no redundant variable IS included in the model, and

c) the model is specified using the correct functional form.

13.2.2 Violation of Basic Assumptions

An economic model is based on certain assumptions. Recall that we made the following assumptions regarding the multiple regression model (see Unit 7):

a) The regression model is linear in parameters

b) $E(X_i u_i) = 0$ (regressor is non-stochastic)

c) $E(u_i) = 0$

d) $E(u_i)^2 = \sigma^2$

e) $E(u_i u_j) = 0$ for $i \neq j$

f) The explanatory variables (X_i) are independent of one another.

Let us look into the implications of the above assumptions. Assumption (a) says that the regression model is linear in parameters. Standard regression model usually takes the following form

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \qquad \ldots (13.1)$$

Equation (13.1) is linear in parameters (there are no such terms as β_i^2, for example) and linear in variables. Examples of non-linear regression models are logarithmic functions, logistic functions, trigonometric functions, exponential functions, etc. For estimation of non-linear models, the OLS method cannot be applied.

Assumption (b) says that X_i and u_i are independent. Thus if we take the X_i values randomly, the joint probability of both that X_i and u_i will not be zero. In order to avoid this problem we assume that X_i is non-stochastic. All explanatory variables are fixed in repeated sampling.

Assumption (c) says that the mean of the error term (u_i) is zero. There could be errors in individual observations; on the whole these errors cancel out. If $E(u_i) \neq 0$, OLS estimator of the intercept term (β_1) will be biased. Estimators of the slope parameters β_2 and β_3 will remain unbiased. For example, suppose $E(u_i) = 3$. In that case $E(Y_i)$ will be

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

Remember that β_i are parameters of the model. They are constants. We have assumed X_i to be fixed across samples. Thus

$$E(Y_i) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + E(u_i) \qquad \ldots (13.2)$$

If $E(u_i) = 3$, we can say that

$$E(Y_i) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + 3$$

Thus the intercept term will be $(\beta_1 + 3)$. Remember that if assumption (d) is violated we have the problem of heteroscedasticity, which is discussed in Unit 11. If assumption (e) is violated we have the problem of autocorrelation, that we have discussed in Unit 12. In case the assumption (f) is violated we have the problem of multicollinearity (see Unit 10).

Check Your Progress 1

1) List the assumptions of the classical regression model.

....

....

....

....

....

2) Do you agree that correct specification of an econometric model is important? Why?

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

3) What are the implications of violations of the basic assumptions classical regression model?

....

....

....

....

4) List three types of specification error that we encounter in an econometric model.

....

....

....

....

....

13.3 CONSEQUENCES OF SPECIFICATION ERRORS

As pointed out earlier, we usually encounter three kinds of problems in an econometric model:

a) Inclusion of irrelevant/redundant variables

b) Omission of relevant variables

c) Incorrect functional form of the model

Each of the above problem results in a different kind of bias. We discuss each of these problems below.

13.3.1 Inclusion of Irrelevant Variable

Let us consider the case where some irrelevant variable is included in the regression model. Suppose the true model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i \qquad \ldots (13.3)$$

But we somehow include a redundant variable, i.e., we estimate the following equation:

$$Y_i = \beta_{0s} + \beta_{1s} X_{1i} + \beta_{2s} X_{2i} + v_i \qquad \ldots (13.4)$$

For the true model (13.3), the slope coefficient is expressed as

$$\hat{\beta}_1 = \frac{\sum y x_1}{\sum x_1^2} \qquad \ldots (13.5)$$

which is unbiased.

For the model (13.4) that we have taken, we obtain

$$\tilde{\beta}_1 = \hat{\beta}_{1s} = \frac{(\sum yx_1)(\sum x_2^2) - (\sum yx_2)(\sum x_1x_2)}{\sum x_1^2 \sum x_2^2 - (\sum x_1x_2)^2} \quad \ldots (13.6)$$

Now the true model in deviation form is

$$y_i = \beta_1 x_1 + (u_i - \bar{u}) \quad \ldots (13.7)$$

Substituting for y_i from (13.7) into (13.6) and simplifying, we obtain

$$\mathrm{E}(\tilde{\beta}_1) = \mathrm{E}(\hat{\beta}_{1s}) = \beta_1 \frac{\sum x_1^2 \sum x_2^2 - (\sum x_1x_2)^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1x_2)^2} \quad \ldots (13.8)$$

From equation (13.8) we find that

$\mathrm{E}(\tilde{\beta}_1) = \beta_1$

Thus, inclusion of an irrelevant variable provides us with unbiased estimator of β_1.The estimator of the redundant variable $\hat{\beta}_{2s}$ is given by

$$\hat{\beta}_{2s} = \frac{(\sum yx_2)(\sum x_1^2) - (\sum yx_1)(\sum x_1x_2)}{\sum x_1^2 \sum x_2^2 - (\sum x_1x_2)^2} \quad \ldots (13.9)$$

If we substitute for y_i from (13.7) in (13.9) and re-arrange terms, we obtain

$$\mathrm{E}(\tilde{\beta}_2) = \mathrm{E}(\hat{\beta}_{2s}) = \beta_2 \frac{(\sum x_1x_2)(\sum x_1^2) - (\sum x_1x_2)(\sum x_1^2)}{\sum x_1^2 \sum x_2^2 - (\sum x_1x_2)^2} \quad \ldots (13.10)$$

Thus, $\mathrm{E}(\tilde{\beta}_2) = \mathrm{E}(\hat{\beta}_{2s}) = 0$

So, we find that $\hat{\beta}_{2s}$ which is absent from the true model has its coefficient 0. Thus we obtain unbiased estimators for both the parameters.

This leads us to conclude that inclusion of irrelevant variables is not that harmful as omission of relevant variables. As an extra variable is added to the model, we observe that there is an increase in R-squared. The variance of the parameters will not be efficient.

Therefore, the specification error in the nature of inclusion of irrelevant variables in the model, will produce unbiased but inefficient least squares estimators of the parameters. The larger variance reduces the precision of the estimates resulting in wider confidence intervals. This may lead to type II error (the error of not rejecting a null hypothesis when the alternative hypothesis is actually true).

13.3.2 Omission of Relevant Variable

Now let us look into the other side of the spectrum – excluding a relevant variable. Since a relevant variable is not included in the model (although it influences the dependent variable) its impact will be included in the residuals. As a result, the residuals will show a systematic pattern rather than being white noise as required by Gauss-Markov theorem. Also, the coefficient of the included variable will be biased.

