ECONOMETRIC

(DMSTT23)

(MSC - STATISTICS)

ACHARYA NAGARJUNA UNIVERSITY

CENTRE FOR DISTANCE EDUCATION

NAGARJUNA NAGAR,

GUNTUR

ANDHRA PRADESH

Lesson 1

INTRODUCTION TO ECONOMETRICS

1.0 Objective:

After studying the lesson the student will have clear idea on the need of a separate discipline of Econometrics and on various steps of the methodology of Econometrics, as it is illustrated by ***the well-known Keynesian consumption function***.

Structure of the Lesson:

1.1 What is Econometrics?

1.2 Why a Separate Discipline?

1.3 The Aims and Methodology of Econometrics

1.4 Self Assessment Questions

1.5 References

1.1 What is Econometrics?

Literally speaking, the word "Econometrics" means *measurement in economics.* Although measurement is an important part of econometrics, the scope of econometrics is much broader, as can be seen from the following:

> *The application of statistical and mathematical methods to the analysis of economic data, with a purpose of giving empirical content to economic theories and verifying them or refuting them*.

In this respect econometrics is distinguished from mathematical economics, which consists of the application of mathematics only, and the theories derived need not be necessarily have an empirical testing.

1.2 Why a Separate Discipline?

Econometrics is an amalgam of economic theory, mathematical economics, economic statistics, and mathematical statistics. Yet the subject deserves to be studied in its own right for the following reasons.

. ***Economic theory*** makes statement or hypotheses that are mostly qualitative in nature. For example, microeconomic theory states that, other things remaining the same, a reduction in the price of a commodity is expected to increase the quantity demanded of that commodity. Thus, economic theory postulates a negative or inverse relationship between the price and quantity demanded of a commodity. But the theory itself does not provide any numerical measure of the relationship between the two; that is, it does not tell by how much the quantity will go up or down as a result of a certain change in the price of the commodity. It is the job of the econometrician to provide such numerical estimates. Stated differently, econometrics gives empirical content to most economic theory.

Mathematical economics is mainly concerned with to express economic theory in mathematical form (equations) without regard to measurability or empirical verification of the theory. Econometrics, on the other hand, is mainly interested in the empirical verification of economic theory. As we shall see, the econometrician often uses the mathematical equations proposed by the mathematical economist but puts these equations in such a form that they lend themselves to empirical testing. And this conversion of mathematical equations into econometric equations requires a great deal of ingenuity and practical skill.

Economic statistics is mainly concerned with collecting, processing, and presenting economic data in the form of charts and tables. These are the jobs of the economic statistician. It is he or she who is primarily responsible for collecting data on gross national product (GNP), employment, unemployment, prices, etc. The data thus collected constitute the raw data for econometric work. But the economic statistician does not go any further, not being concerned with using the collected data to test economic theories. Of course, one who does that becomes an econometrician.

Although ***mathematical statistics*** provides many tools used in the trade, the econometrician often needs special methods in view of the unique nature of most economic data, namely, that the data are not generated as the result of a controlled experiment. The econometrician, like the meteorologist, generally depends on data that cannot be controlled directly.

In econometrics the modeler is often faced with **observational** as opposed to **experimental** data. This has two important implications for empirical modeling in econometrics. First, the modeler is required to master very different skills than those needed for analyzing experimental data. Second, the separation of the data collector and the data analyst requires the modeler to familiarize himself/herself thoroughly with the nature and structure of data in question.

1.3 The Aims and Methodology of Econometrics

The aims of econometrics are:

1. Formulation of econometric models that is formulation of economic models in an empirically testable form. Usually, there are several ways of formulating the econometric model from an economic model because we have to choose the functional form, the specification of the stochastic structure of the variables, and so on. This part constitutes the *specification aspect* of the econometric work.
2. Estimation and testing of these models with observed data. This part constitutes the *inference aspect* of the econometric work.
3. Use of these models for prediction and policy purposes.

How do econometricians proceed in their analysis of an economic problem? That is, what is their methodology? Although there are several schools of thought on econometric methodology, we present here the traditional or classical methodology, which still dominates empirical research in economics and other social and behavioral sciences.

Broadly speaking, traditional econometric methodology proceeds along the following lines:

1. Statement of theory or hypothesis.
2. Specification of the mathematical model of the theory
3. Specification of the statistical, or econometric, model
4. Obtaining the data
5. Estimation of the parameters of the econometric model
6. Hypothesis testing
7. Forecasting or prediction
8. Using the model for control or policy purposes.

To illustrate the preceding steps, let us consider the well-known Keynesian theory of consumption.

1. Statement of Theory or Hypothesis

Keynes stated:

The fundamental psychological law . . . is that men [women] are disposed, as a rule and on average, to increase their consumption as their income increases, but not as much as the increase in their income. In short, Keynes postulated that the **marginal propensity to consume (MPC),** the rate of change of consumption for a unit (say, a rupee) change in income, is greater than zero but less than 1.

2. Specification of the Mathematical Model of Consumption

Although Keynes postulated a positive relationship between consumption and income, he did not specify the precise form of the functional relationship between the two. For simplicity, a mathematical economist might suggest the following form of the Keynesian consumption function:

$$Y = \alpha + \beta X \qquad 0 < \beta < 1 \qquad (1.1)$$

where Y = consumption expenditure and X = income, and where α and β, known as the parameters of the model, are, respectively, the intercept and slope coefficients. The slope coefficient β measures the MPC. This equation, which states that consumption is linearly related to income, is an example of a mathematical model of the relationship between consumption and income that is called the **consumption function** in economics. A model is simply a set of mathematical equations. If the model has only one equation, as in the preceding example, it is called a **single-equation model,** whereas if it has more than one equation, it is known as a **multiple-equation model**. But, in this book, we have confined ourselves to only **single-equation models.** In Eq. (1.1) the variable appearing on the left side of the equality sign is called the **dependent variable** and the variable on the right side is called the **independent** or **explanatory** variable. Thus, in the Keynesian consumption function, Eq. (1.1), consumption (expenditure) is the dependent variable and income is the explanatory variable.

3. Specification of the Econometric Model of Consumption

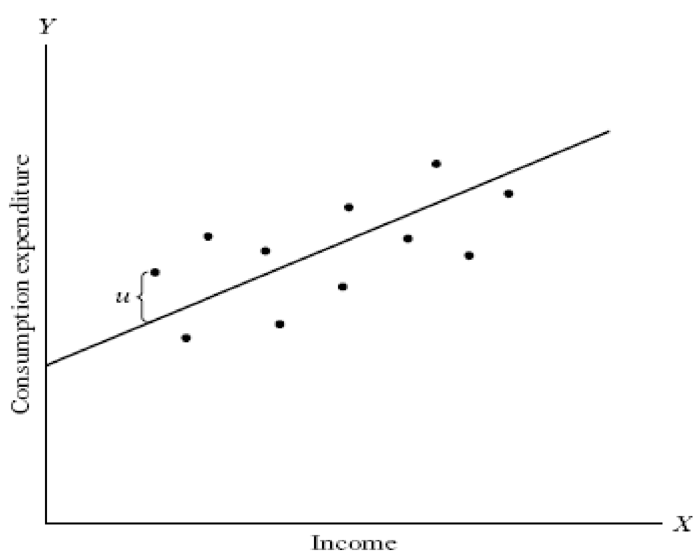
The purely mathematical model of the consumption function given in Eq. (1.1) is of limited interest to the econometrician, for it assumes that there is an *exact* or *deterministic* relationship between consumption and income. But relationships between economic variables are generally inexact. Thus, if we were to obtain data on consumption expenditure and disposable (i.e., after tax) income of a sample of, say, 500 Indian families and plot these data on a graph paper with consumption expenditure on the vertical axis and disposable income on the horizontal axis, we would not expect all 500 observations to lie exactly on the straight line of Eq. (1.1) because, in addition to income, other variables affect consumption expenditure. For example, size of family, ages of the members in the family, family religion, etc., are likely to exert some influence on consumption.

To allow for the inexact relationships between economic variables, the econometrician would modify the deterministic consumption function Eq. (1.1) as follows:

$$Y = \alpha + \beta X + u, \qquad 0 < \beta < 1 \tag{1.2}$$

where u, known as the **disturbance (error) term,** is a **random (stochastic) variable** that has well-defined probabilistic properties. The disturbance term u may well represent all those factors that affect consumption but are not taken into account explicitly.

Eq. (1.2) is an example of an **econometric model.** More technically, it is an example of a **linear regression model,** which is the major concern of this book. The econometric consumption function hypothesizes that the dependent variable Y (consumption) is linearly related to the explanatory variable X (income) but that the relationship between the two is not exact; it is subject to individual variation. The econometric model of the consumption function can be depicted as shown in the following figure.



Econometric model of the Keynesian consumption function.

Figure 1.1

4. Obtaining Data

To estimate the econometric model given in Eq. (1.2), that is, to obtain the numerical values of α and β, we need data. Let us look at the data given in Table I.1, which relate to the U.S. economy for the period 1981–1996. The *Y* variable in this table is the *aggregate* (for the economy as a whole) personal consumption expenditure (PCE) and the *X* variable is gross domestic product (GDP), a measure of aggregate income, both measured in billions of 1992 dollars. Therefore, the data are in "real" terms; that is, they are measured in constant (1992) prices.

TABLE 1.1: DATA ON Y (PERSONAL CONSUMPTION EXPENDITURE) AND X (GROSS DOMESTIC PRODUCT, 1982–1996), BOTH IN 1992 BILLIONS OF DOLLARS

Year	*Y*	*X*	Year	*Y*	*X*
1982	3081.5	4620.3	1990	4132.2	6136.3
1983	3240.6	4803.7	1991	4105.8	6079.4
1984	3407.6	5140.1	1992	4219.8	6244.4
1985	3566.5	5323.5	1993	4343.6	6389.6
1986	3708.7	5487.7	1994	4486.0	6610.7
1987	3822.3	5649.5	1995	4595.3	6742.1
1988	3972.7	5865.2	1996	4714.1	6928.4
1989	4064.6	6062.0			

Source: Economic Report of the President, 1998, Table B–2, p. 282.

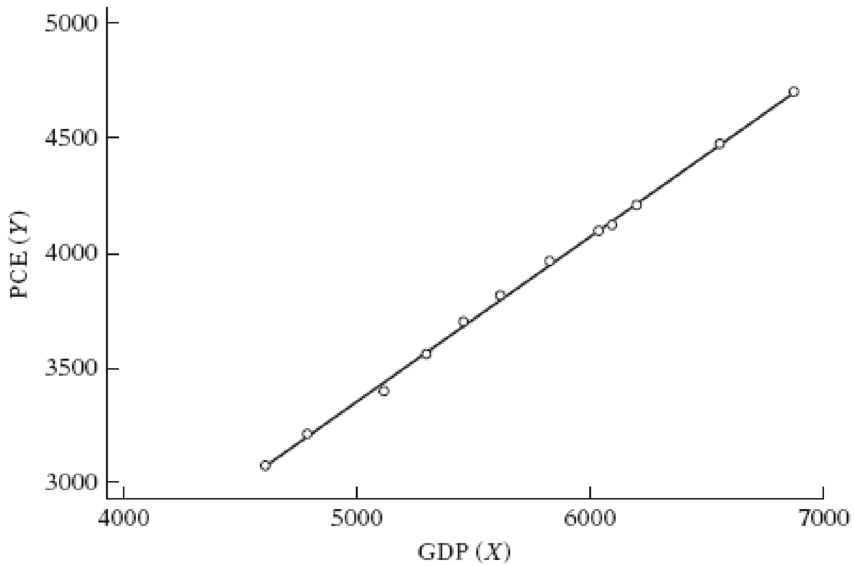
5. Estimation of the Econometric Model

Now that we have the data, our next task is to estimate the parameters of the consumption function. The numerical estimates of the parameters give empirical content to the consumption function. The actual mechanics of estimating the parameters will be discussed in Lesson 2. Now, note that the statistical technique of **regression analysis** is the main tool used to obtain the estimates. Using this technique and the data given in Table1.1, we obtain the following estimates of α and β, namely, −184.08 and 0.7064.

Thus, the estimated consumption function (i.e., regression line)

$$\hat{Y} = -184.08 + 0.7064\, X \tag{1.3}$$

is shown in the following figure.



Personal consumption expenditure (Y) in relation to GDP (X), 1982–1996, both in billions of 1992 dollars.

Figure 1.2

As the above figure shows, the regression line fits the data quite well in that the data points are very close to the regression line. From this figure we see that for the period 1982–1996 the slope coefficient (i.e., the **MPC**) was about 0.70, suggesting that for the sample period an increase in real income of 1 dollar led, *on average*, to an increase of about 0.7 dollar in real consumption expenditure. We say *on average* because the relationship between consumption and income is inexact; as is clear from the above figure; not all the data points lie exactly on the regression line. In simple terms we can say that, according to our data, the *average*, or *mean*, consumption expenditure went up by about 0.7 dollar for a dollar's increase in real income.

6. Hypothesis Testing

Assuming that the fitted model is a reasonably good approximation of reality, we have to develop suitable criteria to find out whether the estimates obtained in, say, Eq. (1.3) are in accord with the expectations of the theory that is being tested. As noted earlier, Keynes expected the MPC to be positive but less than 1. In our example we found the MPC to be about 0.70. But before we accept this finding as confirmation of Keynesian consumption theory, we must enquire whether this estimate is sufficiently below unity to convince us that this is not a chance occurrence or peculiarity of the particular data we have used. In other words, is 0.70 *statistically less than 1?* If it is, it may support Keynes' theory.

Such confirmation or refutation of economic theories on the basis of sample evidence is based on a branch of statistical theory known as **statistical inference (hypothesis testing).** Throughout this book we shall see how this inference process is actually conducted.

7. Forecasting or Prediction

If the chosen model does not refute the hypothesis or theory under consideration, we may use it to predict the future value(s) of the dependent, or forecast variable *Y* on the basis of known or expected future value(s) of the explanatory, or predictor variable *X*.

To illustrate, suppose we want to predict the mean consumption expenditure for 1997. The GDP value for 1997 was 7269.8 billion dollars. Putting this GDP figure on the right-hand side of Eq. (1.3), we obtain

$$\hat{Y}_{1997} = -184.0779 + 0.7064\ (7269.8) = 4951.3167 \qquad (1.4)$$

or about 4951 billion dollars. Thus, given the value of the GDP, the mean, or average, forecast consumption expenditure is about 4951 billion dollars. The actual value of the consumption expenditure reported in 1997 was 4913.5 billion dollars. The estimated model Eq. (1.3) thus *over-predicted* the actual consumption expenditure by about 37.82 billion dollars. We could say the *forecast-error* is about 37.82 billion dollars, which is about 0.76 percent of the actual GDP value for 1997. When we fully discuss the linear regression model in subsequent chapters, we will try to find out if such an error is “small” or “large.” But what is important for now is to note that such forecast errors are inevitable given the statistical nature of our analysis.

There is another use of the estimated model Eq. (1.3). Suppose the President decides to propose a reduction in the income tax. What will be the effect of such a policy on income and thereby on consumption expenditure and ultimately on employment?

Suppose that, as a result of the proposed policy change, investment expenditure increases. What will be the effect on the economy? As macroeconomic theory shows, the change in income following, say, a dollar’s worth of change in investment expenditure is given by the *income multiplier M*, which is defined as

$$M = 1/(1 - MPC) \qquad (1.5)$$

If we use the MPC of 0.70 obtained in Eq. (1.3), this multiplier becomes about $M = 3.33$. That is, an increase (decrease) of a dollar in investment will *eventually* lead to more than a threefold increase (decrease) in income; note that it takes time for the multiplier to work. The critical value in this computation is MPC, for the multiplier depends on it. And this estimate of the MPC can be obtained from regression models such as Eq. (1.3). Thus, a quantitative estimate of MPC provides valuable information for policy purposes. Knowing MPC, one can predict the future course of income, consumption expenditure, and employment following a change in the government’s fiscal policies.

8. Use of the Model for Control or Policy Purposes

Suppose we have the estimated consumption function given in Eq. (1.3). Suppose further the government believes that consumer expenditure of about 4900 (billions of 1992 dollars) will keep the unemployment rate at its current level of about 4.2 percent (early 2000). What level of income will guarantee the target amount of consumption expenditure? If the regression results given in Eq. (1.3) seem reasonable, simple arithmetic will show that

$$4900 = -184.0779 + 0.7064X \qquad (1.6)$$

which gives X = 7197, approximately. That is, an income level of about 7197 (billion) dollars, given an MPC of about 0.70, will produce an expenditure of about 4900 billion dollars. As these calculations suggest, an estimated model may be used for control, or policy, purposes. By appropriate fiscal and monetary policy mix, the government can manipulate the *control variable X* to produce the desired level of the *target variable Y.*

1.4 Self Assessment Questions

1. What is Econometrics? Explain why a separate discipline of econometrics is need?
2. Explain the Role of Econometrics and what are the aims of Econometrics?
3. Discuss the scope, nature and limitations of Econometrics.

1.5 References

1. Gujarati, D.N. (2005): *Basic Econometrics, 4th Ed., Tata McGraw-Hill.*
2. Johnston, J. (1984): *Econometric Methods, 3rd Ed., McGraw-Hill,* New York.
3. Montgomery, D.C., Peck, E.A. and Geoffrey Vining, G. (2003): *Introduction to Linear Regression Analysis, 3rd Ed., Wiley*
4. Draper, N.R., and H. Smith(1998): *Applied Regression Analysis, 3rd Ed., John Wiley & Sons,* New York.
5. G.S. Maddala (2001): *Introduction to Econometrics, 3rd Ed., John Wiley & Sons, Ltd.*
6. Johnston, J. and DiNardo J (1997): *Econometric Methods*, 4th Ed., McGraw Hill.
7. Hill, Carter, William Griffiths, and George Judge(2001): *Undergraduate Econometrics, John Willey & Sons,* New York.
8. *Koutsoyiannis, A(1973): Theory of Econometrics, Harper & Row,* New York.

Lesson 2

SIMPLE REGRESSION ANALYSIS: ESTIMATION

2.0 Objective:

After studying the lesson the student will have clear idea regarding the objective of regression analysis, correlation vs regression, the estimation of the simple linear regression model and properties of the estimators.

Structure of the Lesson:

2.1 Introduction

Regression analysis is one of the most commonly used tools in econometric work. Regression analysis is concerned with describing and evaluating the relationship of a given variable (often called the explained or dependent variable) with one or more other variables (often called the explanatory or independent variables) with a view to estimate and/or predict the (population) mean or average value of the dependent variable in terms of the known or fixed (in repeated sampling) values of the independent variables.

If we are studying the dependence of a variable on only a single explanatory variable, such as that of consumption expenditure on real income, such a study is known as ***simple*** *(two-variable)* ***regression analysis***. However, if we are studying the dependence of one variable on more than one explanatory variable, such as the study of the yield of a particular crop on rainfall, temperature, sunshine, and fertilizer, it is known as **multiple regression analysis.** In other words, in *simple regression* there is only one explanatory variable, whereas in *multiple regression* there are several explanatory variables.

Some illustrations:

1. The well-known *Keynesian consumption function,* which is already explained in Lesson1.
2. A monopolist, who can fix
   X=price or output (but not both), may want to find out the response of
   Y=demand for a product, to changes in price.
   Such an experiment may enable the estimation of the **price elasticity** (i.e., price responsiveness) of the demand for the product and may help determine the most profitable price.
3. A labor economist may want to study the
   Y=rate of change of money wages
   in relation to X=the unemployment rate.
   Such knowledge may be helpful in stating something about the inflationary process in an economy, for increases in money wages are likely to be reflected in increased prices.
4. The marketing director of a company may want to know how
   Y=demand for the company's product
   is related to X=advertising expenditure.
   Such a study will be of considerable help in finding out the **elasticity of demand** with respect to advertising expenditure, that is, the percent change in demand in response to, say, a 1 percent change in the advertising budget. This knowledge may be helpful in determining the "optimum" advertising budget.
5. Finally, an agronomist may be interested in studying the dependence of
   Y= crop yield, say, of wheat,
   on X= rainfall.
   Such a dependence analysis may enable the prediction or forecasting of the average crop yield, given information about the rainfall.

2.2 Regression versus Causation

Although regression analysis deals with the dependence of one variable on other variables, it does not necessarily imply causation. In the words of Kendall and Stuart, "A statistical relationship, however strong and however suggestive, can never establish causal connection and our ideas of causation must come from outside statistics, ultimately from some theory or other."

In the crop-yield example cited previously, there is no *statistical reason* to assume that crop yield does not depend on rainfall. The fact that we treat crop yield as dependent on rainfall (among other things) is due to non-statistical considerations. Common sense suggests that the relationship cannot be reversed, for we cannot control rainfall by varying crop yield.

In all the examples cited in Section 2.1, the point to note is that *a statistical relationship in itself cannot logically imply causation*. To ascribe causality, one must appeal to a priori or theoretical considerations. Thus, in the well-known *Keynesian consumption function,* discussed in **Lesson1**, one can invoke economic theory in saying that consumption expenditure depends on real income.

2.3 Regression versus Correlation

Closely related to but conceptually very much different from regression analysis is *correlation analysis*, where the primary objective is to measure the *strength* or *degree* of *linear* *association* between two variables. The *correlation coefficient* measures this strength of (linear) association. For example, we may be interested in finding the correlation (coefficient) between smoking and lung cancer, between scores on statistics and mathematics examinations, between high school grades and college grades, and so on. In regression analysis, as already noted, we are not primarily interested in such a measure. Instead, we try to estimate or predict the average value of one variable on the basis of the fixed values of other variable. Thus, we may want to know whether we can predict the average score on a statistics examination by knowing a student's score on a mathematics examination.

Regression and correlation have some fundamental differences that are worth mentioning. In regression analysis there is an asymmetry in the way the dependent (explained) and independent (explanatory) variables are treated. The dependent variable is assumed to be statistical, random, or stochastic, that is, to have a probability distribution. The independent variable, on the other hand, is assumed to have fixed values (in repeated sampling), which was made explicit in the definition of regression. Thus, we assumed that the variable age was fixed at given levels and height measurements were obtained at these levels. In correlation analysis, on the other hand, we treat both the variables symmetrically; there is no distinction between the dependent and independent variables. After all, the correlation between scores on mathematics and statistics examinations is the same as that between scores on statistics and mathematics examinations. Moreover, both variables are assumed to be random. Most of the correlation theory is based on the assumption of randomness of variables, whereas in regression theory we have to assume that the dependent variable is stochastic but the independent variable, need not be stochastic always and in most of the occasions, may be non-stochastic or fixed variable.