Suppose the true equation (in deviation form) is

$$y = \beta_1 x_1 + \beta_2 x_2 + u \quad \ldots (13.11)$$

Instead of estimating equation (13.11) suppose we omitted x_2. The following equation is estimated,

$$y = \beta_1^* x_1 + e \qquad \text{… (13.12)}$$

Equation (13.12) is a case of omitted variable, and hence incorrect model specification. In the model with omitted variable (incorrect model) the estimate of β_1^* is

$$\hat{\beta}_1^* = \frac{\sum x_1 y}{\sum x_1^2} \qquad \text{… (13.13)}$$

In order to calculate the bias in the estimated value of β_1 in the incorrect model (equation (13.12)) as compared to the true model (equation (13.11)), we take the following steps:

Substituting the expression of y from the true model in (13.11), we get

$$\hat{\beta}_1^* = \frac{\sum x_1(\beta_1 x_1 + \beta_2 x_2 + u)}{\sum x_1^2} = \beta_1 + \beta_2 \frac{\sum x_1 x_2}{\sum x_1^2} + \frac{\sum x_1 u}{\sum x_1^2} \qquad \text{… (13.14)}$$

Since E $(\sum x_1 u) = 0$ we get

$$\text{E}\,(\hat{\beta}_1^*) = \beta_1 + b_{21}\beta_2 \qquad \text{… (13.15)}$$

where $b_{21} = \frac{\sum x_1 x_2}{\sum x_1^2}$ is the regression coefficient from a regression of X_2 (omitted variable) on X_1.

Thus $\hat{\beta}_1^*$ is a biased estimator for β_1 and the bias is given by

Bias = (coefficient of the excluded variable) × (regression coefficient in a regression of the excluded variable on the included variable) … (13.16)

In the deviation form, the three-variable population regression model can be written as

$$y_i = \beta_2 x_{2i} + \beta_3 x_{3i} + (u_i - \bar{u}) \qquad \text{… (13.17)}$$

First multiplying by x_2 and then by x_3, the usual normal equations are

$$\sum y_i x_{2i} = \beta_2 \sum x_{2i}^2 + \beta_3 \sum x_{2i} x_{3i} + \sum x_{2i}(u_i - \bar{u}) \qquad \text{… (13.18)}$$

$$\sum y_i x_{3i} = \beta_2 \sum x_{2i} x_{3i} + \beta_3 \sum x_{3i}^2 + \sum x_{3i}(u_i - \bar{u}) \qquad \text{… (13.19)}$$

Dividing (13.18) by $\sum x_{2i}^2$ on both sides, we obtain

$$\frac{\sum y_i x_{2i}}{\sum x_{2i}^2} = \beta_2 + \beta_3 \frac{\sum x_{2i} x_{3i}}{\sum x_{2i}^2} + \frac{\sum x_{2i}(u_i - \bar{u})}{\sum x_{2i}^2} \qquad \text{… (13.20)}$$

Thus we have

$$b_{y2} = \frac{\sum y_i x_{2i}}{\sum x_{2i}^2}$$

$$b_{32} = \frac{\sum x_{2i} x_{3i}}{\sum x_{2i}^2}$$

Hence (13.20) can be written as

$$b_{y2} = \beta_2 + \beta_3 \, b_{32} + \frac{\sum x_{2i}(u_i - \bar{u})}{\sum x_{2i}^2} \qquad \dots (13.21)$$

Taking the expected value of (13.21) we obtain

$$E(b_{y2}) = \beta_2 + \beta_3 \, b_{32} \qquad \dots (13.22)$$

Similarly, if x_2 is omitted from the model, the bias in $E(b_{y3})$ can be calculated.

The variance of β_1^* (parameter of the incorrect model) can also be derived by using the formula for variance. As it is a bit complex, we do not present it here. You should note that the variance of β_1^* is higher than that of β_1. An implication of the above is that usual tests of significance concerning parameters are invalid, if some of the relevant variables are excluded from a model.

Thus we know that

(i) When an irrelevant variable is included in the model: (a) the estimators of parameters are unbiased, (b) efficiency of the estimators decline, and (c) estimator of the error variance is unbiased. Thus conventional tests of hypothesis are valid. The inferences drawn could be somewhat erroneous.

(ii) When a relevant variable is dropped from the model: (a) estimators of parameters are biased, (b) efficiency of estimators decline, and (c) estimator of error variance is biased. Thus conventional tests of hypothesis are invalid. The inferences drawn are faulty.

13.3.3 Incorrect Functional Form

Apart from inclusion of only relevant variables in an econometric model, another specification error pertains to functional form. There is a tendency the part of researchers to assume a linear relationship between variables. This however is not always true. If the true relationship is non-linear and we take a linear regression model for estimation, we will not be able to draw correct inferences. There are test statistics available to choose among functional forms. We will discuss these test statistics in Unit 14.

Check Your Progress 2

1) Explain the consequences of inclusion of an irrelevant variable.

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

…………………………………………………………………………....

2) Explain the consequences of excluding a relevant variable.

....

....

....

....

....

....

13.4 ERROR OF MEASUREMENT IN VARIABLES

So far we have assumed the variables in the econometric model under study are measured correctly. It means that there are no measurement errors in both explained and explanatory variables. Sometimes we do not have data on the variables that we want to use in the model. This could be for various reasons such as non-response error, reporting error, and computing error. A classic example of measurement error pertains to the variable permanent income used in the Milton Friedman model. Measurement error in variables is a serious problem in econometric studies. There are two types of measurement errors:

(i) Measurement error in dependent variable, and

(ii) Measurement error in independent variable.

13.4.1 Measurement Error in Dependent Variable

Let us consider the following model:

$$Y_i^* = \alpha + \beta X_i + u_i \qquad \dots (13.23)$$

where Y_i^* is permanent consumption expenditure

X_i is current income, and

u_i is the stochastic disturbance term.

(we place a star mark (*) on the variable that is measured with errors)

Since Y_i^* is not directly measureable, we may use an observable expenditure variable Y_i such that

$$Y_i = Y_i^* + e_i \qquad \dots (13.24)$$

where e_i denote measurement error in Y_i^*.

Therefore, instead of estimating

$Y_i^* = \alpha + \beta X_i + u_i$, we estimate

$$Y_i = \alpha + \beta X_i + u_i + e_i$$

$$= \alpha + \beta X_i + (u_i + e_i)$$

Let us re-write the above equation as

$$Y_i = \alpha + \beta X_i + v_i \quad \ldots (13.25)$$

where $v_i = u_i + e_i$

In equation (13.25) we take v_i as a composite error term comprising population disturbance term (u_i) and measurement error term (e_i).

Let us assume that the following classical assumptions hold

a) E(u_i) = E(e_i) = 0

b) Cov (X_i, u_i) = 0

c) Cov (u_i, e_i) = 0

An implication of (c) above is that the stochastic error term and the measurement error term are uncorrelated. Thus expected value of the composite error term is zero; $E(v) = 0$. By extending the logic given in Unit 4, we can say that $E(\hat{\beta}) = \beta$. It implies that $\hat{\beta}$ is *unbiased*.

Now let us look into the issue of variance in the case of measurement error in the dependent variable. As you know, variance of the estimator $\hat{\beta}$ in a two variable regression model (13.23) is given by

$$\text{Var}(\hat{\beta}) = \frac{\sigma_u^2}{\sum x_i^2},$$

For the composite error term, this will translate into

$$\text{Var}(\hat{\beta}) = \frac{\sigma_u^2 + \sigma_e^2}{\sum x_i^2} = \frac{\sigma_v^2}{\sum x_i^2} \quad \ldots (13.26)$$

Thus we see that the variance of the error term is larger if there is measurement error in the dependent variable. This leads to inefficiency of the estimators. They are not best linear unbiased estimators (BLUE).

13.4.2 Measurement Error in Independent Variable

There could be measurement error in explanatory variables. Let us assume the true regression model to be estimated is

$$Y_i = \alpha + \beta X_i^* + u_i \quad \ldots (13.27)$$

Suppose we do not have data on variable X_i^*. On the other hand, suppose we have data on X_i. In that case, instead of observing X_i^*, we observe

$$X_i = X_i^* + w_i \quad \ldots (13.28)$$

where w_i represents error of measurement in X_i^*.

In the permanent income hypothesis model, for example,

$$Y_i = \alpha + \beta X_i^* + u_i$$

where Y_i is current consumption expenditure

X_i^* is permanent income

u_i is stochastic disturbance term (equation error)

From equation (13.27) and (13.28) we find that

$$Y_i = \alpha + \beta(X_i - w_i) + u_i \qquad \ldots (13.29)$$

$$= \alpha + \beta X_i + (u_i - \beta w_i)$$

$$= \alpha + \beta X_i + z_i \qquad \ldots (13.30)$$

where $z_i = (u_i - \beta w_i)$. You should notice that z_i is made up of two terms: stochastic error and measurement error.

Now, let us assume that w_i has zero mean; it is serially independent; and it is uncorrelated with u_i. Even in that case, the composite error term z_i is not independent of the explanatory variable X_i.

$$\text{Cov}\,(z_i, X_i) = \text{E}[z_i - E(z_i)][X_i - E(X_i)]$$

$$= \text{E}(u_i - \beta w_i)(w_i)$$

$$= \text{E}(-\beta w_i^2)$$

$$= -\beta\sigma_w^2 \qquad \ldots (13.31)$$

From (13.31) we find that the independent variable and the error term are correlated. This violates the basic assumption of the classical regression model that the explanatory variable is uncorrelated with the stochastic disturbance term. In such a situation the OLS estimators are not only biased but also inconsistent, that is they remain biased even if the sample size *n* increases infinitely.

Check Your Progress 3

1) Explain the consequences measurement error in the dependent variable.

....

....

....

....

....

....

2) Explain the consequences of measurement error in the explanatory variable.

....

....

....

....

....

....

3) Measurement error in the dependent variable is a lesser evil than measurement error in the explanatory variable.

....

....

....

....

....

....

13.5 LET US SUM UP

Correct specification of an econometric model determines the accuracy of the estimates obtained. Therefore, correct specification of an econometric model is very important. Economic theory and logic guide us in specification of econometric models.

In order to correctly specify an econometric model all relevant explanatory variables should be included in the model. No relevant explanatory variable should be excluded from the model. Further, the functional form of the model should be correct.

At times we do not get appropriate variable required in an econometric model. In such cases there could be cases where either dependent variable or independent variable is measured with certain error. Measurement error in dependent variable is a lesser evil than the measurement error in the independent variable.

13.6 ANSWERS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) The basic assumptions of the classical regression model are as follows:

   a) The regression model is linear in parameters

   b) $E(X_i u_i) = 0$ (regressor is non-stochastic)

   c) $E(u_i) = 0$

   d) $E(u_i)^2 = \sigma^2$

   e) $E(u_i u_j) = 0$ for $i \neq j$

   f) The explanatory variables (X_i) are independent of one another.

2) Go through Section 13.2. It is important because incorrect specification has serious implications on desirable properties of the estimators.

3) Go through Sub-Section 13.2.2 and answer.

4) The important specific issues are: inclusion of irrelevant/redundant variables; omission of relevant variables; and incorrect functional form of the model

Check Your Progress 2

1) The estimator is unbiased but inefficient. See Sub-Section 13.3.1.

2) The estimator is biased as well as inefficient. See Sub-Section 13.3.2.

Check Your Progress 3

1) Go through Sub-Section 13.4.1 and answer.

2) Go through Sub-Section 13.4.2 and answer.

3) If there is measurement error in dependent variable the estimator is unbiased but inefficient. Measurement error in explanatory variable results in biased estimator. See Section 13.4 for details.

UNIT 14 TESTS FOR SPECIFICATION ERROR*

Structure

14.1 INTRODUCTION

In the previous Unit we highlighted the consequences of specification errors. There could be three types of specification errors; inclusion of an irrelevant variable, exclusion of a relevant variable, and incorrect functional form. When the econometric model is not specified correctly, the coefficient estimates, the confidence intervals, and the hypothesis tests are misleading and inconsistent. In view of this, econometric models should be correctly specified.

While building a model we face a lot of difficulties in specifying a model correctly. In some cases economic theory is quite transparent about the dependent variables and the independent variables. In some other cases still it is in a hypothesis stage. Researchers are still working in that area to confirm the hypothesis suggested by others. In such cases, what we have a dependent variable and a set of explanatory variables. Out of these explanatory variables we have to select the most appropriate ones.

* Dr. Sahba Fatima, Independent Researcher, Lucknow.

Econometric theory suggests certain criteria and test statistics. On the basis of these criteria we select the most appropriate econometric model. We describe some of these criteria below.

14.2 OBJECTIVES

After going through this Unit, you should be in a position to

- identify econometric models that are not specified correctly;
- take remedial measures for correcting the specification error; and
- evaluate the performance of competing models.

14.3 TESTS FOR IDENTIFYING THE MOST EFFICIENT MODEL

As pointed out above, econometric models should be specified correctly. Any spurious relationship should be identified and excluded from the model. There are certain tests for this purpose. These tests can be used under specific circumstances in conjunction with practical understanding of the variables and an enlightened study of it through the related literature. Following tests are most commonly used for model testing and evaluation.

14.3.1 The R^2 Test and Adjusted-R^2 Test

We have discussed the concept of coefficient of determination (R^2) in Unit 4. As you know, the coefficient of determination indicates the explanatory power of a model. If, for example, $R^2 = 0.76$ we can infer that 76 per cent variation in the dependent variable is explained by the explanatory variable in the model.

We define R^2 as follows:

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS} \qquad \text{... (14.1)}$$

where TSS = Total Sum of Squares

ESS = Explained Sum of squares

RSS = Residual Sum of Squares

As you know,

$$TSS = RSS + ESS \qquad \text{... (14.2)}$$

Dividing both sides of equation (14.2) by TSS, we find that

$$\frac{RSS}{TSS} + \frac{ESS}{TSS} = 1 \qquad \text{... (14.3)}$$

Since $R^2 = \frac{ESS}{TSS}$, we observe that R^2 lies between 0 and 1 necessarily. Its closeness to 1 indicates better fit of the model. If R^2 is close to one, RSS is much smaller compared to ESS. Therefore, very little residual will be left. Thus a model with higher R^2 is preferred. You should however keep in mind that a very high R^2 indicates the presence of multicollinearity in the model. If the R^2 is high but the t-ratio of the coefficients are not statistically significant you should check for multicollinearity. The R^2 is calculated on the basis of the sample data.

Thus the explanatory variables included the model are considered for estimation of R^2. Variables not included in the model do not account for the variation in the dependent variable.