2.4 Terminology and Notation

Before we proceed to a formal analysis of regression theory, let us spell out briefly on the matter of terminology and notation. In the literature, the terms *dependent variable* and *independent variable* are described variously. A representative list is:

a) Dependent variable	Independent variable(s)
↕	↕
b) Predictand	Predictor(s)
↕	↕
c) Regressand	Regressor(s)
↕	↕
d) Explained variable	Explanatory variable(s)
↕	↕
e) Effect variable	Causal variable(s)
↕	↕
f) Endogenous variable	Exogenous variable(s)
↕	↕
g) Target variable	Control variable(s)

Unless stated otherwise, the letter Y will denote the dependent/explained variable and the X's (X_1, X_2, . . ., X_k) will denote the independent/explanatory variables, X_k being the k^{th} explanatory variable. The subscript i or t will denote the ith or the t^{th} observation or value. X_{ki} (or X_{kt}) will denote the i^{th} (or t^{th}) observation on variable X_k. As a matter of convention, the observation subscript i will be used for **cross sectional data** (i.e., data collected at one point in time) and the subscript t will be used for **time series data** (i.e., data collected over a period of time).

The term *random* is a synonym for the term *stochastic***.** A random or stochastic variable is a variable that can take on any set of values, positive or negative, with a given probability.

2.5 Simple Linear Regression Equation

We may postulate the relationship between a dependent variable Y and an independent variable X as

$$Y = f(X) \tag{2.1}$$

which indicates that the variable X is influencing the other variable Y. Here the function $f(X)$ may be either a linear function or a non-linear function. Let us confine ourselves, in this lesson, to only a linear function. Hence, we may write the Eq. (2.1) as

$$Y = \alpha + \beta X \tag{2.2}$$

where α and β are the unknown parameters and are very often called as intercept and slope coefficients. Here, it may be noted that the above Eq. (2.2) is a deterministic or mathematical liner relationship, which is not suitable for measuring and testing the relationship between economic variables. Therefore, we have to convert the deterministic Eq. (2.2) into a stochastic equation by introducing a stochastic term into the equation. Thus, the linear relationship Eq. (2.2) may, now, be expressed as a stochastic linear relationship given by

$$Y = \alpha + \beta X + u \tag{2.3}$$

where u is a stochastic or random variable (often called as error term or disturbance term) with known p.d.f. In the above equation $\alpha + \beta X$ is the deterministic component of Y and u is the stochastic or random component. *Eq. (2.3) is known as a* ***simple linear regression equation***. The unknown parameters α and β are called as regression coefficients or regression parameters, which are to be estimated from the data on Y and X .

2.6 The Significance of the Stochastic Disturbance term

The following are the various reasons for inclusion of the stochastic disturbance term u in the simple linear model.

1. *Unpredictable element of randomness in human behavior/responses:* For instance, if Y=consumption expenditure of a household and X=disposable income of the household, there is an unpredictable element of randomness in each household's consumption. The household does not behave like a machine. In one month the people in the household are on a spending spree. In another month they are tightfisted.
2. *Effect of a large number of omitted variables.* Again, in our example, X is not the only variable influencing Y. The family size, tastes of the family, spending habits, and so on, affect the variable Y. The error u is a *catchall* for the effects of all these variables, some of which may not even be quantifiable, and some of which may not even be identifiable. But it is quite possible that the joint influence of all or some of these variables may be so small and at best nonsystematic or random that as a practical matter and for cost considerations it does not pay to introduce them into the model explicitly.
3. *Unavailability of data:* Even if we know what some of the excluded variables are and therefore consider a multiple regression rather than a simple regression, we may not have quantitative information about these variables. It is a common experience in empirical analysis that the data, which we would like to have ideally, often are not available. For example, in principle we could introduce family wealth as an explanatory

variable in addition to the income variable to explain family consumption expenditure. But unfortunately, information on family wealth generally is not available. Therefore, we may be forced to omit the wealth variable from our model despite its great theoretical relevance in explaining consumption expenditure.

4. *Measurement error* in Y. In our example this refers to measurement error in the household consumption. That is, we cannot measure it accurately. The disturbance term u may represent these errors of measurement.

5. *Wrong functional form:* Even if we have theoretically correct variables explaining a phenomenon and even if we can obtain data on these variables, very often we do not know the form of the functional relationship between the regressand and the regressors. Is consumption expenditure a linear (in variables) function of income or a nonlinear (in variables) function? In two-variable models the functional form of the relationship can often be judged from the scatter diagram. But in a multiple regression model, it is not easy to determine the appropriate functional form, for graphically we cannot visualize

For all these reasons, the stochastic disturbance u assume an extremely critical role in regression analysis, which we will see as we progress.

2.7 The Simple Linear Regression Model

If we have *n* observations on Y and X, we can write Eq. (2.3) as

$$Y_i = \alpha + \beta X_i + u_i \qquad \forall i \tag{2.4}$$

Now, our objective is to get estimates of the unknown parameters α and β of the above equation based on the given n sets of observations on Y and X. In order to do this we have to make some assumptions about the error terms u_i s, which are given below.

1. *Zero mean*. $E(u_i) = 0 \;\; \forall i$. Equivalently $E(Y_i) = \alpha + \beta X_i \quad \forall\, i$
2. *Homoscedasticity or Common Variance.* $\text{var}(u_i) = E(u_i^2) = \sigma^2 \quad \forall\, i$.
3. *No autocorrelation between the disturbances.* In other words, u_i and u_j (*i ≠ j*) are uncorrelated. i.e. $\text{cov}(u_i, u_j) = E(u_i u_j) - E(u_i)E(u_j) = E(u_i u_j) = 0 \quad \forall\, i \neq j$.
4. *X values are fixed in repeated sampling.* Values taken by the regressor *X* are considered fixed in repeated samples. More technically, *X* is assumed to be *non-stochastic.*
5. *Zero covariance between u_i and X_i*, or $\text{cov}(X_i, u_i)$ = 0. It will be automatically fulfilled if *X* variable is non-random or non-stochastic
6. *The number of observations n must be greater than the number of parameters to be estimated.* Alternatively, the number of observations *n* must be greater than the number of explanatory variables.
7. *The regression model is correctly specified.* Alternatively, there is no **specification bias or error** in the model used in empirical analysis.

The set of n equations given in Eq. (2.4) along with the above assumptions is called ***simple linear regression model***.

2.8 Principle of Least Squares Estimation

Now, let us consider the ***simple linear regression model***, which is explained in the above section, given by

$$
\begin{aligned}
&Y_i = \alpha + \beta X_i + u_i \qquad i = 1, 2, ..., n \\
&E(u_i) = 0 \quad \text{for all } i \\
&cov(u_i, u_j) = E(u_i u_j) = \begin{cases} 0 & \text{for } i \neq j; i, j = 1, 2, \ldots, n \\ \sigma^2 & \text{for } i = j; i, j = 1, 2, \ldots, n \end{cases}
\end{aligned} \tag{2.5}
$$

where α, β, and σ^2 are unknown parameters.

Let us suppose that $\hat{\alpha}$ and $\hat{\beta}$ are some arbitrary estimates of the unknown parameters α and β of the above model. Then the unknown regression model Eq. (2.5) may be replaced with an estimated regression model as

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + e_i \qquad i = 1, 2, ..., n \tag{2.6}$$

where e_i is the difference between observed Y_i and estimated $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ often called as *'residual'.*

Now, the ***method of principle of least squares*** is that the $\hat{\alpha}$ and $\hat{\beta}$ should be chosen so as the residual sum of squares

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2 \tag{2.7}$$

is least. In order to minimize $\sum_{i=1}^{n} e_i^2$, the partial derivatives of it with respect to $\hat{\alpha}$ and $\hat{\beta}$ are set to be equal to zero. So that

$$
\begin{aligned}
&\frac{\partial}{\partial \hat{\alpha}} \left(\sum_{i=1}^{n} e_i^2 \right) = -2 \sum_{i=1}^{n} \left(Y_i - \hat{\alpha} - \hat{\beta} X_i \right) = 0 \\
&\frac{\partial}{\partial \hat{\beta}} \left(\sum_{i=1}^{n} e_i^2 \right) = -2 \sum_{i=1}^{n} X_i \left(Y_i - \hat{\alpha} - \hat{\beta} X_i \right) = 0
\end{aligned} \tag{2.8}
$$

Simplifying these equations, we get

$$
\begin{aligned}
&\sum_{i=1}^{n} Y_i = n\hat{\alpha} + \hat{\beta} \sum_{i=1}^{n} X_i \\
&\sum_{i=1}^{n} X_i Y_i = \hat{\alpha} \sum_{i=1}^{n} X_i + \hat{\beta} \sum_{i=1}^{n} X^2{}_i
\end{aligned} \tag{2.9}
$$

The pair of equations given above are popularly known as ***the normal equations*** of the straight line $Y = \hat{\alpha} + \hat{\beta} X$.

Dividing the first equation of Eq. (2.9) by n, we get

$$\overline{Y} = \hat{\alpha} + \hat{\beta} \overline{X}, \qquad \text{where } \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \ , \ \ \overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{2.10}$$

If we replace $\hat{\alpha}$, using Eq. (2.10), in second equation of Eq. (2.9), we get

$$\sum_{i=1}^{n} X_i Y_i = (\bar{Y} - \hat{\beta}\bar{X})\sum_{i=1}^{n} X_i + \hat{\beta}\sum_{i=1}^{n} X_i^2$$

$$= n\bar{X}\ \bar{Y} + \hat{\beta}\left(\sum_{i=1}^{n} X_i^2 \text{-n}\, \bar{X}^2\right) \tag{2.11}$$

Rearranging Eq. (2.11), we get

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i \text{ - } \bar{X}\ \bar{Y}}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 \text{ - } \bar{X}^2} = \frac{\operatorname{cov}(X, Y)}{\operatorname{var}(X)} \tag{2.12}$$

From Eq. (2.10) we get

$$\hat{\alpha} = \bar{Y} \text{ - } \hat{\beta}\bar{X} \tag{2.13}$$

Eqs. (2.12) and (2.13) are known as the least squares estimators of the parameters α and β respectively.

Remark 1: From first Eq. (2.8), we have

$$\sum_{i=1}^{n}\left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right) = 0$$

$$\boxed{\Rightarrow \sum_{i=1}^{n} e_i = 0 \quad \textit{i.e. The sum of the residuals is zero}} \tag{2.14}$$

Remark 2: From Eq. (2.10), we may note that the estimated regression line passes through the point of means $(\bar{X}, \bar{Y})$.

2.9 Properties of Least Squares Estimators

From Eq. (2.12), we may write that

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i \text{-} \bar{X}\ \bar{Y}}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 \text{-} \bar{X}^2}$$

$$= \frac{\sum_{i=1}^{n} X_i Y_i \text{- n}\bar{X}\ \bar{Y}}{\sum_{i=1}^{n} X_i^2 \text{-n}\bar{X}^2}$$

$$\hat{\beta}=\frac{\sum_{i=1}^{n}x_i y_i}{\sum_{i=1}^{n}x_i^2},\quad \text{where } \mathrm{x}_i = X_i\text{-}\bar{X} \text{ and } \mathrm{y}_i = Y_i\text{-}\bar{Y}$$

$$=\frac{\sum_{i=1}^{n}x_i Y_i}{\sum_{i=1}^{n}x_i^2} - \frac{\bar{Y}\sum_{i=1}^{n}x_i}{\sum_{i=1}^{n}x_i^2}$$

$$=\frac{\sum_{i=1}^{n}x_i Y_i}{\sum_{i=1}^{n}x_i^2}\ ,\quad \left(\because \sum_{i=1}^{n}x_i=\sum_{i=1}^{n}\left(X_i\text{-}\bar{X}\right)=\sum_{i=1}^{n}X_i\text{-}n\bar{X}=n\bar{X}\text{-}n\bar{X}=0\right)$$

$$=\sum_{i=1}^{n}w_i Y_i \qquad \text{where } w_i=\frac{x_i}{\sum_{i=1}^{n}x_i^{2}} \tag{2.15}$$

and we may note that

$$\sum_{\mathrm{i=1}}^{\mathrm{n}}w_i=0,\quad \sum_{i=1}^{n}w_i^2=\frac{1}{\sum_{i=1}^{n}x_i^2},\quad \sum_{\mathrm{i=1}}^{\mathrm{n}}w_i x_i=\sum_{\mathrm{i=1}}^{\mathrm{n}}w_i X_i=1 \tag{2.16}$$

From Eq. (2.15) we may notice that $\hat{\beta}$ is a linear function of the actual observations Y_i, i=1,2,…,n.

Substituting Eq. (2.15) in Eq. (2.13) and rearranging it we get

$$\hat{\alpha}=\sum_{i=1}^{n}\left(\frac{1}{n}-\bar{X}w_i\right)Y_i \tag{2.17}$$

Thus from Eqs. (2.15) and (2.17), the least squares estimators $\hat{\alpha}$ and $\hat{\beta}$ **are linear estimators.**

Using Eq. (2.4) in Eq. (2.15) we ge

Substituting Eq. (2.4) in Eq. (2.15), we get

$$\hat{\beta}=\sum w_i Y_i=\sum w_i(\alpha+\beta X_i+u_i)=\beta+\sum_{i=1}^{n}w_i u_i \qquad \text{(using Eq.(2.16))} \tag{2.18}$$

Taking expectation on both sides, we get

$$\mathrm{E}(\hat{\beta})=\beta+\sum_{i=1}^{n}w_i\mathrm{E}(u_i)=\beta \qquad (\because \mathrm{E}(u_i)=0\ ) \tag{2.19}$$

Thus $\hat{\beta}$ is a linear unbiased estimator of β

Similarly, subtitling Eq. (2.4) in Eq. (2.17) we get

$$\hat{\alpha} = \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{X}w_i\right)(\alpha + \beta X_i + u_i)$$

$$= \alpha - \alpha\bar{X}\sum_{i=1}^{n} w_i + \beta\bar{X} - \beta\bar{X}\sum_{i=1}^{n} w_i X_i + \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{X}w_i\right)u_i$$

$$= \alpha + \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{X}w_i\right)u_i \qquad (\text{using Eq.}(2.16)) \qquad (2.20)$$

Taking expectation on both sides, we get

$$\mathrm{E}(\hat{\alpha}) = \alpha + \sum\left(\frac{1}{n} - \bar{X}w_i\right)\mathrm{E}(u_i) = \alpha \qquad (\because \mathrm{E}(u_i) = 0\ \forall\ \mathrm{i}) \qquad (2.21)$$

Thus $\hat{\alpha}$ is linear unbiased estimator of α

The variance of $\hat{\beta}$ is given by

$$\mathrm{var}(\hat{\beta}) = E\left[\left(\hat{\beta} - \beta\right)^2\right]$$

$$= E\left[\left(\sum_{i=1}^{n} w_i u_i\right)^2\right] \qquad (\text{using Eq.}(2.18))$$

$$= E\left[\sum_{i=1}^{n} w_i^2 u_i^2 + \sum_{i \neq j=1}^{n} w_i w_j u_i u_j\right]$$

$$= \sum_{i=1}^{n} w_i^2 E\left(u_i^2\right) \qquad \left(\because E\left(u_i u_j\right) = 0\right)$$

$$= \sigma^2 \sum_{i=1}^{n} w_i^2 \qquad \left(\because E\left(u_i^2\right) = \sigma^2\right)$$

Using Eq. (2.16), we get

$$\boxed{\mathrm{var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}} \qquad (2.22)$$

We derive in a similar fashion the variance of $\hat{\alpha}$ as

$$\begin{aligned}
\operatorname{var}(\hat{\alpha}) &= E\left[\left(\hat{\alpha}-\alpha\right)^2\right] \\
&= \left[\sum_{i=1}^{n}\left(\frac{1}{n}-\bar{X}w_i\right)u_i\right]^2 && \left(\text{from Eq. (2.20)}\right) \\
&= \sum_{i=1}^{n}\left(\frac{1}{n}-\bar{X}w_i\right)^2 E\left(u_i^2\right) && \left(\because \mathrm{E}\left(u_i u_j\right)=0\ \forall \mathrm{i}\neq \mathrm{j}\right) \\
&= \left(\frac{1}{n}+\bar{X}^2\sum_{i=1}^{n} w_i^2-\frac{2\bar{X}}{n}\sum_{i=1}^{n} w_i\right)\sigma^2 && \left(\because \mathrm{E}\left(u_i^2\right)=\sigma^2\right) \\
&= \left(\frac{1}{n}+\frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2}\right)\sigma^2 && \left(\text{using Eq. (2.16)}\right)
\end{aligned}$$

and rearranging slightly, we get

$$\operatorname{var}(\hat{\alpha}) = \frac{\sum_{i=1}^{n} x_i^2 + n\bar{X}^2}{n\sum_{i=1}^{n} x_i^2}\ \sigma^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 + n\bar{X}^2}{n\sum_{i=1}^{n} x_i^2}\ \sigma^2 = \frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}\ \sigma^2 \qquad (2.23)$$

From Eq. (2.20) we have

$$\begin{aligned}
\hat{\alpha}-\alpha &= \sum_{i=1}^{n}\left(\frac{1}{n}-\bar{X}w_i\right)u_i \\
&= \bar{u}-\bar{X}\sum_{i=1}^{n} w_i u_i \\
&= \bar{u}-\bar{X}\left(\hat{\beta}-\beta\right) && \left(\text{using Eq. (2.18)}\right)
\end{aligned}$$

Now the covariance between $\hat{\alpha}$ and $\hat{\beta}$ is

$$\begin{aligned}
\operatorname{cov}\left(\hat{\alpha},\hat{\beta}\right) &= \mathrm{E}\left\{\left(\hat{\alpha}-\alpha\right)\left(\hat{\beta}-\beta\right)\right\} = \mathrm{E}\left\{\left[\bar{u}-\bar{\mathrm{X}}\left(\hat{\beta}-\beta\right)\right]\left(\hat{\beta}-\beta\right)\right\} \\
&= -\bar{\mathrm{X}}\mathrm{E}\left\{\left(\hat{\beta}-\beta\right)^2\right\} = -\bar{\mathrm{X}}\operatorname{var}\left(\hat{\beta}\right) && \left(\because \mathrm{E}\left\{\left(\hat{\beta}-\beta\right)\bar{u}\right\}=0\right) \\
&= -\frac{\bar{\mathrm{X}}\sigma_u^2}{\sum_{i=1}^{n} x_i^2} && \left(\text{from Eq. (2.22)}\right) \qquad (2.24)
\end{aligned}$$

The least-squares estimator $\hat{\beta}$ is BLUE (Gauss-Markov theorem for simple linear regression model):

In the above, we have already shown that $\hat{\beta}$ is a linear unbiased estimator of β. Now, in order to show that $\hat{\beta}$ is BLUE, we have to yet show that $\hat{\beta}$ has minimum variance among the class of all unbiased estimators.

Now, let us consider an arbitrary linear estimator of β given by

$$b = \sum_{i=1}^{n} c_i Y_i \tag{2.25}$$

where the problem is to choose the weights c_i s such that

$$\mathrm{E}(b) = \beta$$

and to make the $\mathrm{var}(b)$ as small as possible. Using Eq. (2.4) in Eq. (2.25), we get

$$\begin{aligned} b &= \sum_{i=1}^{n} c_i(\alpha + \beta X_i + u_i) \\ &= \alpha \sum_{i=1}^{n} c_i + \beta \sum_{i=1}^{n} c_i X_i + \sum_{i=1}^{n} c_i u_i \end{aligned}$$

Taking expectation on both sides, we get

$$E(b) = \alpha \sum_{i=1}^{n} c_i + \beta \sum_{i=1}^{n} c_i X_i \qquad (\because E(u_i) = 0)$$

and $$E(b) = \beta \quad \Leftrightarrow \sum_{i=1}^{n} c_i = 0 \text{ and } \sum_{i=1}^{n} c_i X_i = 1 \tag{2.26}$$

Under conditions given in Eq. (2.26), the above equation becomes

$$b = \beta + \sum_{i=1}^{n} c_i u_i$$

and $$\begin{aligned} \mathrm{var}(b) &= E\left[(b - \beta)^2 \right] \\ &= E\left[\left(\sum_{i=1}^{n} c_i u_i \right)^2 \right] \\ &= \sum_{i=1}^{n} c_i^2 E(u_i^2) \qquad \left(\because E(u_i u_j) = 0 \ \ \forall i \neq j \right) \\ &= \sigma^2 \sum_{i=1}^{n} c_i^2 \end{aligned} \tag{2.27}$$

The problem now is to minimize $\mathrm{var}(b)$ subject to the conditions given in Eq. (2.26).
But, from Eq. (2.22) the variance of OLS estimator $\hat{\beta}$ is

$$\mathrm{var}(\hat{\beta}) = \sigma^2 \sum_{i=1}^{n} w_i^2 \quad \text{where } w_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2} \tag{2.28}$$

for comparing $\mathrm{var}(b)$ with $\mathrm{var}(\hat{\beta})$, we write

$$c_i = w_i + (c_i - w_i)$$

$$\sum_{i=1}^{n} c_i^2 = \sum_{i=1}^{n} w_i^2 + \sum_{i=1}^{n} (c_i - w_i)^2 + 2\sum_{i=1}^{n} w_i (c_i - w_i) \tag{2.29}$$

Consider

$$\sum_{i=1}^{n} w_i\left(c_i - w_i\right) = \sum_{i=1}^{n} w_i c_i - \sum_{i=1}^{n} w_i^2$$

$$= \frac{\sum_{i=1}^{n} x_i c_i}{\sum_{i=1}^{n} x_i^2} - \frac{1}{\sum_{i=1}^{n} x_i^2} \qquad \left(\because \text{ from Eq. (2.28)}\right)$$

$$= \frac{\sum_{i=1}^{n} c_i X_i}{\sum_{i=1}^{n} x_i^2} - \frac{\bar{X}\sum_{i=1}^{n} c_i}{\sum_{i=1}^{n} x_i^2} - \frac{1}{\sum_{i=1}^{n} x_i^2} \qquad \left(\because x_i = X_i - \bar{X}\right)$$

$$= \frac{1}{\sum_{i=1}^{n} x_i^2} - \frac{1}{\sum_{i=1}^{n} x_i^2} \qquad \left(\text{using Eq. (2.26)}\right)$$

$$= 0 \tag{2.30}$$

Substituting Eq. (2.30) in Eq. (2.29), we get

$$\sum_{i=1}^{n} c_i^2 = \sum_{i=1}^{n} w_i^2 + \sum_{i=1}^{n}\left(c_i - w_i\right)^2 \tag{2.31}$$

Substituting Eq. (2.31) in Eq. (2.27) we get

$$\text{var}(b) = \sigma^2\left[\sum_{i=1}^{n} w_i^2 + \sum_{i=1}^{n}\left(c_i - w_i\right)^2\right]$$

$$= \text{var}(\hat{\beta}) + \sigma^2\sum_{i=1}^{n}\left(c_i - w_i\right)^2$$

Since $\sum_{i=1}^{n}\left(c_i - w_i\right)^2 \geq 0$, we have

$$\text{var}(b) \geq \text{var}(\hat{\beta}) \tag{2.32}$$

Equality hold only when $c_i = w_i$ for all *i*, in which case obviously $b = \hat{\beta}$

Thus $\hat{\beta}$ is a linear unbiased estimator of β with minimum variance among the class of all linear unbiased estimators. Therefore, by definition, $\hat{\beta}$ is BLUE of β.