There is a tendency of the R^2 to increase if more explanatory variables are added. Thus, we are tempted to add more explanatory variables to increase the explanatory power of the model. If we add irrelevant explanatory variables in a model, the estimators are unbiased, but there is an increase in the variance of the estimators. This makes forecast and analysis on the basis of such models unreliable.

In order to overcome this difficulty, we use the 'adjusted-R^2'. It is denoted by $\bar{R}^2$ and defined as follows:

$$\bar{R}^2 = 1 - \frac{ESS/(n-k)}{TSS/(n-1)} = 1 - (1 - R^2)\frac{n-1}{n-k} \qquad \ldots (14.4)$$

where n is the number of observations and k is the number of regressors. As you know the TSS has a degree of freedom of $(n-1)$ while the ESS has a degree of freedom of $(n-k)$. Thus, $\bar{R}^2$ takes into account the degrees of freedom of the model. The $\bar{R}^2$ penalises the addition of explanatory variables. It is observed that there is an increase in $\bar{R}^2$ only if the t-value (absolute number) of the additional explanatory variable is greater than 1. Hence, superfluous variables can be identified and eliminated from the model. The restriction here is to regress all the independent variable against the same dependent variable.

Remember that we can compare the $\bar{R}^2$ of two models only if the dependent variable is the same. For example, we cannot compare two models if in one model the explanatory variable is Y and in the other model the explanatory variable in *log*Y.

14.3.2 Akaike Information Criterion (AIC)

Another method for identifying the mis-specification in a model is Akaike Information Criterion (AIC). This method also penalises the addition of regressors as we can see from the formula below:

$$AIC = e^{2k/n} \sum \frac{\hat{u}_i^2}{n} = e^{2k/n} \frac{RSS}{n} \qquad \ldots (14.5)$$

where k is the number of regressors (explanatory variables) and n is the number of observations.

We can further simplify equation (14.5) as

$$\ln AIC = \left(\frac{2k}{n}\right) + \ln\left(\frac{RSS}{n}\right) \qquad \ldots (14.6)$$

where $\ln AIC$ is the natural log of AIC, and $\frac{2k}{n}$ is the penalty factor.

Remember that the model with a lower value of $\ln AIC$ is considered to be better. Thus, when we compare two models by using the AIC criterion, the model with lower value of AIC has a better specification. The logic is simple. An econometric model that reduces the residual sum of squares is a better specified model.

14.3.3 Schwarz Information Criterion

The Schwarz Information Criterion (SIC) also relies on the RSS, like the AIC criterion mentioned above. This method also is popular for analysing correct specification of an econometric model. The SIC is defined as follows:

$$SIC = n^{k/n}\frac{\sum \hat{u}^2}{n} = n^{k/n}\frac{RSS}{n} \quad \dots (14.7)$$

If we take in log-form, equation (14.7) is given as

$$\ln SIC = \frac{k}{n}\ln n + \ln\left(\frac{RSS}{n}\right) \quad \dots (14.8)$$

where $[(k/n)\ln n]$ is the penalty factor. Note that the SIC criterion imposes a harsher penalty for inclusion of explanatory variable compared to the AIC criterion.

14.3.4 Mallow's C_p Criterion

When we do not include all the relevant variables in a model, the estimators are biased. The Mallow's C_p Criterion evaluates such bias to find out whether there is significant deviation from the unbiased estimators. Thus, the Mallow's C_p Criterion helps us in selecting the best among competing econometric models.

If some of the explanatory variables are dropped from a model, there is an increase in the residual sum of squares (RSS). Let us assume that the true model has k regressors. For this model, $\hat{\sigma}^2$ is the estimator of true σ^2. Now, suppose we drop p regressors from the model. The residual sum of squares obtained from the truncated model is RSS_p. The Mallow's C_p Criterion is based on the following formula:

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} - (n - 2p) \quad \dots (14.9)$$

where n is the number of observations.

While choosing a model according to the C_p criterion, the model with the lowest C_p value is preferred.

14.4 CAUTION ABOUT MODEL SELECTION CRITERIA

We have emphasized earlier that econometric models should be based on economic theory and logic. Therefore, while constricting an econometric model, you should go by the theoretical appropriateness of including or excluding a variable. In order to have a correctly specified model, a thorough understanding of the theoretical concepts and the related literature is necessary. Also, the model that we fit will only be as good as the data that we have collected. If the data collected does not suffer from, say, multicollinearity or autocorrelation, we are likely to have a more robust model.

As mentioned earlier, the criteria for selecting an appropriate model primarily rests on the theory behind it and the strength of the collected data. Many a time, we observe certain relationship between two variables. Such relationship however may be superficial or spurious. Let us take an example. At a traffic light, cars stop when the signal is red. It does not mean that cars cannot move when there is red light in front of them. It also does not mean that traffic light has some damaging effect on moving cars. The reason is observance of traffic rules. Unless we look into the traffic rules and go by observation only, our reasoning will be wrong. The dependent variable and the independent variable both may be affected by another variable. In such cases the relationship is confounded.

You should note one more issue regarding selection of econometric models. Different test criteria may suggest different models. For example, economic logi suggests that there could two possible econometric models (say, model A and model B) for a particular issue. You may come across a situation such that $\bar{R}^2$ test suggests model A and AIC criterion suggest model B. In such situations you should carry out a number of tests and then only chose the best model.

Adjusted R-squared, Mallows C_p, p-values, etc. may point to different regression equations without much clarity to the econometrician. Thus, we conclude that none of the methods for model selection listed above are adequate by itself. There is no substitute to theoretical understanding of the related literature, accurately collected data, practical understanding of the problem, and common sense while specifying an econometric model. We will discuss further on the model selection criteria in the course BECC 142: Applied Econometrics.

Check Your Progress 1

1) Explain why $\bar{R}^2$ is a better criterion than R^2 in model specification.

....

....

....

....

....

....

2) Explain how the AIC and BIC criteria are applied in selection of econometric models.

....

....

....

....

....

3) What precaution you should take while selecting an econometric model?

....

....

....

....

....

14.5 LET US SUM UP

Selection of an appropriate econometric model is a difficult task. We have to take into account the economic theory and logic behind the econometric model. There could be many competing models for a particular issue.

There a certain criteria on the basis of which the best econometric model is selected. These criteria could be $\bar{R}^2$, AIC, BIC, and Mallow's C_p. We have described the formulae for these test criteria in the Unit.

14.6 ANSWERS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) In Sub-Section 14.3.1 we have compared between R^2 and $\bar{R}^2$. The $\bar{R}^2$ takes into account the degrees of freedom.
2) You should describe the test statistics used in AIC and BIC criteria (see Section 14.3). The model with lowest value of test statistics is preferred.
3) Go through Section 14.4 and answer.

Table A1: Normal Area Table

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Table A2: Critical Values of Chi-squared Distribution

df\area	0.1	0.05	0.025	0.01	0.005
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.071	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188
11	17.275	19.675	21.920	24.725	26.757
12	18.549	21.026	23.337	26.217	28.300
13	19.812	22.362	24.736	27.688	29.819
14	21.064	23.685	26.119	29.141	31.319
15	22.307	24.996	27.488	30.578	32.801
16	23.542	26.296	28.845	32.000	34.267
17	24.769	27.587	30.191	33.409	35.718
18	25.989	28.869	31.526	34.805	37.156
19	27.204	30.144	32.852	36.191	38.582
20	28.412	31.410	34.170	37.566	39.997
21	29.615	32.671	35.479	38.932	41.401
22	30.813	33.924	36.781	40.289	42.796
23	32.007	35.172	38.076	41.638	44.181
24	33.196	36.415	39.364	42.980	45.559
25	34.382	37.652	40.646	44.314	46.928
26	35.563	38.885	41.923	45.642	48.290
27	36.741	40.113	43.195	46.963	49.645
28	37.916	41.337	44.461	48.278	50.993
29	39.087	42.557	45.722	49.588	52.336
30	40.256	43.773	46.979	50.892	53.672

Table A3: Critical Values of *t* Distribution

Df\p	0.25	0.10	0.05	0.025	0.01	0.005
1	1.0000	3.0777	6.3138	12.7062	31.8205	63.6567
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.7649	1.6377	2.3534	3.1825	4.5407	5.8409
4	0.7407	1.5332	2.1318	2.7765	3.7470	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0321
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9980	3.4995
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.6955	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.6912	1.3406	1.7531	2.1315	2.6025	2.9467
16	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.6892	1.3334	1.7396	2.1098	2.5669	2.8982
18	0.6884	1.3304	1.7341	2.1009	2.5524	2.8784
19	0.6876	1.3277	1.7291	2.0930	2.5395	2.8609
20	0.6870	1.3253	1.7247	2.0860	2.5280	2.8453
20	0.6870	1.3253	1.7247	2.0860	2.5280	2.8453
21	0.6864	1.3232	1.7207	2.0796	2.5177	2.8314
22	0.6858	1.3212	1.7171	2.0739	2.5083	2.8188
23	0.6853	1.3195	1.7139	2.0687	2.4999	2.8073
24	0.6849	1.3178	1.7109	2.0639	2.4922	2.7969
25	0.6844	1.3163	1.7081	2.0595	2.4851	2.7874
26	0.6840	1.3150	1.7056	2.0555	2.4786	2.7787
27	0.6837	1.3137	1.7033	2.0518	2.4727	2.7707
28	0.6834	1.3125	1.7011	2.0484	2.4671	2.7633
29	0.6830	1.3114	1.6991	2.0452	2.4620	2.7564
30	0.6828	1.3104	1.6973	2.0423	2.4573	2.7500
inf	0.6745	1.2816	1.6449	1.9600	2.3264	2.5758

Table A4: Critical Values of *F* Distribution

(5% level of significance)

df2/df1	1	2	3	4	5	6	7	8	9	10
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396
3	10.128	9.552	9.277	9.117	9.014	8.941	8.887	8.845	8.812	8.786
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
5	6.608	5.786	5.410	5.192	5.050	4.950	4.876	4.818	4.773	4.735
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
8	5.318	4.459	4.066	3.838	3.688	3.581	3.501	3.438	3.388	3.347
9	5.117	4.257	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
10	4.965	4.103	3.708	3.478	3.326	3.217	3.136	3.072	3.020	2.978
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
15	4.543	3.682	3.287	3.056	2.901	2.791	2.707	2.641	2.588	2.544
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
21	4.325	3.467	3.073	2.840	2.685	2.573	2.488	2.421	2.366	2.321
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.237
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.266	2.220
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165
40	4.085	3.232	2.839	2.606	2.450	2.336	2.249	2.180	2.124	2.077
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993
120	3.920	3.072	2.680	2.447	2.290	2.175	2.087	2.016	1.959	1.911
inf	3.842	2.996	2.605	2.372	2.214	2.099	2.010	1.938	1.880	1.831

Table A4: Critical Values of *F* Distribution (Contd.)

(5% level of significance)

df2/df1	12	15	20	24	30	40	60	120	INF
1	243.906	245.950	248.013	249.052	250.095	251.143	252.196	253.253	254.314
2	19.413	19.429	19.446	19.454	19.462	19.471	19.479	19.487	19.496
3	8.745	8.703	8.660	8.639	8.617	8.594	8.572	8.549	8.526
4	5.912	5.858	5.803	5.774	5.746	5.717	5.688	5.658	5.628
5	4.678	4.619	4.558	4.527	4.496	4.464	4.431	4.399	4.365
6	4.000	3.938	3.874	3.842	3.808	3.774	3.740	3.705	3.669
7	3.575	3.511	3.445	3.411	3.376	3.340	3.304	3.267	3.230
8	3.284	3.218	3.150	3.115	3.079	3.043	3.005	2.967	2.928
9	3.073	3.006	2.937	2.901	2.864	2.826	2.787	2.748	2.707
10	2.913	2.845	2.774	2.737	2.700	2.661	2.621	2.580	2.538
11	2.788	2.719	2.646	2.609	2.571	2.531	2.490	2.448	2.405
12	2.687	2.617	2.544	2.506	2.466	2.426	2.384	2.341	2.296
13	2.604	2.533	2.459	2.420	2.380	2.339	2.297	2.252	2.206
14	2.534	2.463	2.388	2.349	2.308	2.266	2.223	2.178	2.131
15	2.475	2.403	2.328	2.288	2.247	2.204	2.160	2.114	2.066
16	2.425	2.352	2.276	2.235	2.194	2.151	2.106	2.059	2.010
17	2.381	2.308	2.230	2.190	2.148	2.104	2.058	2.011	1.960
18	2.342	2.269	2.191	2.150	2.107	2.063	2.017	1.968	1.917
19	2.308	2.234	2.156	2.114	2.071	2.026	1.980	1.930	1.878
20	2.278	2.203	2.124	2.083	2.039	1.994	1.946	1.896	1.843
21	2.250	2.176	2.096	2.054	2.010	1.965	1.917	1.866	1.812
22	2.226	2.151	2.071	2.028	1.984	1.938	1.889	1.838	1.783
23	2.204	2.128	2.048	2.005	1.961	1.914	1.865	1.813	1.757
24	2.183	2.108	2.027	1.984	1.939	1.892	1.842	1.790	1.733
25	2.165	2.089	2.008	1.964	1.919	1.872	1.822	1.768	1.711
26	2.148	2.072	1.990	1.946	1.901	1.853	1.803	1.749	1.691
27	2.132	2.056	1.974	1.930	1.884	1.836	1.785	1.731	1.672
28	2.118	2.041	1.959	1.915	1.869	1.820	1.769	1.714	1.654
29	2.105	2.028	1.945	1.901	1.854	1.806	1.754	1.698	1.638
30	2.092	2.015	1.932	1.887	1.841	1.792	1.740	1.684	1.622
40	2.004	1.925	1.839	1.793	1.744	1.693	1.637	1.577	1.509
60	1.917	1.836	1.748	1.700	1.649	1.594	1.534	1.467	1.389
120	1.834	1.751	1.659	1.608	1.554	1.495	1.429	1.352	1.254
inf	1.752	1.666	1.571	1.517	1.459	1.394	1.318	1.221	1.000

Table A4: Critical Values of *F* Distribution (contd.)