Hence Gauss-Markov theorem is proved in case of simple linear regression model.

The least-squares estimator of σ^2:

We have the simple linear model

$$Y_i = \alpha + \beta X_i + u_i \qquad i = 1, 2, ..., n \tag{2.33}$$

with the assumptions

1. $E(u_i) = 0 \qquad \forall i$.
2. $\text{var}(u_i) = E(u_i^2) = \sigma^2 \quad \forall\ i$.
3. $\text{cov}(u_i, u_j) = E(u_i u_j) - E(u_i)E(u_j) = E(u_i u_j) = 0 \quad \forall\ i \neq j$.

If we average the above equation over the n sample values we obtain

$$\bar{Y} = \alpha + \beta\bar{X} + \bar{u} \tag{2.34}$$

Subtracting Eq. (2.34) from Eq. (2.33) we get

$$y_i = \beta x_i + \left(u_i - \bar{u}\right)$$

suppose $\hat{\beta}$ is least squares estimator of β then we have $\hat{y}_i = \hat{\beta}x_i$. Hence the residual

$$e_i = \mathrm{y}_i - \hat{\mathrm{y}}_i = \beta x_i + \left(u_i - \bar{u}\right) - \hat{\beta}x_i = -\left(\hat{\beta} - \beta\right)x_i + \left(u_i - \bar{u}\right)$$

Therefore

$$\sum_{i=1}^{n} e_i^2 = \left(\hat{\beta} - \beta\right)^2 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n}\left(u_i - \bar{u}\right)^2 - 2\left(\hat{\beta} - \beta\right)\sum_{i=1}^{n} x_i\left(u_i - \bar{u}\right)$$

Taking expected values of each term on the right-hand side gives

$$E\left[\left(\hat{\beta} - \beta\right)^2 \sum_{i=1}^{n} x_i^2\right] = Var\left(\hat{\beta}\right)\sum_{i=1}^{n} x_i^2 = \sigma^2 \qquad \text{using(2.22)}$$

$$E\left[\sum_{i=1}^{n}\left(u_i - \bar{u}\right)^2\right] = E\left[\sum_{i=1}^{n} u_i^2 - \tfrac{1}{n}\left(\sum_{i=1}^{n} u_i\right)^2\right]$$

$$= \sum_{i=1}^{n} E(u_i^2) - \tfrac{1}{n}\left(\sum_{i=1}^{n} E(u_i^2)\right) \qquad (\because \mathrm{E}(\mathrm{u_i u_j}) = 0)$$

$$= \sum_{i=1}^{n} var(u_i) - \tfrac{1}{n}\left(\sum_{i=1}^{n} var(u_i)\right) \qquad (\because \mathrm{E}(\mathrm{u_i}) = 0)$$

$$= \mathrm{n}\sigma^2 - \tfrac{1}{n}(n\sigma^2)$$

$$= (\mathrm{n}\text{-}1)\sigma^2$$

$$E\left[\left(\hat{\beta} - \beta\right)\sum_{i=1}^{n} x_i\left(u_i - \bar{u}\right)\right] = E\left[\frac{\sum x_i u_i}{\sum x_i^2}\sum x_i\left(u_i - \bar{u}\right)\right] \qquad \text{using(2.18)}$$

$$= E\left[\frac{\sum x_i u_i}{\sum x_i^2}\left(\sum x_i u_i - \bar{u}\sum x_i\right)\right]$$

$$= E\left[\frac{\left(\sum x_i u_i\right)^2}{\sum x_i^2}\right] \qquad \left(\because \sum x_i = 0\right)$$

$$= \frac{\sum x_i^2 E(u_i^2)}{\sum x_i^2} \qquad (\because E(u_i u_j) = 0)$$

$$= \sigma^2$$

Hence, $E\left[\sum_{i=1}^{n} e_i^2\right] = \sigma^2 + (\mathrm{n}\text{-}1)\sigma^2 - 2\sigma^2 = (\mathrm{n}\text{-}2)\sigma^2$

Thus the least squares estimator

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} \quad \text{is an unbiased estimator of } \sigma^2 \tag{2.35}$$

2.10 The Coefficient of Determination r^2:A Measure of "Goodness of Fit":

We now consider the **goodness of fit** of the fitted regression line to a set of data; that is, we shall find out how "well" the regression line fits the data. If all the sample observations (Y_i, X_i), $i = 1, 2, ..., n$, lie on the fitted regression line, then we say that the regression fit is "perfect" fit, but this is a very rare case. Generally, there will be some positive residuals and some negative residuals and let us hope that these residuals, around the fitted regression line, are as small as possible. The **coefficient of determination** r^2 (simple regression case) or R^2 (multiple regression case) is a summary measure that tells how well the regression line fits the data.

We have the residual,

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\alpha} - \hat{\beta} X_i \qquad i = 1, 2, ..., n \tag{2.36}$$

But we have the OLS estimator of α

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

which is substituted in Eq. (2.36) to get

$$\begin{aligned} e_i &= Y_i - \bar{Y} + \hat{\beta}\bar{X} - \hat{\beta}X_i \qquad i = 1, 2, ..., n \\ &= Y_i - \bar{Y} - \hat{\beta}\left(X_i - \bar{X}\right) \\ &= y_i - \hat{\beta}x_i \ , \quad \text{where } x_i = X_i - \bar{X} \ \& \ y_i = Y_i - \bar{Y} \end{aligned} \tag{2.37}$$

squaring Eq. (2.37) on both sides and summing over the sample we get

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} y_i^2 - 2\hat{\beta}\sum_{i=1}^{n} x_i y_i + \hat{\beta}^2 \sum_{i=1}^{n} x_i^2$$

Thus, the residual sum of squares is quadratic function of $\hat{\beta}$.

But we know that

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

There fore

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} y_i^2 - 2\hat{\beta}^2 \sum_{i=1}^{n} x_i^2 + \hat{\beta}^2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i^2 - \hat{\beta}^2 \sum_{i=1}^{n} x_i^2$$

which may be rewritten as

$$\sum_{i=1}^{n} y_i^2 = \hat{\beta}^2 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(\hat{\beta} x_i\right)^2 + \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2 \tag{2.38}$$

which is a famous decomposition of sum of squares and is usually writes as

$$\boxed{\text{TSS=ESS+RSS}} \tag{2.39}$$

where

TSS = total sum of squared deviation in Y variable

$= \sum_{i=1}^{n} y_i^2$

ESS = explained sum of squares from the regression of Y on X

$= \sum_{i=1}^{n} \hat{y}_i^2$

RSS = residual or unexplained sum of squares from the regression of Y on X

$= \sum_{i=1}^{n} e_i^2$

Now substituting $\hat{\beta} = \frac{\sum_{i=1}^{n} x_i \mathrm{y}_i}{\sum_{i=1}^{n} x_i^2}$ in Eq. (2.38), we get

$$\sum_{i=1}^{n} y_i^2 = \frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2} + \sum_{i=1}^{n} e_i^2 = \frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2 y_i^2} \sum_{i=1}^{n} y_i^2 + \sum_{i=1}^{n} e_i^2 \tag{2.40}$$

But, by definition the simple correlation coefficient 'r' is given by

$$r = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 y_i^2}} \tag{2.41}$$

Substituting Eq. (2.41) in Eq. (2.40), we get

$$\sum_{i=1}^{n} y_i^2 = r^2 \sum_{i=1}^{n} y_i^2 + \sum_{i=1}^{n} e_i^2 \quad \Rightarrow \quad 1 = r^2 + \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} y_i^2}$$

$$r^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} y_i^2} = 1 - \frac{RSS}{TSS} \tag{2.42}$$

$$\text{(OR)} \quad r^2 = \frac{\sum_{i=1}^{n} \hat{y}_i^2}{\sum_{i=1}^{n} y_i^2} = \frac{ESS}{TSS} \qquad (\text{using Eq. (2.39)})$$

In the above equation the square of correlation coefficient 'r^2' is usually called as **coefficient of determination.**

Note: The coefficient of determination obviously lies between 0 and 1. i.e., $0 < r^2 < 1$.

2.11 SELF ASSESSMENT QUESTIONS

1. Derive OLS estimators in a two variable linear model.
2. Explain the significance of the disturbance (error) term in a two variable regression model.
3. Explain the justification for the inclusion of disturbance (error) term in a simple linear model.
4. Show that OLS estimators of intercept and slope in a two variable regression model are unbiased.
5. Prove Gauss-Markov theorem in case of simple linear regression model.
6. Show that OLS estimators of intercept and slope in a two variable regression model are BLUEs.
7. Show that the sum of residuals is zero in a simple linear model.
8. Derive the least squares estimator of a regression coefficient and its variance in a two variable linear model.
9. Derive the normal equations for simple linear model.
10. Derive the least square estimate of the variance of the disturbance term in the single linear model and show that it is unbiased.
11. Distinguish between regression and correlation.
12. Let $\hat{\boldsymbol{\beta}}_{xy}$ and $\hat{\boldsymbol{\beta}}_{yx}$ represents the slopes in the regression of X and Y and Y on X respectively. Show that $\boldsymbol{\beta}_{xy}.\boldsymbol{\beta}_{yx} = r^2$ where $r^2 = r(x, y)$.
13. Define coefficient of determination in simple linear model.
14. Derive the coefficient of determination in simple linear model.
15. Derive an unbiased estimator of the variance of the disturbance term in the single linear model.

2.12 REFERENCES

1. Gujarati, D.N. (2005): *Basic Econometrics, 4th Ed., Tata McGraw-Hill.*
2. Johnston, J. (1984): *Econometric Methods, 3rd Ed., McGraw-Hill,* New York.
3. Montgomery, D.C., Peck, E.A. and Geoffrey Vining, G. (2003): *Introduction to Linear Regression Analysis, 3rd Ed., Wiley*
4. Draper, N.R., and H. Smith(1998): *Applied Regression Analysis, 3rd Ed., John Wiley & Sons,* New York.
5. G.S. Maddala (2001): *Introduction to Econometrics, 3rd Ed., John Wiley & Sons, Ltd.*
6. Johnston, J. and DiNardo J (1997): *Econometric Methods*, 4th Ed., McGraw Hill.
7. Hill, Carter, William Griffiths, and George Judge(2001): *Undergraduate Econometrics, John Wiley & Sons,* New York.
8. *Koutsoyiannis, A(1973): Theory of Econometrics, Harper & Row,* New York.

Lesson 3

SIMPLE REGRESSION ANALYSIS: TESTS OF SIGNIFICANCE AND PREDICTION

3.0 Objective:

The objective of this lesson is to derive the significance tests and confidence in travels for the slope (β) and intercept (α) of the simple linear model. Further, in this lesson, point and interval predictions are derived for the predictor (dependent) variable.

Structure of the Lesson:

3.1 Introduction

The results established in Lesson 2 are based on the assumption that the $u_i \sim iid\left(0,\sigma^2\right)$. Thus u_i is a random variable whose mean is zero and whose variance is σ^2. Further $u_i s$ are independently and identically distributed. But, the probability distribution function of u_i is not specified.

Now to carry out the tests of significance about the parameters of the regression model, we need further assumption about the probability distribution of the *u*'s. The standard assumption is that of normality, which may be justified by appeal to the Central Limit Theorem, since the u's represent the net effect of many separate but unmeasured influences.

Estimation of the regression model is half the battle and testing the fitted regression model is other half. In this lesson we discuss the testing of the regression model along with the prediction or forecasting.

3.2 The Sampling Distributions of the OLS Estimators:

Let us reconsider the two variable regression model (Eq. (2.4) in Lesson 2) along the additional assumption of the normality of the disturbance term u.

$$Y_i = \alpha + \beta X_i + u_i \qquad i = 1, 2, ..., n \tag{3.1}$$

where

i. $E(u_i) = 0 \quad \forall i$

ii. $\mathrm{var}(u_i) = E(u_i^2) = \sigma^2 \quad \forall\ i$.

iii. $\mathrm{cov}(u_i, u_j) = E(u_i u_j) = 0 \quad \forall\ i \neq j$

iv. u_i is normal (additional assumption)

the above assumptions may stated in compact form

$$u_i \overset{iid}{\sim} N\left(0, \sigma^2\right) \qquad \text{for } i = 1, 2, \ldots, n \tag{3.2}$$

From equation (2.18) (Lesson 2), the OLS estimator of β is given by

$$\hat{\beta} = \beta + \sum_{i=1}^{n} w_i u_i, \qquad \text{where } w_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2},\ x_i = \mathrm{X}_i - \bar{\mathrm{X}} \tag{3.3}$$

From the above equation, we may notice that the OLS estimator $\hat{\beta}$ is a linear combination of normal random disturbances $u_i s$. But, we know that every linear combination of a set of independent normal variates is also a normal variate. Thus from Eqs. (3.2) and (3.3), we may say that the sampling distribution of $\hat{\beta}$ is a normal variate whose mean and variance are given by

$$E(\hat{\beta}) = \beta + \sum_{i=1}^{n} w_i E\left(u_i\right) = \beta \qquad \left(\because E\left(u_i\right) = 0\right)$$

and $$\mathrm{var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2} \qquad \left(\text{from Eq. (2.22)}\right)$$

$$\boxed{\text{Therefore sampling distribution of } \hat{\beta} \text{ is } N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}\right)} \tag{3.4}$$

Similarly, from equation (2.20) of lesson 2, the OLS estimator of α

$$\hat{\alpha} = \alpha + \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{X} w_i\right) u_i \tag{3.5}$$

Since $\hat{\alpha}$ is a linear combination of independent normal random disturbances u_i's , $\hat{\alpha}$ is also a normal variate and its mean is given by

$$E\left(\hat{\alpha}\right) = \alpha \qquad \left(\because E\left(u_i\right) = 0\right)$$

From equation (2.23) of lesson 2, variance of $\hat{\alpha}$ is given by

$$\operatorname{var}(\hat{\alpha}) = \left(\frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2} \right) \sigma^2 \qquad (3.6)$$

Therefore the sampling distribution of $\hat{\alpha}$ is

$$\hat{\alpha} \sim N\left(\alpha, \frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2} \sigma^2 \right) \qquad (3.7)$$

Note: The standard deviation of the sampling distribution of an estimator is often referred to a standard error of the estimator.

3.3 Test of the Significance and confidence interval of β :

Since the sampling distribution of $\hat{\beta}$, from the above section, involves the unknown parameter σ^2, it is not operational as it stands. To derive the sampling distribution of $\hat{\beta}$, which does not depend on the unknown σ^2, we have to adopt the following two results.

i. $$\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} \sim \chi^2_{n-2} \qquad (3.8)$$

and ii. $\sum_{i=1}^{n} e_i^2$ is independently distributed of $\hat{\beta}$. (3.9)

From Eq. (3.4) of the above section, we may write

$$\frac{\hat{\beta} - \beta}{\sigma \Big/ \sqrt{\sum_{i=1}^{n} x_i^2}} \sim N(0,1) \qquad (3.10)$$

We know that the t-distribution is the ratio of a standard normal variate to the square root of an independent χ^2 variate divided by its degrees of freedom (d.f.). Therefore, from Eqs.(3.9) & (3.10) we may immediately write

$$t = \frac{(\hat{\beta} - \beta)\sqrt{\sum_{i=1}^{n} x_i^2} \Big/ \sigma}{\sqrt{\sum_{i=1}^{n} e_i^2} \Big/ \sigma\sqrt{n-2}} = \frac{(\hat{\beta} - \beta)\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{\sum_{i=1}^{n} e_i^2} \Big/ \sqrt{n-2}} \sim t \text{ distribution with n-2 d.f.}$$

But, an unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} \tag{3.11}$$

Therefore the above equation will be

$$t = \frac{\hat{\beta} - \beta}{\hat{\sigma} \Big/ \sqrt{\sum_{i=1}^{n} x_i^2}} = \frac{\hat{\beta} - \beta}{SE\left(\hat{\beta}\right)} \sim t\left(n-2\right) \tag{3.12}$$

where $SE\left(\hat{\beta}\right) = \hat{\sigma} \Big/ \sqrt{\sum_{i=1}^{n} x_i^2}$ (3.13)

is known as the standard error of $\hat{\beta}$, which is the square root of the estimated $\operatorname{var}\left(\hat{\beta}\right)$. Thus the standard normal variate given in Eq. (3.10), when σ is replaced by $\hat{\sigma}$, follows student-t distribution with n-2 d.f.

If we set the null hypothesis about the β as

$$H_0 : \beta = \beta_0$$

against the alternative hypothesis

$$H_1 : \beta \neq \beta_0$$

then Eq. (3.12) with $\beta = \beta_0$ may be used as the test statistic for testing the above H_0 and is given by

$$t = \frac{\hat{\beta} - \beta_0}{SE\left(\hat{\beta}\right)} \sim t\left(n-2\right) \tag{3.14}$$

Reject H_0 at 100ε percent level of significance (l.o.s.) if $|t| > t_{\varepsilon/2}\left(n-2\right)$, otherwise accept H_1. Here, $t_{\varepsilon/2}\left(n-2\right)$ is a two-tailed percentile of t distribution with n-2 d.f. at ε l.o.s. and is defined as

$$\Pr\left\{-t_{\varepsilon/2}\left(n-2\right) < t < t_{\varepsilon/2}\left(n-2\right)\right\} = \Pr\left\{|t| < t_{\varepsilon/2}\left(n-2\right)\right\} = 1 - \varepsilon$$

For instance, when $\varepsilon = 5\%$, we chose $t_{0.025}\left(n-2\right)$ such that

$$\Pr\left\{-t_{0.025}\left(n-2\right) < t < t_{0.025}\left(n-2\right)\right\} = 0.95$$

The hypothesis most frequently tested is

$H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$

and the test statistic can be obtained by substituting $\beta_0 = 0$ in Eq. (3.14) and is given by

$$\boxed{t = \frac{\hat{\beta}}{SE\left(\hat{\beta}\right)} \sim t\left(n-2\right) \qquad (3.15)}$$

Now, based on Eq. (3.15), we may draw the conclusion

$$If \quad |t| = \left| \frac{\hat{\beta}}{SE\left(\hat{\beta}\right)} \right| > t_{\varepsilon/2}\left(n-2\right) \text{ reject } H_0 : \beta = 0$$
$$\left(\text{or}\right) \text{ equvalently accept } H_1 : \beta \neq 0 \text{ at } \varepsilon \text{ l.o.s.} \tag{3.16}$$

Note:

1. If Eq. (3.16) is drawn as the conclusion, then we say that the regression coefficient β is significant and in this case, the regression model Eq. (3.1) is said to be well fitted.
2. The two-tailed t value $t_{\varepsilon/2}\left(n-2\right)$ can be obtained from student-t table at given ε and n-2 d.f.

The $100\left(1\text{-}\varepsilon\right)\%$ confidence interval of β:

A $100\left(1\text{-}\varepsilon\right)\%$ confidence interval for β, based on Eq. (3.12), is given by

$$\left(\hat{\beta} - t_{\varepsilon/2}\left(n-2\right)SE\left(\hat{\beta}\right), \quad \hat{\beta} + t_{\varepsilon/2}\left(n-2\right)SE\left(\hat{\beta}\right)\right) \tag{3.17}$$

3.4 Test of the Significance and confidence interval of α:

From Eq. (3.7) of section Eq. (3.2), we may write

$$\frac{\hat{\alpha} - \alpha}{\sigma\sqrt{\sum_{i=1}^{n} X_i^2 \Big/ \left(n\sum_{i=1}^{n} x_i^2\right)}} \sim N\left(0,1\right) \tag{3.18}$$

We know that the t-distribution is the ratio of a standard normal variate to the square root of an independent χ^2 variate divided by its d.f. Therefore, from Eqs. (3.9) and (3.18) we may write

$$t = \frac{\hat{\alpha} - \alpha}{\sigma\sqrt{\sum_{i=1}^{n} X_i^2 \Big/ \left(n\sum_{i=1}^{n} x_i^2\right)}} \frac{\sigma}{\sqrt{\sum_{i=1}^{n} e_i^2 \Big/ n-2}} \sim t\left(n\text{-}2\right)$$

The above equation may be rewritten as

$$t = \frac{\hat{\alpha} - \alpha}{\hat{\sigma}\sqrt{\sum_{i=1}^{n} X_i^2 \Big/ \left(n\sum_{i=1}^{n} x_i^2\right)}} \sim t\left(n-2\right) \tag{3.19}$$

where $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$ is an unbiased estimator of σ^2

Thus the standard normal variate given by Eq. (3.18), in which σ is replaced by $\hat{\sigma}$, follows student-t distribution with n-2 d.f.

The test statistic for testing $H_0 : \alpha = \alpha_0$ vs $H_1 : \alpha \neq \alpha_0$ may be obtained from the above Eq. (3.19) as

$$t = \frac{\hat{\alpha} - \alpha_0}{SE(\hat{\alpha})} \sim t(n-2) \qquad (3.20)$$

where $SE(\hat{\alpha}) = \hat{\sigma}\sqrt{\sum_{i=1}^{n} X_i^2 \Big/ \left(n\sum_{i=1}^{n} x_i^2\right)}$ (3.21)

is the standard error of $\hat{\alpha}$

Reject H_0 at 100ε percent level of significance (l.o.s.) if $|t| > t_{\varepsilon/2}(n-2)$, otherwise accept H_1.

The hypothesis most frequently tested is

$H_0 : \alpha = 0$ vs $H_1 : \alpha \neq 0$

and the test statistic can be obtained by substituting $\alpha_0 = 0$ in Eq. (3.20) and is given by

$$t = \frac{\hat{\alpha}}{SE(\hat{\alpha})} \sim t(n-2) \qquad (3.22)$$

$$\text{If } |t| = \left|\frac{\hat{\alpha}}{SE(\hat{\alpha})}\right| > \mathbf{t}_{\varepsilon/2}(n-2) \text{ Reject } H_0 : \alpha = 0$$
$$(\text{or}) \text{ equivalently Accept } H_1 : \alpha \neq 0 \text{ at } \varepsilon \text{ l.o.s.} \qquad (3.23)$$

Note: If the conclusion Eq. (3.23) is drawn, then we say that the intercept (constant) α is significant.

The $100(1\text{-}\varepsilon)\%$ confidence interval of α :

A $100(1\text{-}\varepsilon)\%$ confidence interval for α , based on Eq. (3.19), is given by

$$\left(\hat{\alpha} - t_{\varepsilon/2}(n-2)SE(\hat{\alpha}),\ \hat{\alpha} + t_{\varepsilon/2}(n-2)SE(\hat{\alpha})\right) \qquad (3.24)$$

3.5 Prediction in the Least-Square Model:

Since $\hat{\alpha}$ and $\hat{\beta}$ are the BLUEs of α and β the optimum point (BLUE) prediction is given by the regression value corresponding to X_f , that is

$$\hat{Y}_f = \hat{\alpha} + \hat{\beta} X_f \qquad (3.25)$$

The true value of Y in the prediction period is given by

$$Y_f = \alpha + \beta X_f + u_f$$

where u_f indicates the value that would be drawn from the disturbance distribution in the prediction period. The prediction error may then be defined as

$$\begin{aligned} e_f &= Y_f - \hat{Y}_f \\ &= u_f - (\hat{\alpha} - \alpha) - (\hat{\beta} - \beta) X_f \end{aligned} \tag{3.26}$$

since $E(u_f) = 0$ and $\hat{\alpha}$ and $\hat{\beta}$ are unbiased estimators of α and β, we have

$$E(e_f) = 0 \Rightarrow E(\hat{Y}_f) = E(Y_f) = \alpha + \beta X_t \qquad (\because \mathrm{E}(\mathrm{u_t}) = 0) \tag{3.27}$$

Thus the least-squares predictor $\hat{Y}_f$ given in Eq. (3.25) is an unbiased predictor of $E(Y_f)$. The variance of the prediction error is given by (using Eq. (3.27))

$$\begin{aligned} \operatorname{var}(e_f) &= E(e_f^2) \\ &= E(u_f^2) + E(\hat{\alpha} - \alpha)^2 + X_f^2 E(\hat{\beta} - \beta)^2 - 2E[u_f(\hat{\alpha} - \alpha)] \\ &\quad - 2X_f E[u_f(\hat{\beta} - \beta)] + 2X_f E[(\hat{\alpha} - \alpha)(\hat{\beta} - \beta)] \end{aligned}$$

But we known that both $(\hat{\alpha} - \alpha)$ and $(\hat{\beta} - \beta)$ are linear functions of $u_1, u_2, \ldots, u_n$ and by assumption u_f is independent of $u_1, u_2, \ldots, u_n$. Therefore u_f is uncorrelated with $(\hat{\alpha} - \alpha)$ and $(\hat{\beta} - \beta)$ and thus

$$E[u_f(\hat{\alpha} - \alpha)] = 0 \textit{ and } E[u_f(\hat{\beta} - \beta)] = 0$$

Therefore the above equation becomes

$$\operatorname{var}(e_f) = \operatorname{var}(u_f) + \operatorname{var}(\hat{\alpha}) + X_f^2 \operatorname{var}(\hat{\beta}) + 2X_f \operatorname{cov}(\hat{\alpha}, \hat{\beta}) \tag{3.28}$$

But $\operatorname{var}(u_f) = \sigma^2$ we have from equation (2.22), (2.23) and (2.24)

$$\operatorname{var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}, \operatorname{var}(\hat{\alpha}) = \left(\frac{\sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} x_i^2} \right) \sigma^2 \textit{ and } \operatorname{cov}(\hat{\alpha}, \hat{\beta}) = -\frac{\bar{X}\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

But, $\operatorname{var}(\hat{\alpha})$ may be rewritten as

$$\operatorname{var}(\hat{\alpha}) = \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2} \right) \sigma^2$$

Substituting $\operatorname{var}(u_f)$, $\operatorname{var}(\hat{\alpha}), \operatorname{var}(\hat{\beta})$ and $\operatorname{cov}(\hat{\alpha}, \hat{\beta})$ in Eq. (3.28) we get

$$\operatorname{var}(e_f) = \sigma^2 \left[1 + \frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} x_i^2} + \frac{X_f^2}{\sum_{i=1}^{n} x_i^2} - \frac{2X_f \overline{X}}{\sum_{i=1}^{n} x_i^2} \right] = \sigma^2 \left[1 + \frac{1}{n} + \frac{\left(X_f - \overline{X}\right)^2}{\sum_{i=1}^{n} x_i^2} \right] \qquad (3.29)$$

The variance of the prediction error is thus at its minimum value when $X_f = \overline{X}$ and increases nonlinearly as X_f departs from $\overline{X}$. From Eq. (3.26) e_f is seen to be a linear function of normal variables and so if is itself distributed normally. Thus

$$\frac{e_f}{\sigma \sqrt{\left[1 + \frac{1}{n} + \frac{\left(X_f - \overline{X}\right)^2}{\sum_{i=1}^{n} x_i^2} \right]}} \sim N(0,1)$$

Replacing the unknown σ by its estimate $\hat{\sigma} = \sqrt{\sum_{i=1}^{n} e^2 \Big/ (n-2)}$ then gives

$$\frac{Y_f - \hat{Y}_f}{\hat{\sigma} \sqrt{\left[1 + \frac{1}{n} + \frac{\left(X_f - \overline{X}\right)^2}{\sum_{i=1}^{n} x_i^2} \right]}} \sim t(n-2) \qquad (3.30)$$

Everything is Eq. (3.30) is known except Y_f and so, in the usual way, we derive **a** $100(1-\varepsilon)$ **percent confidence interval of** Y_f as

$$\boxed{\left(\left(\hat{\alpha} + \hat{\beta} X_f\right) - t_{\varepsilon/2} \hat{\sigma} \sqrt{\left[1 + \frac{1}{n} + \frac{\left(X_f - \overline{X}\right)^2}{\sum_{i=1}^{n} x_i^2} \right]},\ \left(\hat{\alpha} + \hat{\beta} X_f\right) + t_{\varepsilon/2} \hat{\sigma} \sqrt{\left[1 + \frac{1}{n} + \frac{\left(X_f - \overline{X}\right)^2}{\sum_{i=1}^{n} x_i^2} \right]} \right) \qquad (3.31)}$$

where $t_{\varepsilon/2}(n-2)$ indicates the $100\varepsilon/2$ percent point of the t distribution with (n-2) degrees of freedom.

Sometimes interest centers on predicting the *mean* value of Y_f, that is

$$E(Y_f) = \alpha + \beta X_f$$

Rather than Y_f itself, since there is, of course, no way of predicting the value of a single drawing from $p(u)$. The prediction error is now

$$e_f = E(Y_f) - \hat{Y}_f$$
$$= -(\hat{\alpha} - \alpha) - (\hat{\beta} - \beta) X_f$$

which gives

$$\operatorname{var}(e_f) = \sigma_u^2 \left[\frac{1}{n} + \frac{(X_f - \overline{X})^2}{\sum_{i=1}^{n} x_i^2} \right]$$

and so **a** $100(1-\varepsilon)$ **percent confidence interval for** $E(Y_f)$ is

$$\left(\hat{\alpha} + \hat{\beta} X_f - t_{\varepsilon/2} \hat{\sigma} \sqrt{\left[\frac{1}{n} + \frac{(X_f - \overline{X})^2}{\sum_{i=1}^{n} x_i^2} \right]},\ \hat{\alpha} + \hat{\beta} X_f + t_{\varepsilon/2} \hat{\sigma} \sqrt{\left[\frac{1}{n} + \frac{(X_f - \overline{X})^2}{\sum_{i=1}^{n} x_i^2} \right]} \right) \quad (3.32)$$

Note:

1. For the choice of $\varepsilon = 0.05$, the confidence intervals given in Eqs. (3.31) and (3.32) are respectively the 95% confidence intervals for Y_f and $E(Y_f)$.
2. Similarly for choice of $\varepsilon = 0.01$, the confidence intervals given in Eqs. (3.31) and (3.32) are respectively the 99% confidence intervals for Y_f and $E(Y_f)$.

3.6 SELF ASSESSMENT QUESTIONS

1. Derive the test for the significance of regression coefficient is simple linear model.
2. Show that, under the assumption of normality of error term, the least squares estimators of the regression coefficients are normally distributed.
3. Derive the tests for the significance of OLS estimators in a two variable linear regression model.
4. Derive the sampling distribution of the OLS estimators in simple linear model.
5. Derive the confidence intervals for the coefficients of the simple linear model.
6. Obtain the confidence interval for the slope of the simple linear model, making the necessary assumptions.
7. Obtain the confidence interval for the intercept of the simple linear model, making the necessary assumptions.
8. Discuss the point prediction in the simple linear model.
9. Construct interval prediction for the regressand in the simple linear model.
10. Construct interval prediction for the mean of the regressand in the simple linear model.
11. Discuss the forecasting problem in the simple linear model.
12. Obtain the predictor for the mean of the regressand in the simple linear model.
13. Obtain the predictor for the regressand in the simple linear model.

3.7 REFERENCES

1. Gujarati, D.N. (2005): *Basic Econometrics, 4th Ed., Tata McGraw-Hill.*
2. Johnston, J. (1984): *Econometric Methods, 3rd Ed., McGraw-Hill,* New York.
3. Montgomery, D.C., Peck, E.A. and Geoffrey Vining, G. (2003): *Introduction to Linear Regression Analysis, 3rd Ed., Wiley*
4. Draper, N.R., and H. Smith(1998): *Applied Regression Analysis, 3rd Ed., John Wiley & Sons,* New York.
5. G.S. Maddala (2001): *Introduction to Econometrics, 3rd Ed., John Wiley & Sons, Ltd.*
6. Johnston, J. and DiNardo J (1997): *Econometric Methods*, 4th Ed., McGraw Hill.
7. Hill, Carter, William Griffiths, and George Judge(2001): *Undergraduate Econometrics, John Willey & Sons,* New York.
8. *Koutsoyiannis, A(1973): Theory of Econometrics, Harper & Row,* New York.

Lesson 4

OTHER FUNCTIONAL FORMS OF REGRESSION MODELS

4.0 Objective:

The objective of this lesson is to make the student familiar with some commonly used regression models, those may be nonlinear in the variables but are linear in the parameters. Some important functional forms of regression models discussed in this lesson are a) The log-linear model, b) semi log models, and c) reciprocal models.

Structure of the Lesson:

4.1 Introduction

As already mentioned in the earlier lessons, this text book is concerned primarily with models that are linear in the parameters; they may or may not be linear in the variables. In the following sections, we consider some commonly used regression models that may be nonlinear in the variables but are linear in the parameters or that can be made so by suitable transformations of the variables. In particular, we discuss the following regression models:

1. The log-linear or constant elasticity model
2. Semilog regression models
3. Reciprocal regression models.

We discuss the special features of each model, when they are appropriate, and how they are estimated.

In the log-linear model both the regressand and the regressor(s) are expressed in the logarithmic form. The regression coefficient attached to the log of a regressor is interpreted as the elasticity of the regressand with respect to the regressor.

In the semilog model either the regressand or the regressor(s) are in the log form. In the semilog model where the regressand is logarithmic and the regressor X is time, the estimated slope coefficient (multiplied by 100) measures the (instantaneous) rate of growth of the regressand. Such models are often used to measure the growth rate of many economic phenomena. In the semilog model if the regressor is logarithmic, its coefficient measures the absolute rate of change in the regressand for a given percent change in the value of the regressor.

In the reciprocal models, either the regressand or the regressor is expressed in reciprocal, or inverse, form to capture nonlinear relationships between economic variables.

4.2 The Log-Linear Model:

Consider the following model, known as the **exponential regression model:**

$$Y_i = \beta_0 X_i^{\beta} e^{u_i}, \quad i = 1, 2, \ldots, n \tag{4.1}$$

which may be expressed alternatively as

$$\log Y_i = \log \beta_0 + \beta \log X_i + u_i, \qquad i = 1, 2, \ldots, n \tag{4.2}$$

where $\log$ = natural log (i.e., log to the base *e*, and where *e* = 2.718).

If we write Eq. (4.2) as

$$\log Y_i = \alpha + \beta \log X_i + u_i, \qquad i = 1, 2, \ldots, n \tag{4.3}$$

where $\alpha = \log \beta_0$, and this model is linear in the parameters α and β and linear in the logarithms of the variables *Y* and *X*, and can be estimated by OLS regression. Because of this linearity, such models are called **log-log, double-log,** or **log-linear** models.

If the assumptions of the classical linear regression model are fulfilled, the parameters of Eq. (4.3) can be estimated by applying the OLS method to the following model

$$Y_i^* = \alpha + \beta X_i^* + u_i, \qquad i = 1, 2, \ldots, n \tag{4.4}$$

where $Y^* = \log Y_i$ and $X^* = \log X_i$. The OLS estimators $\hat{\alpha}$ and $\hat{\beta}$ will be the BLUEs of α and β respectively and are given by

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i^* Y_i^* - \bar{X}^* \bar{Y}^*}{\frac{1}{n}\sum_{i=1}^{n} X_i^{*2} - \bar{X}^{*2}} \tag{4.5}$$

$$\hat{\alpha} = \bar{Y}^* - \hat{\beta}\bar{X}^* \Rightarrow \hat{\beta}_0 = e^{\hat{\alpha}} \tag{4.6}$$

$$\text{where } \bar{X}^* = \frac{1}{n}\sum_{i=1}^{n} \log X_i$$

$$\bar{Y}^* = \frac{1}{n}\sum_{i=1}^{n} \log Y_i$$

One attractive feature of the log-log model, which has made it popular in applied work, is that the slope coefficient β measures the **elasticity** of *Y* with respect to X , that is, the percentage change in Y for a given (small) percentage change in X. Thus, if Y represents the quantity of a commodity demanded and X its unit price, β measures the price elasticity of demand, a parameter of considerable economic interest.

Two special features of the log-linear model may be noted: The model assumes that the elasticity coefficient between Y and X , β , remains constant throughout, hence the alternative name **constant elasticity model.** In other words, the change in $\log Y$ per unit change in $\log X$ (i.e., the elasticity, β) remains the same no matter at which $\log X$ we measure the elasticity. Another feature of the model is that although $\hat{\alpha}$ and $\hat{\beta}$ are unbiased estimates of α and β, β_0 (the parameter entering the original model) when estimated as $\hat{\beta}_0 = e^{\hat{\alpha}}$ is itself a biased estimator. In most practical problems, however, the intercept term is of secondary importance, and one need not worry about obtaining its unbiased estimate.

In the two-variable model, the simplest way to decide whether the loglinear model fits the data is to plot the scatter diagram of $\log Y$ against $\log X$,and see if the scatter points lie approximately on a straight line.

4.3 The Semi Log Models: Log-Lin and Lin-Log Models

Log–Lin Model (to Measure the Growth Rate):

Economists, businesspeople, and governments are often interested in finding out the rate of growth of certain economic variables, such as population, GNP, money supply, employment, productivity, and trade deficit.

We may recall the following well-known compound interest formula from your introductory course in economics.

$$Y_t = Y_0(1 + r)^t \qquad t = 1, 2, \ldots, n \tag{4.7}$$

Taking natural log on both side

$$\log Y_t = \log Y_0 + t \log(1+r) \qquad t = 1, 2, \ldots, n \tag{4.8}$$

Now letting

$$\alpha = \log Y_0$$

$$\beta = \log(1+r)$$

and adding the disturbance term to Eq. (4.8) we obtain

$$\log Y_t = \alpha + \beta t + u_t \qquad t = 1, 2, \ldots, n \tag{4.9}$$

This model is like any other linear regression model in that the parameters α and β are linear. The only difference is that the regressand is the logarithm of *Y* and the regressor is “time,” which will take values of 1, 2, 3, etc.

Models like Eq. (4.9) are called **semilog models** because only one variable (in this case the regressand) appears in the logarithmic form. For descriptive purposes a model in which the regressand is logarithmic will be called a **log-lin model.** In this model *the slope coefficient measures the constant proportional or relative change in* Y *for a given absolute change in the value of the regressor* (in this case the variable *t*), that is,

$$\beta = \frac{\text{relative change in regressand}}{\text{absolute change in regressor}}$$

The OLS Estimators of the parameters of α and β of Eq. (4.9) are given by

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{t=1}^{n} t \log Y_t \;-\overline{\log Y}\left(\frac{n+1}{2}\right)}{\frac{1}{n}\sum_{i=1}^{n} t^2 - \left(\frac{n+1}{2}\right)^2} \qquad t = 1, 2, \ldots, n \qquad (4.10)$$

$$\hat{\alpha} = \overline{\log Y} - \hat{\beta}\left(\frac{n+1}{2}\right) \qquad Where\ \overline{\log Y} = \frac{1}{n}\sum_{t=1}^{n} \log Y_t \qquad (4.11)$$

It may be noted the above OLS Estimators $\hat{\alpha}$ and $\hat{\beta}$ are the BLUEs of α and β respectively

Lin–Log Model:
Unlike the growth model just discussed, in which we were interested in finding the percent growth in Y for an absolute change in X, suppose we now want to find the absolute change in Y for a percent change in X. A model that can accomplish this purpose can be written as:

$$Y_i = \alpha + \beta \log X_i \; + u_i \quad i = 1, 2, \ldots, n \qquad (4.12)$$

For descriptive purposes we call such a model a **lin–log model.**
The interpretation of the slope coefficient β is

$$\beta = \frac{\text{change in Y}}{\text{change in logX}}$$

$$= \frac{\text{change in Y}}{\text{realative change in X}}$$

The second step follows from the fact that *a change in the log of a number is a relative change.*
The OLS Estimators α and β respectively given by

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i \;\log X_i \; - n\overline{\log X}\overline{Y}}{\frac{1}{n}\sum_{i=1}^{n}\left(\log X_i\right)^2 - \overline{\log X}^2} \qquad (4.13)$$

$$\hat{\alpha} = \overline{\mathrm{Y}} - \hat{\beta}\overline{\log X} \qquad (4.14)$$

$$where \;\; \overline{\log X} = \sum_{i=1}^{n} \log X_i$$

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

It may be noted the above OLS Estimators $\hat{\alpha}$ and $\hat{\beta}$ are the BLUEs of α and β respectively

4.4 The Reciprocal Models

Models of the following type are known as **reciprocal** models.

$$Y_i = \alpha + \beta\left(\frac{1}{X_i}\right) + u_i \qquad i = 1, 2, \ldots, n \qquad (4.15)$$

Although this model is nonlinear in the variable X because it enters inversely or reciprocally, the model is linear in α and β and is therefore a linear regression model.

This model has these features: As X increases indefinitely, the term β (1/X) approaches zero (*note:* β is a constant) and Y approaches the limiting or *asymptotic* value α . Therefore, models like (4.15) have built in them an **asymptote** or limit value that the dependent variable will take when the value of the X variable increases indefinitely.

The OLS Estimators β and α of the reciprocal (4.15) are respectively by

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i \,/ X_i\right) - \overline{X}^{*}\overline{Y}}{\frac{1}{n}\sum_{i=1}^{n} 1\big/X_i^2 - \overline{X}^{*2}}$$

$$\hat{\alpha} = \overline{\mathrm{Y}} - \hat{\beta}\overline{X}^{*} \quad where\ \overline{X}^{*} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{X_i},\ \text{and}\ \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{4.16}$$

It may be noted the above OLS Estimators $\hat{\alpha}$ and $\hat{\beta}$ are the BLUEs of α and β respectively

4.5 Self Assessment Questions

1. Explain the following models
   (i) Log-linear, (ii) Semilog and (ii) Reciprocal
2. Derive the estimators in exponential regression model.
3. Estimate the elasticity in a log-linear (exponential regression) model.
4. Distinguish between linear and log-linear models.
5. Explain the estimation method of a log-linear model.
6. Explain the estimation method of an exponential model.
7. Derive the least squares estimators in linear-log model.
8. Derive the least squares estimators in log-linear model.
9. Derive the least squares estimators in reciprocal regression model.
10. Distinguish between log-linear and semi log-linear models.
11. Distinguish between linear and reciprocal models.

4.6 References

1. Gujarati, D.N. (2005): *Basic Econometrics, 4th Ed., Tata McGraw-Hill.*
2. Johnston, J. (1984): *Econometric Methods, 3rd Ed., McGraw-Hill,* New York.
3. Montgomery, D.C., Peck, E.A. and Geoffrey Vining, G. (2003): *Introduction to Linear Regression Analysis, 3rd Ed., Wiley*
4. Draper, N.R., and H. Smith(1998): *Applied Regression Analysis, 3rd Ed., John Wiley & Sons,* New York.
5. G.S. Maddala (2001): *Introduction to Econometrics, 3rd Ed., John Wiley & Sons, Ltd.*
6. Johnston, J. and DiNardo J (1997): *Econometric Methods*, 4th Ed., McGraw Hill.
7. Hill, Carter, William Griffiths, and George Judge(2001): *Undergraduate Econometrics, John Willey & Sons,* New York.
8. Koutsoyiannis, A(1973): Theory of Econometrics, Harper & Row, New York.

Lesson 5

APPLICATIONS OF SIMPLE LINEAR REGRESSION ANALYSIS

5.0 Objective:

The objective of this lesson is to demonstrate the student the computations involved in simple linear regression analysis, which was explained through Lessons 2- 4 with some practical applications.

Structure of the Lesson:

5.1 Introduction

In Lesson 2, we have discussed the estimation of simple linear model where as in Lesson 3, we have discussed the testing the significance of the estimated regression model. In Lesson 4, we have discussed the estimation of some non-linear models, those can be transformed into simple linear model with some simple transformation of the variables of the model.

Now in this lesson, we demonstrate the simple linear regression analysis discussed in Lessons 2, 3 and 4 with some suitable applications. The computation pertaining to an example of simple linear model are presentation in Section 5.2 and those pertaining to an example of a non-linear model are presented in Section 5.3.

5.2 Estimation of the Consumption function for United States:

The following table gives the data on percapita disposable income (income after deducting income tax) (X) and percapita consumption expenditure (Y) both in constant dollars for the United States. Using this data estimate the consumption function for United States for the period 1970-1984 and using this estimated consumption function predict the percapita consumption for the year 1985 at a given percapita disposable income of 5,100 US dollars.

Table 5.1: Per capita personal consumption expenditure (Y) and per capita disposable personal income (X) (in 1972 dollars) for the United States, 1970-1984

Year	Y	X
1970	3277	3665
1971	3355	3752
1972	3511	3860
1973	3623	4080
1974	3566	4009
1975	3609	4051
1976	3774	4158
1977	3924	4280
1978	4057	4441
1979	4121	4512
1980	4093	4487
1981	4131	4561
1982	4146	4555
1983	4303	4670
1984	4490	4941

Source: Economic Report of the President, 1984, p. 261.