(1% level of significance)

df2/df1	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	4052.181	4999.500	5403.352	5624.583	5763.650	5858.986	5928.356	5981.070	6022.473	6055.847
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388	99.399
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345	27.229
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051
6	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874
7	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814
9	10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398	3.310
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299	3.211
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129
26	7.721	5.526	4.637	4.140	3.818	3.591	3.421	3.288	3.182	3.094
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149	3.062
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120	3.032
29	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198	3.092	3.005
30	7.562	5.390	4.510	4.018	3.699	3.473	3.304	3.173	3.067	2.979
40	7.314	5.179	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632
120	6.851	4.787	3.949	3.480	3.174	2.956	2.792	2.663	2.559	2.472
inf	6.635	4.605	3.782	3.319	3.017	2.802	2.639	2.511	2.407	2.321

Table A4: Critical Values of *F* Distribution (contd.)

(1% level of significance)

df2/df1	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**INF**
1	6106.321	6157.285	6208.730	6234.631	6260.649	6286.782	6313.030	6339.391	6365.864
2	99.416	99.433	99.449	99.458	99.466	99.474	99.482	99.491	99.499
3	27.052	26.872	26.690	26.598	26.505	26.411	26.316	26.221	26.125
4	14.374	14.198	14.020	13.929	13.838	13.745	13.652	13.558	13.463
5	9.888	9.722	9.553	9.466	9.379	9.291	9.202	9.112	9.020
6	7.718	7.559	7.396	7.313	7.229	7.143	7.057	6.969	6.880
7	6.469	6.314	6.155	6.074	5.992	5.908	5.824	5.737	5.650
8	5.667	5.515	5.359	5.279	5.198	5.116	5.032	4.946	4.859
9	5.111	4.962	4.808	4.729	4.649	4.567	4.483	4.398	4.311
10	4.706	4.558	4.405	4.327	4.247	4.165	4.082	3.996	3.909
11	4.397	4.251	4.099	4.021	3.941	3.860	3.776	3.690	3.602
12	4.155	4.010	3.858	3.780	3.701	3.619	3.535	3.449	3.361
13	3.960	3.815	3.665	3.587	3.507	3.425	3.341	3.255	3.165
14	3.800	3.656	3.505	3.427	3.348	3.266	3.181	3.094	3.004
15	3.666	3.522	3.372	3.294	3.214	3.132	3.047	2.959	2.868
16	3.553	3.409	3.259	3.181	3.101	3.018	2.933	2.845	2.753
17	3.455	3.312	3.162	3.084	3.003	2.920	2.835	2.746	2.653
18	3.371	3.227	3.077	2.999	2.919	2.835	2.749	2.660	2.566
19	3.297	3.153	3.003	2.925	2.844	2.761	2.674	2.584	2.489
20	3.231	3.088	2.938	2.859	2.778	2.695	2.608	2.517	2.421
21	3.173	3.030	2.880	2.801	2.720	2.636	2.548	2.457	2.360
22	3.121	2.978	2.827	2.749	2.667	2.583	2.495	2.403	2.305
23	3.074	2.931	2.781	2.702	2.620	2.535	2.447	2.354	2.256
24	3.032	2.889	2.738	2.659	2.577	2.492	2.403	2.310	2.211
25	2.993	2.850	2.699	2.620	2.538	2.453	2.364	2.270	2.169
26	2.958	2.815	2.664	2.585	2.503	2.417	2.327	2.233	2.131
27	2.926	2.783	2.632	2.552	2.470	2.384	2.294	2.198	2.097
28	2.896	2.753	2.602	2.522	2.440	2.354	2.263	2.167	2.064
29	2.868	2.726	2.574	2.495	2.412	2.325	2.234	2.138	2.034
30	2.843	2.700	2.549	2.469	2.386	2.299	2.208	2.111	2.006
40	2.665	2.522	2.369	2.288	2.203	2.114	2.019	1.917	1.805
60	2.496	2.352	2.198	2.115	2.028	1.936	1.836	1.726	1.601
120	2.336	2.192	2.035	1.950	1.860	1.763	1.656	1.533	1.381
inf	2.185	2.039	1.878	1.791	1.696	1.592	1.473	1.325	1.000

Table A5: Durbin-Watson d-statistic   Level of Significance = 0.05   k= no. of regressors

	k=1		k=2		k=3		k=4	
n	dL	dU	dL	dU	dL	dU	dL	dU
6	0.6102	1.4002						
7	0.6996	1.3564	0.4672	1.8964				
8	0.7629	1.3324	0.5591	1.7771	0.3674	2.2866		
9	0.8243	1.3199	0.6291	1.6993	0.4548	2.1282	0.2957	2.5881
10	0.8791	1.3197	0.6972	1.6413	0.5253	2.0163	0.3760	2.4137
11	0.9273	1.3241	0.7580	1.6044	0.5948	1.9280	0.4441	2.2833
12	0.9708	1.3314	0.8122	1.5794	0.6577	1.8640	0.5120	2.1766
13	1.0097	1.3404	0.8612	1.5621	0.7147	1.8159	0.5745	2.0943
14	1.0450	1.3503	0.9054	1.5507	0.7667	1.7788	0.6321	2.0296
15	1.0770	1.3605	0.9455	1.5432	0.8140	1.7501	0.6852	1.9774
16	1.1062	1.3709	0.9820	1.5386	0.8572	1.7277	0.7340	1.9351
17	1.1330	1.3812	1.0154	1.5361	0.8968	1.7101	0.7790	1.9005
18	1.1576	1.3913	1.0461	1.5353	0.9331	1.6961	0.8204	1.8719
19	1.1804	1.4012	1.0743	1.5355	0.9666	1.6851	0.8588	1.8482
20	1.2015	1.4107	1.1004	1.5367	0.9976	1.6763	0.8943	1.8283
21	1.2212	1.4200	1.1246	1.5385	1.0262	1.6694	0.9272	1.8116
22	1.2395	1.4289	1.1471	1.5408	1.0529	1.6640	0.9578	1.7974
23	1.2567	1.4375	1.1682	1.5435	1.0778	1.6597	0.9864	1.7855
24	1.2728	1.4458	1.1878	1.5464	1.1010	1.6565	1.0131	1.7753
25	1.2879	1.4537	1.2063	1.5495	1.1228	1.6540	1.0381	1.7666
26	1.3022	1.4614	1.2236	1.5528	1.1432	1.6523	1.0616	1.7591
27	1.3157	1.4688	1.2399	1.5562	1.1624	1.6510	1.0836	1.7527
28	1.3284	1.4759	1.2553	1.5596	1.1805	1.6503	1.1044	1.7473
29	1.3405	1.4828	1.2699	1.5631	1.1976	1.6499	1.1241	1.7426
30	1.3520	1.4894	1.2837	1.5666	1.2138	1.6498	1.1426	1.7386
31	1.3630	1.4957	1.2969	1.5701	1.2292	1.6500	1.1602	1.7352
32	1.3734	1.5019	1.3093	1.5736	1.2437	1.6505	1.1769	1.7323
33	1.3834	1.5078	1.3212	1.5770	1.2576	1.6511	1.1927	1.7298
34	1.3929	1.5136	1.3325	1.5805	1.2707	1.6519	1.2078	1.7277
35	1.4019	1.5191	1.3433	1.5838	1.2833	1.6528	1.2221	1.7259
36	1.4107	1.5245	1.3537	1.5872	1.2953	1.6539	1.2358	1.7245
37	1.4190	1.5297	1.3635	1.5904	1.3068	1.6550	1.2489	1.7233
38	1.4270	1.5348	1.3730	1.5937	1.3177	1.6563	1.2614	1.7223
39	1.4347	1.5396	1.3821	1.5969	1.3283	1.6575	1.2734	1.7215
40	1.4421	1.5444	1.3908	1.6000	1.3384	1.6589	1.2848	1.7209
41	1.4493	1.5490	1.3992	1.6031	1.3480	1.6603	1.2958	1.7205