Solution:

We will carry out below the regression analysis of per capita consumption expenditure (Y) on per capita disposable income (X) .

Year	Y	X	Y^2	X^2	XY
1970	3277	3665	10738729	13432225	12010205
1971	3355	3752	11256025	14077504	12587960
1972	3511	3860	12327121	14899600	13552460
1973	3623	4080	13126129	16646400	14781840
1974	3566	4009	12716356	16072081	14296094
1975	3609	4051	13024881	16410601	14620059
1976	3774	4158	14243076	17288964	15692292
1977	3924	4280	15397776	18318400	16794720
1978	4057	4441	16459249	19722481	18017137
1979	4121	4512	16982641	20358144	18593952
1980	4093	4487	16752649	20133169	18365291
1981	4131	4561	17065161	20802721	18841491
1982	4146	4555	17189316	20748025	18885030
1983	4303	4670	18515809	21808900	20095010
1984	4490	4941	20160100	24413481	22185090
Total	57980	64022	225955018	275132696	249318631

Computation of regression coefficients:
From the above table we have

n = 15,

$\sum Y$ = 57980 $\sum X$ = 64022

$\sum Y^2$ = 225955018 $\sum X^2$ = 275132696

$\sum XY$ = 249318631

$$\bar{X} = \frac{64022}{15} = 4268.1333 \qquad \bar{Y} = \frac{57980}{15} = 3865.333$$

Substituting the above values in Eq. (2.12) of Lesson 2 we get OLS Estimate of β

$$\hat{\beta} = \frac{249318631/15 - (4268.1333*3865.3333)}{275132696/15 - 4268.1333^2}$$

$$= \frac{123484.0222}{125217.5822} = 0.9862$$

Similarly substituting $\hat{\beta}$, $\bar{X}$ and $\bar{Y}$ in Eq. (2.13) of Lesson 2 we get

$$\hat{\alpha} = 3865.333333\text{-}0.9862*4268.1333 = \text{-}343.90$$

Hence the estimated consumption function for United States of the period 1970-1984 is

$$\hat{Y} = \text{-}343.90 + 0.9862 * X$$

In the estimated regression $\hat{\beta} = 0.9862$ is the estimate of the marginal propensity to consume, which means on the average 98.6 dollars will be the consumption expenditure for every 100 dollars of disposable income.

Computation of $\hat{\sigma}^2$:

From Eq. (2.35) of Lesson 2 the unbiased estimator of $\hat{\sigma}^2$ is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}, \quad \text{where } e_i = Y_i - \hat{Y}_i, \qquad \hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

Computation of residuals $e_i s$:

Y	$\hat{Y}$	$e = Y - \hat{Y}$	e^2
3277	3270.52	-6.48	41.99
3355	3356.32	1.32	1.74
3511	3462.83	-48.17	2320.35
3623	3679.8	56.8	3226.24
3566	3609.78	43.78	1916.69
3609	3651.2	42.2	1780.84
3774	3756.72	-17.28	298.6
3924	3877.04	-46.96	2205.24
4057	4035.81	-21.19	449.02
4121	4105.83	-15.17	230.13
4093	4081.18	-11.82	139.71
4131	4154.16	23.16	536.39

4146	4148.24	2.24	5.02
4303	4261.65	-41.35	1709.82
4490	4528.91	38.91	1513.99
	TOTALS	-0.01	16375.77

$$\hat{\sigma}^2 = \frac{16375.77}{13} = 1259.675$$

The Computation of Coefficient of Determination r^2:

From Eq. (2.42) of Lesson 2 we have the coefficient of determination r^2

$$r^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} y_i^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} y_i^2 - n\overline{Y}^2} = 1 - \frac{16375.77}{225955018 - 15\left(3865.3333\right)^2} = 0.9911$$

Alternative computation of r^2:

Since r is nothing but the correlation coefficient between X and Y, by definition r^2 is

$$r^2 = \frac{\left[\operatorname{cov}(X,Y)\right]^2}{\operatorname{var}(X)\operatorname{var}(Y)} = \beta^2 \frac{\operatorname{var}(X)}{\operatorname{var}(Y)}$$

$$= \beta^2 \frac{\sum_{i=1}^{n} X^2 \Big/ n - \overline{X}^2}{\sum_{i=1}^{n} Y^2 \Big/ n - \overline{Y}^2} = \left(0.9862\right)^2 \frac{275132696/15 - 4268.1333^2}{225955018/15 - 3865.3333^2} = 0.9911$$

Testing the significance of regression coefficient (β) at 5 level of significance

Suppose we want to test the significance β we set the null hypothesis

$$H_0 : \beta = 0 \text{ vs } H_1 : \beta \neq 0$$

The test statistic for testing the above hypothesis is given by (from equation (3.15) of Lesson 3)

$$t = \frac{\hat{\beta}}{SE\left(\hat{\beta}\right)} = \frac{0.9862}{SE\left(\hat{\beta}\right)}$$

where from Eq.(3.13),

$$SE\left(\hat{\beta}\right) = \hat{\sigma} \Big/ \sqrt{\sum_{i=1}^{n} x_i^2} = \hat{\sigma} \Big/ \sqrt{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}$$

$$= \frac{\sqrt{1259.675}}{\sqrt{275132696 - \left(14 * 4268.13333^2\right)}}$$

$$= 0.281$$

$$t = \frac{0.9862}{SE\left(\hat{\beta}\right)} = \frac{0.9862}{0.281} = 3.5096$$

From student t-table, at $\varepsilon = 5\%$ level of significance

$$t_{0.025}(13) = 2.160$$

Since $|t| = 3.5096 > t_{0.025}(13) = 2.160$ from equation Eq. (3.16) of Lesson 3,

we will reject $H_0 : \beta = 0$.

Thus we conclude the regression coefficient β is significantly different from zero, which establishes the linear influence of X on Y. <u>Hence, the model is well fitted.</u>

The 95% confidence interval of β :

From Eq. (3.17) of Lesson 3, the 95% confidence interval of β is

$$\left[\hat{\beta} - t_{0.025}(13)\, SE\left(\hat{\beta}\right),\ \ \hat{\beta} + t_{0.025}(13)\, SE\left(\hat{\beta}\right)\right]$$

$$i.e.\quad [0.9862 - 2.160 * 0.281,\ 0.9862 + 2.160 * 0.281]$$

$$i.e.\quad [0.3792,\ 1.5932]$$

Testing the significance of intercept (α) at 5% level of significance:

Suppose we want to test the significance of α we set the null hypothesis

$$H_0 : \alpha = 0 \text{ vs } H_1 : \alpha \neq 0$$

The test statistic for testing the above hypothesis is given by (from Eq. (3.22) of Lesson 3)

$$t = \frac{\hat{\alpha}}{SE(\hat{\alpha})} = \frac{-343.90}{1245.7163} = -0.2761$$

where

$$SE(\hat{\alpha}) = \hat{\sigma}\sqrt{\sum_{i=1}^{n} X_i^2 \Big/ \left(n\sum_{i=1}^{n} x_i^2\right)} = \hat{\sigma}\sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)}}$$

$$= 1259.675 * \sqrt{\frac{275132696}{14 * \left[275132696 - \left(14 * 4268.13333^2\right)\right]}} = 1245.7163$$

We have $t_{0.025}(13) = 2.160$

Since $|t| = 0.2761 < t_{0.025}(13) = 2.160$

from equation Eq. (3.16) of Lesson 3, we accept $H_0 : \alpha = 0$.

Thus we conclude the intercept or constant term α is not significant.

Note: The insignificance of the constant term α , however, does not influence the above conclusion that the fitted model is a good one.

The point prediction of the consumption expenditure for the year 1985 at given per capita disposable income X_t =5100 US dollars can be obtained from Eq. (3.25) as

$$\hat{Y}_{1985} = \text{-}343.90 + 0.9862 * X_{1985} = \text{-}343.9 + 0.9862 * 5100 = 4685.72$$

95% Interval prediction for Y_{1985} from Eq. (3.31)

$$\left(\hat{Y}_{1985} - t_{0.025}\hat{\sigma}\sqrt{\left[1 + \frac{1}{n} + \frac{\left(X_{1985} - \overline{X}\right)^2}{\sum_{i=1}^{n} x_i^2}\right]},\ \hat{Y}_{1985} + t_{0.025}\hat{\sigma}\sqrt{\left[1 + \frac{1}{n} + \frac{\left(X_{1985} - \overline{X}\right)^2}{\sum_{i=1}^{n} x_i^2}\right]} \right)$$

We have $\frac{1}{n} + \frac{\left(X_{1985} - \overline{X}\right)^2}{\sum_{i=1}^{n} x_i^2} = \frac{1}{15} + \frac{\left(5100 - 4268.1333\right)^2}{275132696 - 15 * 4268.1333^2}$ =0.4351

Therefore, 95% Interval prediction for Y_{1985} is

$$\left(4685.72 - 2.160 * \sqrt{1259.675} * \sqrt{[1 + 0.4351]},\ \ 4685.72 - 2.160 * \sqrt{1259.675} * \sqrt{[1 + 0.4351]}\right)$$

$$= \left(4593.882,\ 4777.558\right)$$

Similarly, a 95% confidence interval for $E(Y_{1985})$ from Eq. (3.32) is

$$\left(\hat{Y}_{1985} - t_{0.025}\hat{\sigma}\sqrt{\left[\frac{1}{n} + \frac{\left(X_{1985} - \overline{X}\right)^2}{\sum_{i=1}^{n} x_i^2}\right]},\ \hat{Y}_{1985} + t_{0.025}\hat{\sigma}\sqrt{\left[\frac{1}{n} + \frac{\left(X_{1985} - \overline{X}\right)^2}{\sum_{i=1}^{n} x_i^2}\right]} \right)$$

$$= \left(4685.72 - 2.160 * \sqrt{1259.675} * \sqrt{0.4351},\ \ 4685.72 + 2.160 * \sqrt{1259.675} * \sqrt{0.4351}\right)$$

$$= \left(4635.1522,\ \ 4736.2878\right)$$

5.3 Estimation of the Expenditure on Durable Goods:

In the following table we have data quarterly data: about expenditure on durable goods (in billions of 1992 dollars) and total personal consumption expenditure (in billions of 1992 dollars) using this data, estimate the following exponential (log-linear) regression model.

$Y_t = \beta_0 X_t^{\beta} e^{u_t}$ Where

where Y =expenditure on durable goods, billions of 1992 dollars.

X =total personal consumption expenditure, billions of 1992 dollars.

And test for its goodness of fit.

Table 5.2: EXPENDITURE ON DURABLE GOODS AND TOTAL PERSONAL CONSUMPTION EXPENDITURE (BILLIONS OF 1992 DOLLARS)

Observation	Y	X	Observation	Y	X
1993-I	504.0	4286.8	1996-I	611.0	4692.1
1993-II	519.3	4322.8	1996-II	629.5	4746.6
1993-III	529.9	4366.6	1996-III	626.5	4768.3
1993-IV	542.1	4398.0	1996-IV	637.5	4802.6
1994-I	550.7	4439.4	1997-I	656.3	4853.4
1994-II	558.8	4472.2	1997-II	653.8	4872.7

1994-III	561.7	4498.2	1997-III	679.6	4947.0
1994-IV	576.6	4534.1	1997-IV	648.8	4981.0
1995-I	575.2	4555.3	1998-I	710.3	5055.1
1995-II	583.5	4593.6	1998-II	729.4	5130.2
1995-III	595.3	4623.4	1998-III	733.7	5181.8
1995-IV	602.4	4650.0			

Source: Economic Report of the President, 1999, Table B-17, p. 347.

Solution:

The exponential regression model

$$Y_t = \beta_0 X_t^{\beta} e^{u_t}, \qquad t = 1, 2, \ldots, n$$

may alternatively be expressed as

$$\log Y_t = \alpha + \beta \log X_t + u_t, \qquad t = 1, 2, \ldots, n$$

where $\alpha = \log \beta_0 \qquad X^* = \log X_t \qquad Y^* = \log Y_t$

The log values of the given data are computed below:

Observation	$Y^* = \log Y_t$	$X^* = \log X_t$	Observation	$Y^* = \log Y_t$	$X^* = \log X_t$
1993-I	6.22258	8.36330	1996-I	6.41510	8.45364
1993-II	6.25248	8.37166	1996-II	6.44493	8.46518
1993-III	6.27269	8.38174	1996-III	6.44015	8.46975
1993-IV	6.29545	8.38891	1996-IV	6.45755	8.47691
1994-I	6.31119	8.39827	1997-I	6.48662	8.48743
1994-II	6.32579	8.40564	1997-II	6.48280	8.49140
1994-III	6.33097	8.41143	1997-III	6.52150	8.50654
1994-IV	6.35715	8.41938	1997-IV	6.47512	8.51339
1995-I	6.35472	8.42405	1998-I	6.56569	8.52815
1995-II	6.36904	8.43242	1998-II	6.59222	8.54290
1995-III	6.38907	8.43889	1998-III	6.59810	8.55291
1995-IV	6.40092	8.44462			

From the given data we have n=23

$\overline{X}^* = 8.4508 \qquad \overline{Y}^* = 6.4070$

$\sum_{t=1}^{n} X_t^{*2} = 1642.6376 \qquad \sum_{t=1}^{n} Y_t^{*2} = 944.4005$

$\sum_{t=1}^{n} \overline{X}^* \overline{Y}^* = 1245.4542$

Using Eqs. (4.5) and (4.6) we can compute

$$\hat{\beta}=\frac{\frac{1}{n}\sum_{i=1}^{n}X_i^*Y_i^*-\overline{X}^*\overline{Y}^*}{\frac{1}{n}\sum_{i=1}^{n}X_i^{*2}-\overline{X}^{*2}}=1.9056$$

$$\hat{\alpha}=\overline{Y}^*-\hat{\beta}\overline{X}^*=\text{-}9.6971 \quad \Rightarrow \hat{\beta}_0=e^{\hat{\alpha}}=e^{-9.6971}=0.00006$$

Computation of r^2**:**

$$r^2=\beta^2\frac{\operatorname{var}\left(X^*\right)}{\operatorname{var}\left(Y^*\right)}=\beta^2\frac{\frac{1}{n}\sum_{i=1}^{n}X_t^{*2}-\overline{X}^{*2}}{\frac{1}{n}\sum_{i=1}^{n}Y_t^{*2}-\overline{Y}^{*2}}=0.985$$

Testing the significance of regression coefficient (β **):**

We set the null hypothesis $H_0:\beta=0$ vs $H_1:\beta\neq 0$

The test statistic for testing the above hypothesis is given by (from equation (3.15) of Lesson 3)

$$t=\frac{\hat{\beta}}{SE\left(\hat{\beta}\right)}=\frac{1.9056}{SE\left(\hat{\beta}\right)}$$

where

$$SE\left(\hat{\beta}\right)=\hat{\sigma}\Big/\sqrt{\sum_{i=1}^{n}x_i^{*2}}=\hat{\sigma}\Big/\sqrt{\sum_{i=1}^{n}X_i^{*2}-n\overline{X}^{*2}}=0.05137$$

Therefore, the t-ratio is $t=\frac{1.9056}{SE\left(\hat{\beta}\right)}=\frac{1.9056}{0.05137}=37.0956$

From student t-table, at $\varepsilon=1\%$ level of significance t-critical value is

$$t_{0.005}\left(21\right)=2.831$$

Since $\left|t\right|=37.0956>t_{0.005}\left(21\right)=2.831,$ from equation Eq. (3.16) of Lesson 3,

we will reject $H_0:\beta=0$ **.**

Thus we conclude the regression coefficient β ***is highly significant, which establishes the linear influence of X on Y.***

<u>Hence, the model is well fitted and the estimated exponential regression model for Expenditure on Durable Goods is</u>

$$\hat{Y}_t=0.00006X_t^{1.9056}$$

5.4 Self Assessment Questions

1. Explain a simple linear model by means of an illustration.
2. Explain the justification for the inclusion of disturbance (error) term in a simple linear model by means of an example.
3. Give some illustrations of simple linear model.
4. Explain the estimation of consumption function.

5. Explain the estimation of demand function.
6. Explain the estimation of supply function.
7. The following data gives age and blood pressure (B.P.) for 12 persons. Obtain the regression of B.P. on age and test for the significance of slope. Further, estimate the blood pressure when the age is 45 years.

Age in years (X)	Blood pressure (Y)	Age in years (X)	Blood pressure (Y)
56	147	55	150
42	125	49	145
72	160	38	115
36	118	42	140
63	149	68	152
47	128	60	155

8. Fit a power curve (log-log model) of the form $Y=aX^b$ from the following data:

X	1	2	3	4	5	6	7	8
Y	1.0	1.2	1.8	2.5	3.6	4.7	6.6	9.1

9. The expected remaining life of an electronic part is believed to be related to the age of the part. The ages of 10 of these parts that were in use on a certain data were recorded in operating hours. When each part burned out, the elapsed time was recorded. The results were as follows:

Age of part (in hrs.)	40	65	90	5	30	10	80	85	70	25
Remaining life (in hrs.)	30	20	10	80	40	65	15	15	20	50

Fit the exponential curve (log-lin model) of the form $Y=ab^X$.

5.5 References

1. Gujarati, D.N. (2005): *Basic Econometrics, 4th Ed., Tata McGraw-Hill.*
2. Johnston, J. (1984): *Econometric Methods, 3rd Ed., McGraw-Hill,* New York.
3. Montgomery, D.C., Peck, E.A. and Geoffrey Vining, G. (2003): *Introduction to Linear Regression Analysis, 3rd Ed., Wiley*
4. Draper, N.R., and H. Smith(1998): *Applied Regression Analysis, 3rd Ed., John Wiley & Sons,* New York.
5. G.S. Maddala (2001): *Introduction to Econometrics, 3rd Ed., John Wiley & Sons, Ltd.*
6. Johnston, J. and DiNardo J (1997): *Econometric Methods*, 4th Ed., McGraw Hill.
7. Hill, Carter, William Griffiths, and George Judge(2001): *Undergraduate Econometrics, John Willey & Sons,* New York.
8. *Koutsoyiannis, A(1973): Theory of Econometrics, Harper & Row,* New York.

Lesson 6

MULTIPLE REGRESSION ANALYSIS: ESTIMATION

6.0 Objective:

In this lesson, the student will be exposed to the multivariate analogue of the concepts used in simple regression analysis discussed in Lesson 2. After studying the lesson the student will have clear idea regarding the general linear regression model and its basic assumptions. Further, the student will learn how to estimate the unknown regression coefficients $\beta_1, \beta_2, ..., \beta_k$ and **σ^2,** variance of the disturbance u, in general linear regression model using the principle of least squares.

Structure of the Lesson:

6.1 Introduction

The two-variable model studied extensively in the previous lessons is often inadequate in practice. In our consumption–income example, for instance, it was assumed implicitly that only income *X* affects consumption *Y*. But, we know that besides income, a number of other variables are also likely to affect consumption expenditure. An obvious example of other variable is wealth of the consumer. As another example, the demand for a commodity is likely to depend not only on its own price but also on the prices of other competing or complementary goods, income of the consumer, social status, etc. Therefore, we need to extend our simple two-variable regression model to cover models involving more than two variables. A regression model that involves more than one exogenous (independent) variable is called a **multiple regression model or general linear model**. Thus, adding more variables leads us to the discussion of multiple regression models, that is, models in which the dependent variable, or regressand, *Y* depends on two or more explanatory variables, or regressors.

Throughout, we are concerned with multiple linear regression models, that is, models linear in the parameters; they may or may not be linear in the variables. The description and assumptions of the general linear model are explained in Section 6.2. The ordinary least squares estimation of the model is described in Section 6.3. Finally, an unbiased estimator of σ^2 based on least squares residuals is derived in Section 6.4.

6.2 General Linear Regression Model and Assumptions

Let us assume that a linear relationship exists between a variable Y (endogenous variable) and (k-1) explanatory variables $X_2, X_3, \ldots, X_k$ and a disturbance term u. If we have a sample of 'n' observations on Y and $X's$ we can write

$$\underset{\sim}{\mathbf{y}} = \beta_1 \underset{\sim}{\mathbf{x_1}} + \beta_2 \underset{\sim}{\mathbf{x_2}} + \ldots + \beta_k \underset{\sim}{\mathbf{x_k}} + \underset{\sim}{\mathbf{u}} \qquad (6.1)$$

where

$$\underset{\sim}{\mathbf{y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix}, \ \underset{\sim}{\mathbf{x}}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \underset{\sim}{\mathbf{x_2}} = \begin{bmatrix} X_{21} \\ X_{22} \\ X_{23} \\ \vdots \\ X_{2n} \end{bmatrix}, \ldots, \underset{\sim}{\mathbf{x_k}} = \begin{bmatrix} X_{k1} \\ X_{k2} \\ X_{k3} \\ \vdots \\ X_{kn} \end{bmatrix} \text{ and } \underset{\sim}{\mathbf{u}} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{bmatrix}$$

Here it may be noted that $\underset{\sim}{\mathbf{x_1}}$ is a column vector of units to allow for an intercept term. The β's are unknown population (model) parameters and are frequently called as regression coefficients. Even if we know their values, the linear combination ($\beta_1 \underset{\sim}{\mathbf{x_1}} + \beta_2 \underset{\sim}{\mathbf{x_2}} + \ldots + \beta_k \underset{\sim}{\mathbf{x_k}}$) would not determine the $\underset{\sim}{\mathbf{y}}$ vector exactly, for economic relations are stochastic, not exact. Thus $\underset{\sim}{\mathbf{u}}$ is a disturbance vector measuring the discrepancies between the linear combination and any actual sample realization of Y values.

Eq. (6.1) may be expressed in matrix form as

$$\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}} \qquad (6.2)$$

where $\mathbf{X} = \begin{bmatrix} \underset{\sim}{\mathbf{x_1}} & \underset{\sim}{\mathbf{x_2}} & \ldots & \underset{\sim}{\mathbf{x_k}} \end{bmatrix}$, $\underset{\sim}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$

The central problem is now to obtain the estimate of the unknown $\underset{\sim}{\boldsymbol{\beta}}$, the vector of regression coefficients. To make any progress with this we need to make some further assumptions about how the observations on Y have been generated.

<u>Assumptions of the linear Model</u>

1. $E(\underset{\sim}{\mathbf{u}}) = \underset{\sim}{0}$ i.e., $E(\underset{\sim}{\mathbf{y}}) = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}}$ (6.3)

   This means that the values of disturbance term will take both positive and negative discrepancies from its expected value and on balance, they will average out at zero i.e, $E(u_i) = 0 \ \ \forall i = 1, 2, \ldots, n$

2. $E(\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}') = \sigma^2 I$ (6.4)

   Since $E(\underset{\sim}{\mathbf{u}}) = \underset{\sim}{0}$, $E(\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}') = \mathrm{var}(\underset{\sim}{\mathbf{u}})$ is the variance - covariance matrix of $\underset{\sim}{\mathbf{u}}$ and this assumption gives

$$\begin{bmatrix} \mathrm{var}(u_1) & \mathrm{cov}(u_1,u_2) & \dots & \mathrm{cov}(u_1,u_n) \\ \mathrm{cov}(u_2,u_1) & \mathrm{var}(u_2) & \dots & \mathrm{cov}(u_2,u_n) \\ . & . & \dots & . \\ \mathrm{cov}(u_n,u_1) & \mathrm{cov}(u_n,u_2) & \dots & \mathrm{var}(u_n) \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ . & . & \dots & . \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

This is a double assumption, namely

i. $\mathrm{var}(u_i) = \sigma^2 \quad \text{for all } i = 1,2,\ldots,n$ i.e. All disturbances have the same variance.

ii. $cov(u_i, u_j) = 0 \quad \text{for all } i \neq j = 1,2,\ldots,n$ i.e. All disturbances are pair wise uncorrelated.

The first property is referred to as homoscedasticity (or homogeneous variances) and its opposite as heteroscedasticity. If the sample observations related to travel expenditures of a cross section of households, the assumption of homoscedasticity would probably not be a reasonable one, since low income families will almost certainly have low average expenditures on travel and also a low variance of actual travel expenditure about the average, while high income families will tend to display both higher mean levels of expenditure and greater variance about the mean. The second part of this assumption - all disturbances being pair wise uncorrelated – is a very strong assumption indeed. Again, in the context of the travel example it means that the size and sign of the disturbance for any one family has no influence on the size and sign of the disturbance for any other family.

3. $\rho(\mathbf{X}) = k$ (6.5)

This assumption states that the explanatory variables do not form a linearly dependent set. For example, if we had just two explanatory variables, X_2 & X_3 and if this assumption was not fulfilled, then there would exist an exact relationship

$$X_3 = c_1 + c_2 X_2$$

which, is substituted in the regression equation

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

gives

$$\begin{aligned} Y &= \beta_1 + \beta_2 X_2 + \beta_3 \left(c_1 + c_2 X_2\right) + u \\ &= \left(\beta_1 + c_1\beta_3\right) + \left(\beta_2 + c_2\beta_3\right) X_2 + u \end{aligned} \tag{6.6}$$

The constants c_1 and c_2 can be determined exactly, and we can estimate the intercept and slope of Eq. (6.6), but it is not possible to obtain the estimates of the three β parameters.

4. $\mathbf{X}$ is a non-stochastic matrix.

It means that if we take another sample of n observations, the $\mathbf{X}$ matrix of explanatory variables remains unchanged, the only source of variation then being in the $\underset{\sim}{\mathbf{u}}$ vector and hence in the $\underset{\sim}{\mathbf{y}}$ vector. However, the social sciences are notoriously difficult for being observational and non-experimental so that in general the X variables are not subject to experimental control by the social scientist. There are three main points to be made about this assumption.

First of all, in spite of the remarks above, there are cases where the X data can be controlled. In a cross-section survey, the sample design may call for the inclusion of certain numbers of families with specific characteristics, and sampling is continued until these specifications are met. Second, even if it is not in fact feasible to control the X data precisely, it is still useful to be able to make statistical inferences which are conditional on the X values actually present in the sample. In this light it is very much an assumption of convenience in that it simplifies dramatically the derivation of several basic statistical results. Third, once these simple results have been derived, it is possible to weaken the assumption to allow the X variables to be stochastic, but distributed independently of the disturbance term, and then see what modifications of the earlier results are required.

6.3 Ordinary Least Squares Estimation of the Regression

Coefficients

The most frequently used estimating technique for the general linear regression model, namely,

$$\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}} \tag{6.7}$$

is the principle of least squares method. If the unknown vector $\underset{\sim}{\boldsymbol{\beta}}$ in the above equation is replaced by an arbitrary estimator $\underset{\sim}{\hat{\boldsymbol{\beta}}}$, then we may define a vector of errors, or residuals

$$\underset{\sim}{\mathbf{e}} = \underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\hat{\boldsymbol{\beta}}} \tag{6.8}$$

The least-squares principle for choosing $\underset{\sim}{\hat{\boldsymbol{\beta}}}$ is to minimize the sum of the squared residuals, namely,

$$\sum_{i=1}^{n} e_i^2 = \underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}} = (\underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\hat{\boldsymbol{\beta}}})'(\underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\hat{\boldsymbol{\beta}}}) \qquad \text{(from Eq. (6.8))}$$

$$= \underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{y}} - \underset{\sim}{\hat{\boldsymbol{\beta}}}'\mathbf{X}'\underset{\sim}{\mathbf{y}} - \underset{\sim}{\mathbf{y}}'\mathbf{X}\underset{\sim}{\hat{\boldsymbol{\beta}}} + \underset{\sim}{\hat{\boldsymbol{\beta}}}'\mathbf{X}'\mathbf{X}\underset{\sim}{\hat{\boldsymbol{\beta}}}$$

$$= \underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{y}} - 2\underset{\sim}{\hat{\boldsymbol{\beta}}}'\mathbf{X}'\underset{\sim}{\mathbf{y}} + \underset{\sim}{\hat{\boldsymbol{\beta}}}'\mathbf{X}'\mathbf{X}\underset{\sim}{\hat{\boldsymbol{\beta}}} \quad \text{(since the transpose a scalar is the same scalar)}$$

In order to minimize $\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}$, we have to differentiate it with respect to $\underset{\sim}{\hat{\boldsymbol{\beta}}}$ which gives

$$\frac{\partial \underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}}{\partial \underset{\sim}{\hat{\boldsymbol{\beta}}}} = -2\mathbf{X}'\underset{\sim}{\mathbf{y}} + 2\mathbf{X}'\mathbf{X}\underset{\sim}{\hat{\boldsymbol{\beta}}} \tag{6.9}$$

Now the Ordinary least squares (OLS) estimator of $\underset{\sim}{\hat{\boldsymbol{\beta}}}$ of the unknown $\underset{\sim}{\boldsymbol{\beta}}$ can be obtained by setting Eq. (6.9) equal to zero vector and solving it for $\underset{\sim}{\hat{\boldsymbol{\beta}}}$. Thus setting Eq. (6.9) to zero vector, we get

$$(\mathbf{X}'\mathbf{X})\underset{\sim}{\hat{\boldsymbol{\beta}}} = \mathbf{X}'\underset{\sim}{\mathbf{y}} \tag{6.10}$$

Which is a non-homogeneous system of *k* linear equations, those are to be solved for *k* unknowns $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ and are often referred to as the OLS normal equations.

The assumption of the GLM, namely, $\rho(\mathbf{X}) = k$ ensures that $\mathbf{X}'\mathbf{X}$ is nonsingular and hence inverse of $\mathbf{X}'\mathbf{X}$ exits. Hence, from Eq. (6.10), the OLS estimator of $\underset{\sim}{\boldsymbol{\beta}}$, is given by

$$\boxed{\underset{\sim}{\hat{\boldsymbol{\beta}}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underset{\sim}{\mathbf{y}}} \tag{6.11}$$

and the vector of errors or residuals given by Eq.(6.8), where $\hat{\underset{\sim}{\boldsymbol{\beta}}}$ is OLS estimator of $\underset{\sim}{\boldsymbol{\beta}}$, is called the vector of OLS residuals. Using Eq. (6.8) in Eq. (6.10) we get

$$(\mathbf{X'X})\hat{\underset{\sim}{\boldsymbol{\beta}}} = \mathbf{X'}\left(\mathbf{X}\hat{\underset{\sim}{\boldsymbol{\beta}}} + \underset{\sim}{\mathbf{e}}\right) = (\mathbf{X'X})\hat{\underset{\sim}{\boldsymbol{\beta}}} + \mathbf{X'}\underset{\sim}{\mathbf{e}}$$

which implies $$\mathbf{X'}\underset{\sim}{\mathbf{e}} = \begin{bmatrix} \underset{\sim}{\mathbf{x}}_1'\underset{\sim}{\mathbf{e}} \\ \underset{\sim}{\mathbf{x}}_1'\underset{\sim}{\mathbf{e}} \\ \underset{\sim}{\mathbf{x}}_1'\underset{\sim}{\mathbf{e}} \\ \vdots \\ \underset{\sim}{\mathbf{x}}_k'\underset{\sim}{\mathbf{e}} \end{bmatrix}_{(kX1)} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{(kX1)} = \underset{\sim}{\mathbf{0}}_{(kX1)} \tag{6.12}$$

This is a fundamental OLS result. The first element in this equation gives (since $\underset{\sim}{\mathbf{X}}_1$ is the vector of units).

(6.13)

$$\underset{\sim}{\mathbf{x}}_1'\underset{\sim}{\mathbf{e}} = 0 \Rightarrow \sum_{i=1}^{n} e_i = 0 \Rightarrow \bar{e} = 0$$

That is, the residuals from the OLS regression always have zero mean

provided that the regression equation contains a constant term. The remaining elements in Eq. (6.12) state that the residual has zero sample correlation with each X variable.

6.4 Estimation of σ²:

As the values of u are not directly observable, it seems plausible to base an estimate of σ^2 on the residual sum of squares (RSS) $\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}$.

We have from Eq. (6.8)

$$\begin{aligned} \underset{\sim}{\mathbf{e}} &= \underset{\sim}{\mathbf{y}} - \mathbf{X}\hat{\underset{\sim}{\boldsymbol{\beta}}} \\ &= \underset{\sim}{\mathbf{y}} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}\underset{\sim}{\mathbf{y}} \qquad \text{where } \mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'} \\ &= \mathbf{M}\underset{\sim}{\mathbf{y}} \end{aligned} \tag{6.14}$$

Here, $\mathbf{M}$ is an important matrix. It can be easily verified that it is symmetric and idempotent (i.e., $\mathbf{M'} = \mathbf{M}$ & $\mathbf{MM'} = \mathbf{M}^2 = \mathbf{M}$). It also follows

$$\mathbf{MX} = \mathbf{X} - \mathbf{X}(\mathbf{X'X})^{-1}(\mathbf{X'X}) = 0 \tag{6.15}$$

From Eqs. (6.7), (6.14) and (6.15)

$$\underset{\sim}{\mathbf{e}} = \mathbf{M}\underset{\sim}{\mathbf{y}} = \mathbf{M}(\mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}}) = \mathbf{M}\underset{\sim}{\mathbf{u}} \qquad (6.16)$$

and $\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}} = \underset{\sim}{\mathbf{u}}'\mathbf{M}'\mathbf{M}\underset{\sim}{\mathbf{u}}$

$$= \underset{\sim}{\mathbf{u}}'\mathbf{M}^2\underset{\sim}{\mathbf{u}} \qquad (\because\ \mathbf{M} \text{ is symmetric})$$

$$= \underset{\sim}{\mathbf{u}}'\mathbf{M}\underset{\sim}{\mathbf{u}} \qquad (\because\ \mathbf{M} \text{ is idemponent}) \qquad (6.17)$$

Taking expectation on both sides, we get

$$E(\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}) = E(\underset{\sim}{\mathbf{u}}'\mathbf{M}\underset{\sim}{\mathbf{u}})$$

$$= E\left(tr\left(\underset{\sim}{\mathbf{u}}'\mathbf{M}\underset{\sim}{\mathbf{u}}\right)\right) \qquad (\because\ \underset{\sim}{\mathbf{u}}'\mathbf{M}\underset{\sim}{\mathbf{u}} \text{ is scalar})$$

$$= E\left(tr\left(\mathbf{M}\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}'\right)\right) \qquad (\because\ \mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA}))$$

$$= \sigma^2 tr(\mathbf{M}) \qquad (\because\ \mathrm{E}(\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}') = \sigma^2\mathbf{I_n})$$

$$= \sigma^2 tr\left[\mathbf{I_n} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]$$

$$= \sigma^2\left[tr(\mathbf{I_n}) - tr\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\right] \qquad (\because\ \mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA}))$$

$$= \sigma^2\left[n - tr\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\right)\right]$$

$$= \sigma^2\left[n - tr(\mathbf{I}_k)\right]$$

$$= \sigma^2(n-k)$$

Thus $E\left(\dfrac{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}}{n-k}\right) = \sigma^2$ and

hence $\dfrac{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}}{n-k}$ is an unbiased estimator of σ^2 which we denote as $\hat{\sigma}^2$ (6.18)

$\hat{\sigma}$ is often referred to as the standard error of the estimate and may be regarded as the standard deviation of the Y values about the regression plane and is given by

$$\hat{\sigma} = \sqrt{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}/(n-k)} \qquad (6.19)$$

6.5 SELF ASSESSMENT QUESTIONS

1. Explain a general linear model (GLM) along with its assumptions. Also give two applications of GLM.
2. Explain the justification for the inclusion of disturbance (error) term in a general linear model.
3. Explain a multiple regression model along with its assumptions.
4. Explain the significance of disturbance (error) term in a general linear model by means of an illustration.
5. Derive the normal equations for a three-variable regression model.
6. Derive the normal equations for multiple regression model.
7. In a general linear (multiple regression) model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \ \boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\mathbf{I}_n)$, derive the OLS estimator of β.

8. In a general linear model $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}, \ \ \boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \mathbf{I}_n)$, derive the OLS estimator of σ^2 and show that it is unbiased.
9. In a general linear model, derive the OLS estimator of the variance of the disturbance (error) term and show that it is unbiased.
10. In a multiple regression model, derive an unbiased estimator of the variance of the disturbance (error) term.
11. Define a multiple linear regression model and mention the standard assumptions on the statistical disturbance term.
12. For a general linear model $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}, \ \ \boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \mathbf{I}_n)$, show that the maximum likelihood estimator and OLS estimator of β are the same.
13. Show that the sum (mean) of the residuals in a GLM is zero.
14. Show that the vector of residuals uncorrelated with the data matrix $\mathbf{X}$.
15. Prove that if a regression is fitted without a constant term, the sum (mean) of the residuals need not be zero.

6.6 REFERENCES

1. Gujarati, D.N. (2005): *Basic Econometrics, 4th Ed., Tata McGraw-Hill.*
2. Johnston, J. (1984): *Econometric Methods, 3rd Ed., McGraw-Hill,* New York.
3. Montgomery, D.C., Peck, E.A. and Geoffrey Vining, G. (2003): *Introduction to Linear Regression Analysis, 3rd Ed., Wiley*
4. Draper, N.R., and H. Smith(1998): *Applied Regression Analysis, 3rd Ed., John Wiley & Sons,* New York.
5. G.S. Maddala (2001): *Introduction to Econometrics, 3rd Ed., John Wiley & Sons, Ltd.*
6. Johnston, J. and DiNardo J (1997): *Econometric Methods*, 4th Ed., McGraw Hill.
7. Hill, Carter, William Griffiths, and George Judge(2001): *Undergraduate Econometrics, John Willey & Sons,* New York.
8. *Koutsoyiannis, A(1973): Theory of Econometrics, Harper & Row,* New York.

Lesson 7

MULTIPLE REGRESSION ANALYSIS: PROPERTIES OF OLS ESTIMATORS

7.0 Objective:

After studying this lesson the student will understand that ordinary least squares estimator $\hat{\underset{\sim}{\boldsymbol{\beta}}}$ of $\underset{\sim}{\boldsymbol{\beta}}$ in general linear model (multiple linear regression model) $\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}}$, is the best linear unbiased estimator of $\underset{\sim}{\boldsymbol{\beta}}$, which is the famous Gauss-Markov theorem. He/she will be knowing the importance of coefficient of the determination of R^2 and adjusted coefficient of determination $\bar{R}^2$

Structure of the Lesson:

7.1 Introduction

This lesson is a continuation of Lesson 6, which is devoted for studying the properties of ordinary least squares estimator of $\underset{\sim}{\boldsymbol{\beta}}$ in the GLM

$$\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}}$$

which is derived in Lesson 6 and is given by

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

In Section 7.2, we will show that each component of $\hat{\underset{\sim}{\boldsymbol{\beta}}}$ is a linear combination of $Y_1, Y_2, \ldots, Y_n$ and also we will show that $\hat{\underset{\sim}{\boldsymbol{\beta}}}$ is linear unbiased estimator of $\underset{\sim}{\boldsymbol{\beta}}$. In Section 7.3, we will establish the famous Gauss-Markov theorem which states that $\hat{\underset{\sim}{\boldsymbol{\beta}}}$ is best linear unbiased estimator of $\underset{\sim}{\boldsymbol{\beta}}$, which means that no other linear unbiased estimator of $\underset{\sim}{\boldsymbol{\beta}}$ has smaller variance than the OLS estimator $\hat{\underset{\sim}{\boldsymbol{\beta}}}$.

In Section 7.4, we will derive the formulae for the coefficient of determination R^2, which measures the proportion of variation in the dependent variable (Y) explained by linear combination of the explanatory variables $(X_2, X_3, \dots, X_k)$ as compared with the total variation in Y.

In Section 7.5, we will discuss the adjusted R^2 or $\bar{R}^2$ which will be useful for knowing the explanatory power of an additional variable.

7.2 OLS Estimators are Linear Unbiased Estimators

We known that the OLS estimator of $\underset{\sim}{\boldsymbol{\beta}}$ in the GLM $\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}}$ is

$$\hat{\underset{\sim}{\boldsymbol{\beta}}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underset{\sim}{\mathbf{y}} \tag{7.1}$$

Since $\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}}$,(7.1) can be written as

$$\hat{\underset{\sim}{\boldsymbol{\beta}}} = \underset{\sim}{\boldsymbol{\beta}} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underset{\sim}{\mathbf{u}} \tag{7.2}$$

Since $\mathbf{X}$ is non-stochastic matrix,

$$E(\hat{\underset{\sim}{\boldsymbol{\beta}}}) = \underset{\sim}{\boldsymbol{\beta}} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\underset{\sim}{\mathbf{u}}) = \underset{\sim}{\boldsymbol{\beta}} \qquad (\because E(\underset{\sim}{\mathbf{u}}) = \underset{\sim}{\mathbf{0}}) \tag{7.3}$$

> Thus the OLS estimator $\hat{\underset{\sim}{\boldsymbol{\beta}}}$ is a linear unbiased estimator of $\underset{\sim}{\boldsymbol{\beta}}$, More specifically, each element of $\hat{\underset{\sim}{\boldsymbol{\beta}}}$ is a linear unbiased estimator of the corresponding element of $\underset{\sim}{\boldsymbol{\beta}}$

The linearity property refers to linearity in $\underset{\sim}{\mathbf{y}}$ (or $\underset{\sim}{\mathbf{u}}$) as is seen in Eqs. (7.1) or (7.2), for each element in $\hat{\underset{\sim}{\boldsymbol{\beta}}}$ is a linear combination of the elements of $\underset{\sim}{\mathbf{y}}$ (or $\underset{\sim}{\mathbf{u}}$), the weights being functions of the $\mathbf{X}$ data matrix which are non-stochastic.

<u>The variance-covariance matrix of $\hat{\underset{\sim}{\boldsymbol{\beta}}}$</u>

From Eqs. (7.2) and (7.3) we have

$$\hat{\underset{\sim}{\boldsymbol{\beta}}} - E(\hat{\underset{\sim}{\boldsymbol{\beta}}}) = \hat{\underset{\sim}{\boldsymbol{\beta}}} - \underset{\sim}{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underset{\sim}{\mathbf{u}}$$

Now by definition the variance-covariance matrix of $\hat{\underset{\sim}{\boldsymbol{\beta}}}$ is

$$\begin{aligned}
\operatorname{var}(\hat{\underset{\sim}{\boldsymbol{\beta}}}) &= E\left[\left(\hat{\underset{\sim}{\boldsymbol{\beta}}} - \underset{\sim}{\boldsymbol{\beta}}\right)\left(\hat{\underset{\sim}{\boldsymbol{\beta}}} - \underset{\sim}{\boldsymbol{\beta}}\right)'\right] \\
&= \mathrm{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad (\because \mathbf{X} \text{ is nonstochastic}) \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{I}_n\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \qquad (\because E[\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}'] = \sigma^2\mathbf{I}_n) \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned} \tag{7.4}$$

The elements on the main diagonal of Eq. (7.4) give the <u>sampling variances</u> of the corresponding elements of $\hat{\underset{\sim}{\boldsymbol{\beta}}}$, and the off-diagonal terms give the sampling co-variances between the two corresponding elements of $\hat{\underset{\sim}{\boldsymbol{\beta}}}$.

7.3 OLS Estimators are BLUEs(Gauss-Markov theorem):

The following proof is somewhat round about, but it has the advantage of establishing a further important result at the same time.
Let $\underset{\sim}{\mathbf{c}}$ denote an arbitrary k-element column vector of known constants and define a scalar quantity μ as

$$\mu = \underset{\sim}{\mathbf{c}}'\underset{\sim}{\boldsymbol{\beta}} \tag{7.5}$$

If we choose $\underset{\sim}{\mathbf{c}} = (0\ 1\ 0\ \dots\ 0)'$, then $\mu = \beta_2$.
Thus we can use Eq. (7.5) to pick out any single element in $\underset{\sim}{\boldsymbol{\beta}}$. Or if we choose

$$\underset{\sim}{\mathbf{c}}' = \begin{bmatrix} 1 & X_{2,n+1} & \cdots & X_{k,n+1} \end{bmatrix}$$

Then $\mu = \beta_1 + \beta_2 X_{2,n+1} + \ldots + \beta_k X_{k,n+1} = E(Y_{n+1}), \quad \text{for } \mathrm{E}\left(u_{n+1}\right) = 0,$
which is expected value of the dependent variable Y in period (n+1), conditional on the X values in that period.