GLOSSARY

Association	:	It refers to the connection or relationship between variables
Alternative Hypothesis	:	It is the hypothesis contrary to the null hypothesis. Null hypothesis and alternative hypothesis are mutually exclusive.
Alternative Hypothesis	:	In hypothesis testing, alternative hypothesis states a condition that is opposite to the null hypothesis. It is expressed as $H_1: \beta_2 \neq 0$, i.e., the slope coefficient is different from zero. It could be positive or negative.
Analysis of Variance (ANOVA)	:	This is a technique that breaks up the total variability of data into two parts one statistical and the other random.
ANCOVA Model	:	This is a model which involves both a quantitative and a dummy variable. The form of such a model will be like: $Y_i = \beta_1 + \beta_2 D + \beta_3 X_i + u_i$.
ANOVA Model	:	This is a regression model containing only a dummy explanatory variable. The functional form of this is like: $Y_i = \beta_1 + \beta_2 D_i + \mu_i$.
Autocorrelation	:	The Classical Linear Regression Model assumes that the random error terms are not related to each other. In other words, there exists no correlation between the error terms associated with each observation. This assumption is referred as the assumption of no autocorrelation.
Base or Benchmark Category	:	The dummy variable which takes the value 0 is referred to as the 'base or benchmark category'.
Continuous Random Variable	:	It refers to a random variable that can take infinite number of values in an interval are called continuous random variables.
Cochrane-Orcutt Procedure	:	This is a transformation procedure suggested by Cochrane-Orcutt. It is helpful in estimating the value of the correlation coefficient between the error terms. The transformation, enables the application of the OLS method, and yields estimates of parameters which enjoy the BLUE property.

Confidence Interval Approach : In order to test the population parameter, a confidence interval can be constructed about the true but unknown mean. If the population parameter lies within the confidence interval, the null hypothesis is accepted; otherwise it is rejected.

Classical Linear Regression Model : It refers to a linear regression model that establishes a linear relationship between the variables, based on certain specified assumptions.

Chow Test : This test visualizes the presence of structural change that may result in differences in the intercept or the slope coefficient or both. This in referred to as parameter instability. For examining this we perform Chow Test

Causal Relationship : The relationship between the variables where one can figure out the cause and the effect between the two variables.

Confidence Interval : It is the range of values that determines the probability that the value of the parameter lies within the interval.

Chi-square Distribution : Chi-square distribution is the distribution which is the sum of squares of k independent standard normal random variables.

Composite or Two-Sided Hypothesis : In hypothesis testing, a composite hypothesis covers a set of values that are not equal to the given or stated null hypothesis.

Confidence Interval : It refers to the probability that a population parameter falls within the set of critical values taken from the Table.

Discrete Random Variable : It refers to random variables that can assume only countable values.

Distribution Function : Distribution function of a real valued random variable gives a value at any given sample point in the sample space.

Deterministic Component : It represents the systematic component of the regression equation. It is the expected value of the dependent variable for given values of the explanatory variable.

Econometric Model : These are statistical models specifying relationship between relationships between various economic quantities.

Differential Intercept Coefficient : In the ANOVA model $Y_i = \beta_1 + \beta_2 D_i + \mu_i$, since there is no continuous regression line involved, the slope coefficient β_2 actually measures by how much the value of the intercept term differs between the two categories (e.g. male/female) under consideration. For this reason, β_2 is more appropriately called as the 'differential intercept coefficient'.

Dummy Variable Trap : Response to a dummy variable like gender (male/female), caste (general/SC-ST/OBC), etc. are called as categories. Depending on the 'number' of such categories, we must consider including the number of dummy variables in the regression carefully. Usually, this should be 'one less than the number of categories'. Failing to do this will land us in a situation called as the 'dummy variable trap'. This means we will face a situation of multicollinearity with no unique estimates, or efficient estimates, of the parameters. The general rule for introducing the number of dummies is that, if there are *m* attributes or categories, the number of dummy variables introduced should be '$m - 1$'.

Dummy Variables : There are variables which are qualitative in nature. Also known as dummy variables, these variables are referred differently like: indicator variables, binary variables, categorical variables, dichotomous variables.

Durbin *h*-statistic : The Durbin- Watson technique fails to operate when the regression model involves the lagged value of dependent variable as one of the explanatory variables. In such models, the h – statistic, also suggested by Durbin, is useful to identify the presence of autocorrelation in the regression model.

Durbin-Watson Test (*d*-statistic) : The test helps detect a first order autocorrelation. The test statistic employed is:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

Estimator	:	A method of arriving at an *estimate* of a parameter.
Estimation of Parameters	:	This process deals with estimating the values of parameters based on measured empirical data that has a random component.
Estimation	:	The process of estimating any population parameter.
F-Distribution	:	It is a right-skewed distribution used for analysis of variance. F-statistic is used for comparing statistical models and to identify the model that best fits the population.
Forecasting	:	Forecasting is a technique that predicts the future trends by using historical data. The method of forecasting is generally used to extrapolate the parameters such as GDP or unemployment.
Goodness of Fit	:	An overall goodness of fit that tells us how well the estimated regression line fits the actual Y values. Such a measure is known as the coefficient of determination, denoted by R^2. It is the ratio of explained sum of squares (ESS) to total sum of squares (TSS).
Glejser Test	:	The Glejser Test is similar to the Park Test. Obtaining e_i from the original model, Glejser suggests regressing the absolute values of e_i, i.e., $\lvert e_i \rvert$ on the X variable expected to be closely associated with the heteroscedastic variance σ_i^2.
Goldfeld-Quandt Test	:	In this method of testing for heteroscedasticity, we first arrange the observations in increasing order of X_i variable. Next we exclude C observations in the middle of dataset. Thus, $(n - C)/2$ observations in the first part and $(n - C)/2$ observations in the last part constitute two groups. We then proceed to obtain the respective residual sum of squares RSS_1 and RSS_2. The RSS_1 represents the RSS for the regression corresponding to the smaller X_i values and RSS_2 to that of the larger X_i values. We conduct F-test to check for the presence of heteroscedasticity.
Gauss Markov Theorem	:	Under the assumptions of classical linear regression model, the least squares estimators are Best Linear Unbiased Estimate (BLUE). This means, in the class of all unbiased linear estimators, the OLS estimates have the minimum or least variance.