We wish to consider the class of linear unbiased estimators of μ. Thus define a scalar m which will serve as a linear estimator of μ, such that

$$m = \underset{\sim}{\mathbf{a}}'\underset{\sim}{\mathbf{y}} = \underset{\sim}{\mathbf{a}}'\mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{a}}'\underset{\sim}{\mathbf{u}} \qquad (\because \underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}}) \tag{7.6}$$

where $\underset{\sim}{\mathbf{a}}$ is some n-element column vector. The definition ensures linearity. To ensure unbiased ness we have

$$E(m) = \underset{\sim}{\mathbf{a}}'\mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{a}}'E(\underset{\sim}{\mathbf{u}}) = \underset{\sim}{\mathbf{a}}'\mathbf{X}\underset{\sim}{\boldsymbol{\beta}} = \underset{\sim}{\mathbf{c}}'\underset{\sim}{\boldsymbol{\beta}} \qquad (\because E(\underset{\sim}{\mathbf{u}})=0) \tag{7.7}$$

only if $\quad \underset{\sim}{\mathbf{a}}'\mathbf{X} = \underset{\sim}{\mathbf{c}}'$ (7.8)

The variance of m is given by

$$\begin{aligned}
\mathrm{var(m)} &= E\left[m - E\left(m\right)\right]^2 \\
&= E\left[\underset{\sim}{\mathbf{a}}'\underset{\sim}{\mathbf{u}}\right]^2 && \left(\textit{from}\ \text{Eqs. (7.6)\&(7.7)}\right) \\
&= E(\underset{\sim}{\mathbf{a}}'\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}'\underset{\sim}{\mathbf{a}}) && \left(\because \underset{\sim}{\mathbf{a}}'\underset{\sim}{\mathbf{u}} \text{ is scalar}\right) \\
&= \underset{\sim}{\mathbf{a}}'\left(\sigma^2\mathbf{I_n}\right)\underset{\sim}{\mathbf{a}} && \left(\because \mathrm{E}\left(\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}'\right) = \sigma^2\mathbf{I_n}\right) \\
&= \sigma^2\underset{\sim}{\mathbf{a}}'\underset{\sim}{\mathbf{a}}
\end{aligned} \tag{7.9}$$

Now we should minimize (7.9) subject to the condition

$$\mathbf{X}'\underset{\sim}{\mathbf{a}} = \underset{\sim}{\mathbf{c}} \ \ i.e\ \ \mathbf{X}'\underset{\sim}{\mathbf{a}} - \underset{\sim}{\mathbf{c}} = \underset{\sim}{\mathbf{0}} \qquad \left(\text{from Eq. (7.8)}\right)$$

This is equivalent to minimize the function

$$\phi = \underset{\sim}{\mathbf{a}}'\underset{\sim}{\mathbf{a}} - 2\lambda'\left(\mathbf{X}'\underset{\sim}{\mathbf{a}} - \underset{\sim}{\mathbf{c}}\right) \tag{7.10}$$

with respect to $\underset{\sim}{\mathbf{a}}$ and $\underset{\sim}{\boldsymbol{\lambda}}$ and thus we obtain

$$\frac{\partial \phi}{\partial \underset{\sim}{\mathbf{a}}} = 0 \Rightarrow 2\underset{\sim}{\mathbf{a}} - 2\mathbf{X}\underset{\sim}{\boldsymbol{\lambda}} = \underset{\sim}{0} \tag{7.11}$$

$$\frac{\partial \phi}{\partial \underset{\sim}{\boldsymbol{\lambda}}} = 0 \Rightarrow 2\left(\mathbf{X}'\underset{\sim}{\mathbf{a}} - \underset{\sim}{\mathbf{c}}\right) = \underset{\sim}{0} \tag{7.12}$$

Pre-multiplying Eq. (7.11) by $\mathbf{X}'$, we get

$$\mathbf{X}'\underset{\sim}{\mathbf{a}} - \mathbf{X}'\mathbf{X}\underset{\sim}{\boldsymbol{\lambda}} = \underset{\sim}{\mathbf{0}} \Rightarrow \underset{\sim}{\mathbf{c}} = \mathbf{X}'\mathbf{X}\underset{\sim}{\boldsymbol{\lambda}} \Rightarrow \underset{\sim}{\boldsymbol{\lambda}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\underset{\sim}{\mathbf{c}} \qquad (\text{using Eq. } (7.12))$$

Substituting back $\underset{\sim}{\boldsymbol{\lambda}}$ in Eq. (7.11) we get $\underset{\sim}{\mathbf{a}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\underset{\sim}{\mathbf{c}}$

Hence, from Eq. (7.6) the linear estimator $m = \underset{\sim}{\mathbf{a}}'\underset{\sim}{\mathbf{y}}$ of $\mu = \underset{\sim}{\mathbf{c}}'\underset{\sim}{\boldsymbol{\beta}}$ has minimum variance for the choice of

$$\underset{\sim}{\mathbf{a}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\underset{\sim}{\mathbf{c}} \tag{7.13}$$

Therefore by definition

$$m = \underset{\sim}{\mathbf{a}}'\underset{\sim}{\mathbf{y}} = \underset{\sim}{\mathbf{c}}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{y}}$$

$$= \underset{\sim}{\mathbf{c}}'\underset{\sim}{\hat{\boldsymbol{\beta}}} \qquad \left(\because \text{ the OLS estimator } \underset{\sim}{\hat{\boldsymbol{\beta}}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{y}}\right)$$

becomes the BLUE of $\mu = \underset{\sim}{\mathbf{c}}'\underset{\sim}{\boldsymbol{\beta}}$.

Thus $\underset{\sim}{\mathbf{c}}'\underset{\sim}{\hat{\boldsymbol{\beta}}}$ is the BLUE of $\underset{\sim}{\mathbf{c}}'\underset{\sim}{\boldsymbol{\beta}}$ and as a consequence $\underset{\sim}{\hat{\boldsymbol{\beta}}}$ is the BLUE of $\underset{\sim}{\boldsymbol{\beta}}$

(7.14)

If follows directly that

1. Each OLS coefficient is the best linear unbiased estimator of the corresponding regression coefficient.
2. The BLUE of any linear combination of the β's is the same linear combination of the $\hat{\beta}$'s.
3. The BLUE of $E(Y_s)$ is $\hat{\beta}_1 + \hat{\beta}_2 X_{2s} + \ldots + \hat{\beta}_k X_{ks}$.

Note:
The above result is often called as Gauss-Markov theorem, which may be stated as follows:

In the general linear model $\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}}$, $E(\underset{\sim}{\mathbf{u}}) = 0$ and $\text{var}(\underset{\sim}{\mathbf{u}}) = \sigma^2\mathbf{I}_n$, the ordinary least squares estimator $\underset{\sim}{\hat{\boldsymbol{\beta}}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{y}}$ is the best linear unbiased estimator (BLUE) of $\underset{\sim}{\boldsymbol{\beta}}$ or more specifically for any arbitrary $\underset{\sim}{\mathbf{c}}$, $\underset{\sim}{\mathbf{c}}'\underset{\sim}{\hat{\boldsymbol{\beta}}}$ is the BLUE of $\underset{\sim}{\mathbf{c}}$, $\underset{\sim}{\mathbf{c}}'\underset{\sim}{\hat{\boldsymbol{\beta}}}$.

7.4 Coefficient of Determination R^2

We have the multiple linear regression model or GLM

$$\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}},\ \mathrm{E}\left(\underset{\sim}{\mathbf{u}}\right) = \underset{\sim}{0} \text{ and } \mathrm{var}\left(\underset{\sim}{\mathbf{u}}\right) = \sigma^2 \mathbf{I}_n$$

and the OLS estimator of $\underset{\sim}{\boldsymbol{\beta}}$ is

$$\hat{\underset{\sim}{\boldsymbol{\beta}}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underset{\sim}{\mathbf{y}} \tag{7.15}$$

Decomposing the $\underset{\sim}{\mathbf{y}}$ vector into the part explained by the regression and the unexplained part,

$$\underset{\sim}{\mathbf{y}} = \hat{\underset{\sim}{\mathbf{y}}} + \underset{\sim}{\mathbf{e}} = \mathbf{X}\hat{\underset{\sim}{\boldsymbol{\beta}}} + \underset{\sim}{\mathbf{e}} \tag{7.16}$$

It follows that

$$\begin{aligned}
\underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{y}} &= \left(\hat{\underset{\sim}{\mathbf{y}}} + \underset{\sim}{\mathbf{e}}\right)'\left(\hat{\underset{\sim}{\mathbf{y}}} + \underset{\sim}{\mathbf{e}}\right) \\
&= \hat{\underset{\sim}{\mathbf{y}}}'\hat{\underset{\sim}{\mathbf{y}}} + \underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}} + 2\hat{\underset{\sim}{\mathbf{y}}}'\underset{\sim}{\mathbf{e}} \\
&= \hat{\underset{\sim}{\boldsymbol{\beta}}}'\mathbf{X}'\mathbf{X}\hat{\underset{\sim}{\boldsymbol{\beta}}} + \underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}} + 2\hat{\underset{\sim}{\boldsymbol{\beta}}}'\mathbf{X}'\underset{\sim}{\mathbf{e}} \qquad (\text{ from Eq.(7.16)})
\end{aligned} \tag{7.17}$$

But we have

$$\begin{aligned}
\mathbf{X}'\underset{\sim}{\mathbf{e}} &= \mathbf{X}'(\underset{\sim}{\mathbf{y}} - \mathbf{X}\hat{\underset{\sim}{\boldsymbol{\beta}}}) \\
&= \mathbf{X}'\underset{\sim}{\mathbf{y}} - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underset{\sim}{\mathbf{y}} \qquad (\text{ from Eq. (7.15) }) \\
&= \mathbf{X}'\underset{\sim}{\mathbf{y}} - \mathbf{X}'\underset{\sim}{\mathbf{y}} = \underset{\sim}{\mathbf{0}}
\end{aligned} \tag{7.18}$$

using Eq. (7.18), Eq. (7.17) becomes

$$\underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{y}} = \hat{\underset{\sim}{\boldsymbol{\beta}}}'\mathbf{X}'\mathbf{X}\hat{\underset{\sim}{\boldsymbol{\beta}}} + \underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}} = \hat{\underset{\sim}{\boldsymbol{\beta}}}'\mathbf{X}'\underset{\sim}{\mathbf{y}} + \underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}} \qquad (\text{ from Eq.(7.15)}) \tag{7.19}$$

However, $\underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{y}} = \sum_{i=1}^{n} Y_i^2$ is the sum of squares of the actual Y values. But normally our interest is analyzing the variation in Y, measured by the sum of the squared deviations from the sample mean, namely,

$$\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2$$

Thus, subtracting $n\overline{Y}^2$ from each side of the decomposition (7.19) gives a revised decomposition,

$$\begin{aligned}
\left(\underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{y}} - n\overline{Y}^2\right) &= \left(\hat{\underset{\sim}{\boldsymbol{\beta}}}'\mathbf{X}'\underset{\sim}{\mathbf{y}} - n\overline{Y}^2\right) + \underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}} \\
TSS \quad &= \quad ESS \quad + RSS
\end{aligned} \tag{7.20}$$

where TSS indicates the total sum of squares in Y; and ESS and RSS are the explained and residual (unexplained) sum of squares respectively..

> The coefficient of determination R^2 is defined as the ratio of ESS to TSS and is
>
> $$R^2 = \frac{ESS}{TSS} = \frac{\hat{\underset{\sim}{\boldsymbol{\beta}}}'\mathbf{X}'\underset{\sim}{\mathbf{y}} - \mathrm{n}\overline{\mathrm{y}}^2}{\underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{y}} - \mathrm{n}\overline{\mathrm{y}}^2}$$

(7.21)

Thus R^2 measures the proportion of the total variation in Y explained by the linear combination of the regressors and obviously lies between 0 and 1 (since $0 \le ESS \le TSS$). Most computer programs routinely produce R^2 , along with the estimated GLM. It may be noted from Eq. (7.21), the R^2 will never decrease with the addition of any variable to the set of regressors (both TSS and $n\bar{y}^2$ in ESS will always be the same for any given set of Y values. The quantity $\hat{\underset{\sim}{\boldsymbol{\beta}}}'\mathbf{X}'\underset{\sim}{\mathbf{y}} = \hat{\underset{\sim}{\boldsymbol{\beta}}}'\mathbf{X}'\mathbf{X}\hat{\underset{\sim}{\boldsymbol{\beta}}}$ is a positive definite quadratic form and hence it will be always positive and will be increased when a new variable is added to the set of existing regressors). If the added variable is totally irrelevant the ESS simply remains constant.

Notes:

1. The positive square root of R^2 is defined as the multiple correlation coefficients.

2. Both R and R^2 lies between 0 and 1. i.e., $0 \le R^2 \le 1$ and $0 \le R \le 1$. If it is 1, the fitted regression line explains. 100 percent of the variation in Y. On the other hand, if it is 0, the model does not explain any of the variation in Y. Typically, however, R^2 lies between these extreme values.

3. The fit of the model is said to be "better'' the closer R^2 is to 1. Recall that in the two-variable case we defined the quantity r as the coefficient of correlation and indicated that it measures the degree of (linear) association between two variables. The three-or-more-variable analogue of r is the coefficient of **multiple correlation,** denoted by R, and it is a measure of the degree of association between Y and all the explanatory variables jointly. Although r can be positive or negative, R is always taken to be positive. In practice, however, R is of little importance. The more meaningful quantity is R^2.

4. The value of R^2 (or R) closer to 1 indicates a higher value of ESS, which results in the case of a strong linear relationship between the dependent variable Y and the set of independent variables $X_2, X_3, \ldots, X_k$. On the other hand, when the linear relationship is very weak, we get a smaller value of ESS when closer to '0'. Thus ^{R2} measures the goodness of fit of the models

7.5 Adjusted R^2 or $\bar{R}^2$ and its use

From Eqs. (7.20) and (7.21), R^2 may also be written as

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}}{\underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{y}} - n\bar{y}^2} \tag{7.22}$$

An important property of R^2 is that it is a non-decreasing function of the number of explanatory variables or regressors present in the model; as the number of regressors increases, R^2 almost invariably increases and never decreases. Stated differently, an additional X variable will not decrease R^2. From Eq.(7.22) we may note that adding any extra explanatory variable can never increase the RSS and thus can never decrease the R^2, since that expression of R^2 takes no account of the number of the number of explanatory variables in the model.

It is sometimes useful to compute an R^2, adjusted for degrees of freedom, especially when comparing the explanatory power of different numbers of explanatory variables. From Eq. (7.22) R^2 may be re-written as

$$R^2 = 1 - \frac{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}} / n}{\left(\underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{y}} - \mathrm{n}\bar{\mathrm{y}}^2\right) / n} \qquad (7.23)$$

The adjusted R^2 is defined as

$$\bar{R}^2 = 1 - \frac{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}} / (n-k)}{\left(\underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{y}} - \mathrm{n}\bar{\mathrm{y}}^2\right) / (n-1)} \qquad (7.24)$$

The rationale behind the adjustment is that k parameters have been used in fitting the regression plane from which the residual sum of squares is measured, and one parameter, the sample mean, has been estimated in computing TSS. These provide unbiased estimators of σ^2 and variance of Y .

From Eqs. (7.22) and (7.24), $\bar{R}^2$ may be re-written as

$$\bar{R}^2 = 1 - \frac{n-1}{n-k} \frac{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}}{\left(\underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{y}} - \mathrm{n}\bar{\mathrm{y}}^2\right)} = 1 - \frac{n-1}{n-k}\left(1 - R^2\right) \qquad (7.25)$$

It is immediately apparent from the above equation that for k > 1,

$\bar{R}^2 < R^2$, for $1 - \bar{R}^2 = \frac{n-1}{n-k}\left(1 - R^2\right) > 1 - R^2 \quad (\because n-1 > n-k \Rightarrow \frac{n-1}{n-k} > 1)$

- which implies that as the number of X variables increases, the adjusted R^2 increases less than the unadjusted R^2.
- Further from Eq.(7.25), it is possible for the adjusted coefficient of determination $\bar{R}^2$ to decline if an additional variable produces too small a reduction in 1- R^2 to compensate for the increase in $\frac{n-1}{n-k}$.
- Therefore, generally, $\bar{R}^2$ is also reported by most statistical packages along with the conventional R^2.

Thus $\bar{R}^2$ will be useful to examine a whether an additional explanatory variable has more specifically, significant influence on the Response variable. When a new explanatory variable is added to the existing explanatory variables and if it produces an increase in $\bar{R}^2$ then we may say that the new variable has some significant influence on Response or dependent variable. Otherwise, the new variable has no significant influence and hence we may discard that new variable from the analysis.

Besides R^2 and adjusted R^2 as goodness of fit measures, other criteria are often used to judge the adequacy of a regression model. Two of these are **Schwarz criterion** and **Akaike's Information criterion,** which are given below and are used to select between competing models.

Schwarz criterion : $SC = \log \frac{e'e}{n} + \frac{k}{n} \log n$

Akaike information criterion: $AIC = \log \frac{e'e}{n} + \frac{2k}{n}$

7.6 Self Assessment Questions

1. Show that OLS estimators are BLUEs in a general linear model.
2. State and prove Gauss-Markov theorem and also discuss the importance of this theorem in linear estimation.
3. Stating clearly the underlying assumptions, prove Gauss-Markov theorem.
4. Stating the assumptions clearly show that the OLS estimator of $\boldsymbol{\beta}$, in the general linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$ is best linear unbiased.
5. Under certain conditions show that the ordinary least squares estimators are best linear unbiased.
6. Prove that the OLS estimator of $\boldsymbol{\beta}$, in the general linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$ is unbiased linear estimator.
7. In a multiple regression model, show that the OLS estimator of the variance of the disturbance (error) term is unbiased.
8. Define the coefficient of determination R^2 and derive a formula for it.
9. Define the coefficient of multiple correlation R and derive a formula for it.
10. Distinguish between R^2 and adjusted R^2 and explain the use of adjusted R^2.
11. Prove that the coefficient of determination R^2 is the square of simple correlation between $\mathbf{y}$ and $\hat{\mathbf{y}}$, where $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.
12. Derive the variance-covariance matrix of the OLS estimator of $\boldsymbol{\beta}$, in the GLM $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$

7.7 References

1. Gujarati, D.N. (2005): *Basic Econometrics, 4th Ed., Tata McGraw-Hill.*
2. Johnston, J. (1984): *Econometric Methods, 3rd Ed., McGraw-Hill,* New York.
3. Montgomery, D.C., Peck, E.A. and Geoffrey Vining, G. (2003): *Introduction to Linear Regression Analysis, 3rd Ed., Wiley*
4. Draper, N.R., and H. Smith(1998): *Applied Regression Analysis, 3rd Ed., John Wiley & Sons,* New York.
5. G.S. Maddala (2001): *Introduction to Econometrics, 3rd Ed., John Wiley & Sons, Ltd.*
6. Johnston, J. and DiNardo J (1997): *Econometric Methods*, 4th Ed., McGraw Hill.
7. Hill, Carter, William Griffiths, and George Judge(2001): *Undergraduate Econometrics, John Willey & Sons,* New York.
8. *Koutsoyiannis, A(1973): Theory of Econometrics, Harper & Row,* New York.

Lesson 8

MULTIPLE REGRESSION ANALYSIS: THE PROBLEM OF INFERENCE AND PREDECTION

8.0 Objective:

Testing the estimated GLM is an important aspect of the multiple regression analysis. In this lesson, the student will be exposed for checking the goodness of fit of the estimated GLM as well as the testing the significance and construction of confidence intervals for individual regression coefficients. Further, from this lesson the student will understand how to predict or forecast the dependent variable using the well fitted model. Using both point and interval prediction methods.

Structure of the Lesson:

8.1 Introduction

This lesson, a continuation of Lesson 3, extends the ideas of hypothesis testing and interval estimation developed there for simple linear model to GLM. Although in many ways the concepts developed in Lesson 3 can be applied straightforwardly to the multiple regression model, a few additional features are unique to such models, and it is these features that will receive more attention in this Lesson.

If our sole objective is point estimation of the parameters of the regression models, as we have seen in Lessons 6 and 7, the method of ordinary least squares (OLS) does not require any assumption about the probability distribution of the disturbances u_i*s*. But if our objective is estimation as well as inference, then we need to assume that the u_i follow some probability distribution in addition to the standard assumptions of the GLM. As in case of simple regression models, for multiple regression models also, we assume that the u_i follow the normal distribution with zero mean and constant variance σ^2. With the normality assumption of the disturbances u_i's, we are able to develop the procedures for hypothesis testing and interval estimation of the β parameters. In Section 8.2, we develop the test procedure for examining the significance of each β parameter along with the construction of its confidence interval and in Section 8.3, we develop the test procedure for the significance of complete regression. Section 8.4 and 8.5 are devoted for developing the test procedure for a set of linear hypothesis $R\underset{\sim}{\boldsymbol{\beta}} = \underset{\sim}{\mathbf{r}}$. In section 8.6, we have discussed the problem of prediction.

8.2 Testing the Significance of the Individual Regression Coefficients:

From Lesson 6, we have the multiple linear regression model (GLM)

$$\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}} \tag{8.1}$$

with the assumption

i) $\mathrm{E}\left(\underset{\sim}{\mathbf{u}}\right) = \underset{\sim}{0}$.

ii) $\mathrm{var}\left(\underset{\sim}{\mathbf{u}}\right) = \mathrm{E}\left(\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}'\right) = \sigma^2 \mathbf{I}_n$.

iii) The data matrix $\mathbf{X}$ is nonstochastic and full rank matrix.

Now, let us make an additional assumption namely

iv) Suppose that the elements of $\underset{\sim}{\mathbf{u}}$ are normal so that $\underset{\sim}{\mathbf{u}}$ is a multivariate normal vector.