Hypothesis : It is a tentative statement that we propose to test. It is based on the limited evidence. Hypothesis is formulated on the basis of economic theory or some logic.

Homoscedasticity : A crucial assumption of the Classical Linear Regression Model (CLRM) in that the error term u_i in the population regression function (PRF) is homoscedastic, i.e., they have the same variance σ^2. Such an assumption is referred to as the assumption of homoscedasticity.

Heteroscedasticity : If the variance of u_i is σ_i^2, i.e., it varies from one observation to another, then the situation is referred to as a case of heteroscedasticity.

Interactive Dummy : This is a variable like DX in which there is one dummy variable and one quantitative variable. It is considered in the multiplicative form to enable us to see whether the slope coefficients of two groups are same or different. The functional form of this type of regression is $Y_i = \beta_1 + \beta_2 D_i + \beta_3 X_i + \beta_4(D_i X_i) + u_i$.

Jarque-Bera (J-B) Test : This is an asymptotic or large sample test based on OLS residuals in order to test the normality of the error term. Coefficient of skewness: S, i.e., the asymmetry of PDF. Measure of tallness or height of population distribution function: K

For normal distribution S = 0, K = 3

Jarque and Bera constructed J-Statistics given by

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

Linear Regression : In linear regression models the functional form of the relationship between the variables is linear.

Mathematical Model : A description of system using mathematical concepts

Multicollinearity : The classical linear regression model assumes that there is no perfect multicollinearity, implying no exact linear relationship among the explanatory variables, included in multiple regression models.

MWD test	:	This is the test for the selection of the appropriate functional form for regression as proposed by Mackinnon, White and Davidson. The test is hence known as the MWD Test.
Null Hypothesis	:	The null hypothesis (also called Strawman hypothesis) states that there is no relationship between the variables. The coefficients are deliberately chosen as zero to find out whether Y is related to X at all. If X really belongs in the model, we would fully expect to reject the zero-null hypothesis H_0 in favour of the alternatives hypothesis H_1 that it is not zero.
Near or imperfect multicollinearity	:	The case when two or more explanatory variables are not exactly linear this reinforces the fact that collinearity can be high but not perfect. "High collinearity" refers to the case of "near" or imperfect" or high multicollinearity.
Null Hypothesis	:	It is the hypothesis that there is no significant difference between specified population, the observed difference is mainly due to sampling or experimental error.
Normal Distribution	:	It is a very common probability distribution. The curve is bell-shaped and the area under the normal curve is 1.
Ordinary Least Squares Method	:	Ordinary Least Squares (OLS) is a method for estimation of the unknown parameters in a linear regression model. The OLS method minimizes the sum of the squares of the errors.
Parameters	:	It is a measurement of any variable. A numerical quantity that characterizes a given population
Prediction	:	A regression model explains the variation in the dependent variable on the basis of explanatory variables. Given the values of the explanatory variables, we predict the value of the dependent variable. The predicted value is different from the actual value.
Parameter	:	A quantity or statistical measure for a given population that is fixed. The mean and the variance of a population are population parameters.

p- value	:	It is the lowest level of significance when the null hypothesis can be rejected.
Power of Test	:	The power of any test of statistical significance is defined as the probability that it will reject a false null hypothesis. The value of the power of test is given by $(1 - \beta)$.
Population Regression Function (PRF)	:	A population regression function hypothesizes a theoretical relationship between a dependent variable and a set of independent or explanatory variables. It is a linear function. The function defines how conditional expectation of a variable Y responds to the changes in independent variable X.
Perfect multicollinearity	:	The case of perfect multicollinearity mainly reflects the situation when the explanatory variables and perfectly correlated with each other implying the coefficient of correlation between the explanatory variables is 1.
Park-Test	:	If there is heteroscedasticity in a dataset, the heteroscedastic variance σ_i^2 may be systematically related to one or more of the explanatory variables. In such cases, we can regress σ_i^2 on one or more of such *X*- variables. Such an approach, adopted in the Park-test, helps detect the presence of heteroscedasticity.
Random Variable	:	A variable which takes on values which are numerical outcomes of a random phenomenon.
Regression	:	A regression analysis is concerned with the study of the relationship the explained or dependent variable and the independent or explanatory variables.
Residual Term	:	The actual value of Y is obtained by adding the residual term to the estimated value of Y. The residual term is the estimated value of the random error term of the population regression function.
Ridge Regression	:	The ridge regressions are the method of resolving the problem of multicollinearity. In the ridge regression, the first step is to standardize the variables both dependent and independent by subtracting the respective means and dividing by their standard deviations.

Statistical Inference : It refers to the process of deducing properties of underlying probability distribution of the parameters by analysing data.

Standard Normal Distribution : It refers to a normal distribution with mean 0 and standard deviation 1.

Statistical Inference : It refers to the method of drawing inference about the population parameter on the basis of random sampling.

Statistical Hypothesis: : It is an assumption about a population parameter. This assumption may or may not be true. This statistical hypothesis is either accepted or rejected on the basis of hypothesis testing.

Stochastic Error : The error term represents the influence of those variables that are not included in the regression model. It is evident that even if we try to include all the factors that influence the dependent variable, there exists some intrinsic randomness between the two variables.

Subsidiary or Auxiliary Regressions : When one explanatory variables X is regressed on each of the remaining X variable and the corresponding R^2 is computed. Each of these regressions is referred as subsidiary or auxiliary regression.

***t*- Distribution** : It refers to a continuous probability distribution that is obtained while estimating mean of normally distributed population where sample size is small and population standard deviation is unknown.

Test of significance Approach : The method of inference used to either reject or accept the null hypothesis. This approach makes use of test statistic to make any statistical inference.

Test Statistic : A test statistic is a standardized value that is computed from a sample during the hypothesis testing. On the basis of test statistics one can either reject or accept the null hypothesis.

Type I Error: : In the statistical hypothesis testing, type I error is the incorrect rejection of true null hypothesis. The value is given by alpha level of significance.

Type II Error : The error that occurs when we accept a null hypothesis that is actually false. It is the probability of accepting the null hypothesis when it is false.

Variance Inflation Factor (VIF) : R^2 obtained variables auxiliary regression may not be completely realiable and is not reliable indicator of collinearity. In this method we modify the formula of var (b_2) and (b_3), var $(b_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1-R_2^2)}$

White's General Heteroscedasticity Test : This is a method to test the presence of heteroscedasticity in a regression model. In this, the residuals obtained from original regression are squared and regressed on the original variables, their squared values and their cross-products. Additional powers of original *X* variables can be added.

SOME USEFUL BOOKS

Dougherty, C. (2011). *Introduction to Econometrics*, Fourth Edition, Oxford University Press

Gujarati, D. N. and D.C. Porter (2010). *Essentials of Econometrics*, Fourth Edition, McGraw Hill

Kmenta, J. (2008). *Elements of Econometrics*, Second Edition, Khosla Publishing House

Maddala, G.S., and Kajal Lahiri (2012). *Introduction to Econometrics*, Fourth Edition, Wiley

Wooldridge, J. M. (2014). *Introductory Econometrics: A Modern Approach*, Cengage Learning, Fifth Edition