The above all four assumptions may now be restated in compact form as

$$\underset{\sim}{\mathbf{u}} \sim N\left(\underset{\sim}{\mathbf{0}},\ \sigma^2 \mathbf{I}_n\right) \tag{8.2}$$

From Eq. (7.1) of Lesson 7, the ordinary least squares (OLS) estimator of $\underset{\sim}{\boldsymbol{\beta}}$ is given by

$$\hat{\underset{\sim}{\boldsymbol{\beta}}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underset{\sim}{\mathbf{y}} \tag{8.3}$$

From Eqs. (8.1) and (8.3), we have

$$\hat{\underset{\sim}{\boldsymbol{\beta}}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{y}} = \underset{\sim}{\boldsymbol{\beta}} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{u}} \tag{8.4}$$

From Eqs. (7.3) and (7.4) of Lesson 7, we have

$$E\left(\hat{\underset{\sim}{\boldsymbol{\beta}}}\right) = \underset{\sim}{\boldsymbol{\beta}} \text{ and } \quad \mathrm{var}\left(\hat{\underset{\sim}{\boldsymbol{\beta}}}\right) = \sigma^2 \mathbf{X}'\mathbf{X} \tag{8.5}$$

From Eq. (8.4) we may notice that each element of $\hat{\underset{\sim}{\boldsymbol{\beta}}}$ is a linear combination of the elements of $\underset{\sim}{\mathbf{u}}$, which is the multivariate normal vector. But, we know that every linear combination of a set of normal variates is also a normal variate and hence, $\hat{\underset{\sim}{\boldsymbol{\beta}}}$ is also a multivariate normal variate and from Eq. (8.5) it immediately follows that

$$\hat{\underset{\sim}{\boldsymbol{\beta}}} \sim N\left(\underset{\sim}{\boldsymbol{\beta}}, \sigma^2 \mathbf{X}'\mathbf{X}\right) \tag{8.6}$$

Result 1: The sampling distribution of the residual sum of squares $\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}$, where $\underset{\sim}{\mathbf{e}} = \underset{\sim}{\mathbf{y}} - \mathbf{X}\hat{\underset{\sim}{\boldsymbol{\beta}}}$ can be shown as

$$\frac{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}}{\sigma^2} \sim \chi^2_{n-k}$$

Proof: We have from Eqs. (6.16) and (6.17) of Lesson 6,

$$\underset{\sim}{\mathbf{e}} = \mathbf{M}\underset{\sim}{\mathbf{u}} \text{ and } \underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}} = \underset{\sim}{\mathbf{u}}'\mathbf{M}\underset{\sim}{\mathbf{u}} \tag{8.7}$$

where $\mathbf{M} = \mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ is an idempotent matrix and it's trace is given by

$$\text{tr}(\mathbf{M})=\text{tr}\left(\mathbf{I}_n-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\quad=\text{tr}(\mathbf{I}_n)-\text{tr}\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)$$
$$=\text{tr}(\mathbf{I}_n)-\text{tr}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\right)\quad=\text{tr}(\mathbf{I}_n)-\text{tr}(\mathbf{I}_k)$$
$$=\text{n-k}$$

But for every idempotent matrix its rank is nothing but its trace and therefore

$$\rho(\mathbf{M})=n-k \tag{8.8}$$

Since $\mathbf{M}$ is an idempotent matrix of rank n-k, there exist an orthogonal matrix (Eigen vector matrix) $\mathbf{P}$ such that

$$\mathbf{P'MP}=\mathbf{E}_{n-k} \tag{8.9}$$

where $\mathbf{E_{n\text{-}k}}$ is the Eigen root diagonal matrix, which is diagonal with *(n-k)* units and *k* zeros on its main diagonal. This is due to the fact any idempotent matrix will have 1's or 0's as eigen roots.

The orthogonal matrix $\mathbf{P}$ may be used to define a transformation from $\underset{\sim}{\mathbf{u}}$ to $\underset{\sim}{\mathbf{v}}$, namely

$\underset{\sim}{\mathbf{u}}=\mathbf{P}\underset{\sim}{\mathbf{v}}$ or $\underset{\sim}{\mathbf{v}}=\mathbf{P}'\underset{\sim}{\mathbf{u}}$ (since $\mathbf{P}$ is orthogonal matrix, $\mathbf{P}^{-1}=\mathbf{P}'$)

Using this transformation in Eq. (8.7) we get

$$\begin{aligned}\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}&=\underset{\sim}{\mathbf{v}}'\mathbf{P}'\mathbf{MP}\underset{\sim}{\mathbf{v}}\\&=\underset{\sim}{\mathbf{v}}'\mathbf{E_{n\text{-}k}}\underset{\sim}{\mathbf{v}} && \left(\text{from Eq.}(8.9)\right)\\&=v_1^2+v_2^2+\ldots+v_{n-k}^2 && \left(\begin{array}{l}\because\ \text{the first } n-k \text{ diagonal elements of } \mathbf{E_{n\text{-}k}}\\ \text{are 1's and remaing } k \text{ diagonal elements of } \mathbf{E_{n\text{-}k}} \text{ are 0's}\end{array}\right)\end{aligned}$$

But the mean vector and variance-covariance matrix of $\underset{\sim}{\mathbf{v}}$ are

$$E\left(\underset{\sim}{\mathbf{v}}\right)=\mathbf{P}'E\left(\underset{\sim}{\mathbf{u}}\right)=\underset{\sim}{\mathbf{0}}$$

and $\quad E\left(\underset{\sim}{\mathbf{v}}\underset{\sim}{\mathbf{v}}'\right)=\mathbf{P}'E\left(\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}'\right)\mathbf{P}=\sigma^2\mathbf{P}'\mathbf{I}_n\mathbf{P}=\sigma^2\mathbf{P}'\mathbf{P}=\sigma^2\mathbf{I}_n \quad \left(\because \mathbf{P}'\mathbf{P}=\mathbf{I}_n\right)$

Therefore, $\quad \underset{\sim}{\mathbf{v}}\sim N\left(\underset{\sim}{\mathbf{0}},\sigma^2\mathbf{I}_n\right)$

Thus the elements of $\underset{\sim}{\mathbf{v}}$ are $v_i \overset{iid}{\sim} N\left(0,\ \sigma^2\right)$ and therefore

$$\frac{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}}{\sigma^2}=\frac{v_1^2+v_2^2+\ldots+v_{n-k}^2}{\sigma^2}\sim\chi^2_{n-k}\quad\left(\because\left(\frac{v_i}{\sigma}\right)^2\ \text{is a chi-square variate}\right) \tag{8.10}$$

Hence the result.

Result 2: The OLS residual vector $\underset{\sim}{\mathbf{e}}=\underset{\sim}{\mathbf{y}}-\mathbf{X}\hat{\underset{\sim}{\boldsymbol{\beta}}}$ is distributed independently of $\hat{\underset{\sim}{\boldsymbol{\beta}}}$. The OLS estimator

Proof: The covariance matrix between $\underset{\sim}{\mathbf{e}}$ and $\hat{\underset{\sim}{\boldsymbol{\beta}}}$ is

$$E\left[\left(\underset{\sim}{\mathbf{e}}-E(\underset{\sim}{\mathbf{e}})\right)\left(\hat{\underset{\sim}{\boldsymbol{\beta}}}-\underset{\sim}{\boldsymbol{\beta}}\right)'\right] = E\left[\underset{\sim}{\mathbf{e}}\left(\hat{\underset{\sim}{\boldsymbol{\beta}}}-\underset{\sim}{\boldsymbol{\beta}}\right)'\right] \qquad \left(\because E(\underset{\sim}{\mathbf{e}})=E[\mathbf{M}\underset{\sim}{\mathbf{u}}]=\mathbf{M}E[\underset{\sim}{\mathbf{u}}]=\underset{\sim}{\mathbf{0}}\right)$$

$$= E\left[\mathbf{M}\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right] \qquad (\text{from Eq. (3.4)})$$

$$= \left[\left(\mathbf{I}_{\mathbf{n}}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\sigma^2\mathbf{I}_{\mathbf{n}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right] \quad \left(\because E(\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}')=\sigma^2\mathbf{I}_{\mathbf{n}}\right)$$

$$= \sigma^2\left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right]$$

$$= \mathbf{0}_{nXk}$$

From Eqs. (8.6) and (8.7), we may notice that each $\hat{\underset{\sim}{\boldsymbol{\beta}}}$ and $\underset{\sim}{\mathbf{e}}$ are multivariate normal vectors and further from the above we have just seen they are uncorrelated. Therefore $\hat{\underset{\sim}{\boldsymbol{\beta}}}$ and $\underset{\sim}{\mathbf{e}}$ are independently distributed and hence $\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}$ is distributed independently of $\hat{\underset{\sim}{\boldsymbol{\beta}}}$.

From (8.6) we have $\hat{\beta}_i \sim N\left(\beta_i,\sigma^2 a_{ii}\right)$ where a_{ii} is i^{th} diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ and hence

$$\frac{\hat{\beta}_i-\beta_i}{\sigma\sqrt{a_{ii}}} \sim N(0,1)$$

From Eq. (8.10), we have $\frac{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}}{\sigma^2} \sim \chi^2$ with *(n-k)* d.f. and independently distributed with $\hat{\beta}_i$

Thus by definition of student t-distribution,

$$t = \frac{\hat{\beta}_i-\beta_i}{\sigma\sqrt{a_{ii}}} \Bigg/ \sqrt{\frac{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}}{\sigma^2(n-k)}}$$

$$= \frac{\hat{\beta}_i-\beta_i}{\sqrt{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}/(n-k)}\sqrt{a_{ii}}}$$

$$= \frac{\hat{\beta}_i-\beta_i}{\hat{\sigma}\sqrt{a_{ii}}} \qquad \left(\because\ \hat{\sigma}^2 = \frac{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}}{(n-k)}\right)$$

$$= \frac{\hat{\beta}_i-\beta_i}{SE(\hat{\beta}_i)} \sim t_{n-k}, \qquad \text{where } SE(\hat{\beta}_i)=\hat{\sigma}\sqrt{a_{ii}} \tag{8.11}$$

Under $H_o : \beta_i = 0$

$$\boxed{t = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \sim t_{n-k}, \quad \text{where } SE(\hat{\beta}_i)=\hat{\sigma}\sqrt{a_{ii}}} \tag{8.12}$$

Thus we may use the above *t* as a test statistic for $H_o : \beta_i = 0$.

Now, based on Eq. (8.12), we may form the decision rule

Decision Rule :

$$If \quad |t| = \left|\frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}\right| > t_{\varepsilon/2}(n-k) \quad \text{reject } H_0 : \beta_i = 0$$

(8.13)

$$\text{(or) equvalently} \qquad \text{accept } H_1 : \beta_i \neq 0 \text{ at } \varepsilon \text{ l.o.s.}$$

Here, $t_{\varepsilon/2}(n-2)$ is a two-tailed percentile of t distribution with n-k d.f. at ε l.o.s. and is defined as

$$\Pr\left\{-t_{\varepsilon/2}(n-k) < t < t_{\varepsilon/2}(n-k)\right\} = \Pr\left\{|t| < t_{\varepsilon/2}(n-k)\right\} = 1-\varepsilon$$

For instance, when $\varepsilon = 5\%$, we chose $t_{0.025}(n-k)$ such that

$$\Pr\left\{-t_{0.025}(n-k) < t < t_{0.025}(n-k)\right\} = 0.95$$

The $100(1\text{-}\varepsilon)\%$ confidence interval of β_i :

From Eq.(8.11) we can construct $(1-\varepsilon)\%$ confidence interval for β_i as follows:

We have by definition, the $(1-\varepsilon)\%$ confidence interval for a student '*t*' variate with *n-k* d.f. is

$$\Pr\left\{-t_{\varepsilon/2}(n-k) < t < t_{\varepsilon/2}(n-k)\right\} = 1-\varepsilon$$

$$\Rightarrow \Pr\left\{-t_{\varepsilon/2}(n-k) < \frac{\hat{\beta}_i - \beta_i}{SE(\hat{\beta}_i)} < t_{\varepsilon/2}(n-k)\right\} = 1-\varepsilon \qquad \text{(from Eq. (8.11))}$$

$$\Rightarrow \Pr\left\{-t_{\varepsilon/2}(n-k) SE(\hat{\beta}_i) < \hat{\beta}_i - \beta_i < t_{\varepsilon/2}(n-k) SE(\hat{\beta}_i)\right\} = 1-\varepsilon \qquad \text{(Since } SE(\hat{\beta}_i) > 0\text{)}$$

$$\Rightarrow \Pr\left\{t_{\varepsilon/2}(n-k) SE(\hat{\beta}_i) > \beta_i - \hat{\beta}_i > -t_{\varepsilon/2}(n-k) SE(\hat{\beta}_i)\right\} = 1-\varepsilon \qquad \text{(multiplying with minus)}$$

$$\Rightarrow \Pr\left\{\hat{\beta}_i + t_{\varepsilon/2}(n-k) SE(\hat{\beta}_i) > \beta_i > \hat{\beta}_i - t_{\varepsilon/2}(n-k) SE(\hat{\beta}_i)\right\} = 1-\varepsilon$$

$$\Rightarrow \Pr\left\{\hat{\beta}_i - t_{\varepsilon/2}(n-k) SE(\hat{\beta}_i) < \beta_i < \hat{\beta}_i + t_{\varepsilon/2}(n-k) SE(\hat{\beta}_i)\right\} = 1-\varepsilon$$

which implies

$$\left(\hat{\beta}_i - t_{\varepsilon/2}(n-k) SE(\hat{\beta}_i),\ \ \hat{\beta}_i + t_{\varepsilon/2}(n-k) SE(\hat{\beta}_i)\right)$$
$$\text{is } 100(1\text{-}\varepsilon) \text{ percent confidence interval of } \beta_i$$

(8.14)

8.3 Testing the Significance of the Complete Regression:

Here the null hypothesis is

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_k = 0 \text{ i.e., } \underset{\sim}{\boldsymbol{\beta}} = 0$$

$$\text{where } \underset{\sim}{\boldsymbol{\beta}} = \begin{pmatrix} \beta_2 & \beta_3 & \cdots & \beta_k \end{pmatrix}' \qquad (8.15)$$

We may develop a test procedure as follows:
Let us consider the model in deviation form i.e.,

$$y_i = \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + \left(u_i - \bar{u}\right) \qquad \forall i=1,2,\ldots,n \qquad (8.16)$$

The model in matrix notation as

$$\underset{\sim}{\mathbf{y}}_* = \mathbf{X}_* \underset{\sim}{\boldsymbol{\beta}} + \left(\underset{\sim}{\mathbf{u}} - \underset{\sim}{\bar{\mathbf{u}}}\right) = \mathbf{X}_* \underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}}_*, \text{ where } \underset{\sim}{\mathbf{u}}_* = \underset{\sim}{\mathbf{u}} - \underset{\sim}{\bar{\mathbf{u}}}$$

where

$$\underset{\sim}{\boldsymbol{\beta}} = \begin{pmatrix} \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{pmatrix}, \underset{\sim}{\mathbf{y}} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} Y_1 - \bar{Y} \\ Y_2 - \bar{Y} \\ \vdots \\ Y_n - \bar{Y} \end{pmatrix}$$

$$\mathbf{X}_* = \begin{pmatrix} x_{21} & x_{31} & \cdots & x_{k1} \\ x_{22} & x_{32} & \cdots & x_{k2} \\ \vdots & \vdots & & \vdots \\ x_{2n} & x_{3n} & \cdots & x_{kn} \end{pmatrix} = \begin{pmatrix} X_{21} - \bar{X}_2 & X_{31} - \bar{X}_3 & \cdots & X_{k1} - \bar{X}_k \\ X_{22} - \bar{X}_2 & X_{32} - \bar{X}_3 & \cdots & X_{k2} - \bar{X}_k \\ \vdots & \vdots & & \vdots \\ X_{2n} - \bar{X}_2 & X_{3n} - \bar{X}_3 & \cdots & X_{kn} - \bar{X}_k \end{pmatrix}$$

Here thus $y's$ and $x's$ are in deviation form.
We have

$$\hat{\underset{\sim}{\boldsymbol{\beta}}} = \left(\mathbf{X}_*'\mathbf{X}_*\right)\mathbf{X}_*'\underset{\sim}{\mathbf{y}}_*$$

$$= \underset{\sim}{\boldsymbol{\beta}} + \left(\mathbf{X}_*'\mathbf{X}_*\right)\mathbf{X}_*'\underset{\sim}{\mathbf{u}}_*$$

and $\hat{\underset{\sim}{\boldsymbol{\beta}}} \sim N\left(\underset{\sim}{\boldsymbol{\beta}}, \sigma^2 \left(\mathbf{X}_*'\mathbf{X}_*\right)^{-1}\right)$ (8.17)

Since each component of $\hat{\underset{\sim}{\boldsymbol{\beta}}}$ is a linear combination of $u's$ of $\underset{\sim}{\mathbf{u}}_*$, where $\underset{\sim}{\mathbf{u}}_* \sim N\left(0, \sigma^2 \mathbf{I}_n\right)$.

From Eq. (8.17) we may write $\hat{\underset{\sim}{\boldsymbol{\beta}}} - \underset{\sim}{\boldsymbol{\beta}} \sim N\left(\underset{\sim}{\mathbf{0}}, \sigma^2 \left(\mathbf{X}_*'\mathbf{X}_*\right)^{-1}\right)$

But we have a result that if a normal vector $\underset{\sim}{\mathbf{z}} \sim N(0, \Sigma)$ then $\underset{\sim}{\mathbf{z}}'\Sigma^{-1}\underset{\sim}{\mathbf{z}} \sim \chi^2_p$ where $p = \rho(\Sigma)$

Therefore

$$\frac{1}{\sigma^2}\left(\hat{\underset{\sim}{\boldsymbol{\beta}}} - \underset{\sim}{\boldsymbol{\beta}}\right)' \left(\mathbf{X}_*'\mathbf{X}_*\right)\left(\hat{\underset{\sim}{\boldsymbol{\beta}}} - \underset{\sim}{\boldsymbol{\beta}}\right) \sim \chi^2_{k-1}$$

$$\left(\mathbf{X}_* \text{ is full rank matrix and hence } p = \rho\left(\mathbf{X}_*'\mathbf{X}_*\right) = \text{order of } \mathbf{X}_* = \text{k-1}\right)$$

Already we have a result

$$\frac{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}}{\sigma^2} \sim \chi^2_{n-k}$$

and is independently distributed with $\hat{\underset{\sim}{\boldsymbol{\beta}}}$.

Therefore by definition of *F*-distribution, we have the ratio of two chi-square varieties divided by the respective d.f. is a *F* variate and hence

$$\mathrm{F} = \frac{\left(\hat{\underset{\sim}{\boldsymbol{\beta}}} - \underset{\sim}{\boldsymbol{\beta}}\right)' \left(\mathbf{X}_*'\mathbf{X}_*\right)\left(\hat{\underset{\sim}{\boldsymbol{\beta}}} - \underset{\sim}{\boldsymbol{\beta}}\right) \Big/ (k-1)}{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}} \big/ (n-k)} \sim \mathrm{F}_{\text{k-1,n-k}}$$

under $\mathrm{H}_0 : \underset{\sim}{\boldsymbol{\beta}} = \underset{\sim}{\mathbf{0}}$,

$$\mathrm{F} = \frac{\hat{\underset{\sim}{\boldsymbol{\beta}}}' \left(\mathbf{X}_*'\mathbf{X}_*\right)' \hat{\underset{\sim}{\boldsymbol{\beta}}} \Big/ (k-1)}{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}} \big/ (n-k)} \sim \mathrm{F}_{\text{k-1,n-k}}$$

$$\Rightarrow F = \frac{\hat{\underset{\sim}{\boldsymbol{\beta}}}' \mathbf{X}_*'\underset{\sim}{\mathbf{y}}_* \Big/ (k-1)}{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}} \big/ (n-k)} \sim \mathrm{F}_{\text{k-1,n-k}} \quad \left(\because \hat{\underset{\sim}{\boldsymbol{\beta}}} = \left(\mathbf{X}_*'\mathbf{X}_*\right)^{-1} \mathbf{X}_*'\underset{\sim}{\mathbf{y}}_*\right) \tag{8.18}$$

But by definition we have the coefficient of determination

$$R^2 = \frac{ESS}{TSS} = \frac{\hat{\underset{\sim}{\boldsymbol{\beta}}}' \mathbf{X}_* \underset{\sim}{\mathbf{y}}_*}{\underset{\sim}{\mathbf{y}}_*'\underset{\sim}{\mathbf{y}}_*} \Rightarrow \hat{\underset{\sim}{\boldsymbol{\beta}}}' \mathbf{X}_*'\underset{\sim}{\mathbf{y}}_* = R^2 \underset{\sim}{\mathbf{y}}_*'\underset{\sim}{\mathbf{y}}_* \tag{8.19}$$

We also have

$$1\text{-}R^2 = \frac{RSS}{TSS} = \frac{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}}{\underset{\sim}{\mathbf{y}}_*'\underset{\sim}{\mathbf{y}}_*} \Rightarrow \underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}} = \left(1\text{-}R^2\right) \underset{\sim}{\mathbf{y}}_*'\underset{\sim}{\mathbf{y}}_* \tag{8.20}$$

Substituting Eqs. (8.19) and (8.20) in Eq. (8.18) we get

$$\mathrm{F} = \frac{R^2 \underset{\sim}{\mathbf{y}}_*'\underset{\sim}{\mathbf{y}}_* \Big/ (k-1)}{\left(1-R^2\right) \underset{\sim}{\mathbf{y}}_*'\underset{\sim}{\mathbf{y}}_* \Big/ (n-k)} = \frac{R^2 \big/ (k-1)}{\left(1-R^2\right) \big/ (n-k)} \sim \mathrm{F}_{\text{k-1,n-k}} \tag{8.21}$$

Decision Rule:

Testing the overall significance of a regression in terms of R^2 - Alternative but equivalent test to the test (8.18).

Given the *k*-variable regression model:

$$Y_i = \beta_i + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i$$

To test the hypothesis

$H_0 : \beta_2 = \beta_3 = \cdots = \beta_k = 0$ vs H_1: Not all slope coefficients are simultaneously zero

compute $$\mathrm{F} = \frac{R^2 \big/ (k-1)}{\left(1-R^2\right) \big/ (n-k)}$$

If $F > F_\varepsilon\left(k-1, n-k\right)$, reject H_0; otherwise you may accept H_0 where $F_\varepsilon\left(k-1, n-k\right)$ is the critical F value at the ε level of significance and ($k-1$) numerator df and (n - k) denominator d.f.

8.4 Set of linear Hypothesis

Consider a set of q linear hypotheses about the elements of $\underset{\sim}{\boldsymbol{\beta}}$,

$$\mathbf{R}\underset{\sim}{\boldsymbol{\beta}} = \underset{\sim}{\mathbf{r}} \tag{8.22}$$

where $\mathbf{R}$ is a known matrix of order $q\text{X}k$ and $\underset{\sim}{\mathbf{r}}$ is a known q-element vector. We also assume $\mathbf{R}$ to have rank q that is there is no linear dependency between the hypotheses. It is extremely important to understand the range of various hypotheses represented by Eq. (8.22). We illustrate them with some examples

1. **Testing the Significance of Individual regression coefficient:**
   Suppose we wish to test $\text{H}_0 : \beta_i = 0$ vs $\text{H}_0 : \beta_i \neq 0$. Then we have to choose $\mathbf{R} = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix}$ and $\underset{\sim}{\mathbf{r}} = 0$ in Eq.(8.22). Here $\mathbf{R}$ contains only a single row $(q = 1)$ with a unit in the i^th position and 0's everywhere else, and $\underset{\sim}{\mathbf{r}}$ is the scalar zero.

2. **Testing the equality of two regression coefficients:**
   Suppose we wish to test the hypothesis $H_0 : \beta_2 - \beta_3 = 0$ i.e., $\beta_2 = \beta_3$. Then choose $\mathbf{R} = \begin{bmatrix} 0 & 1 & -1 & \cdots & 0 \end{bmatrix}$ and $\underset{\sim}{\mathbf{r}} = 1$

3. **Testing a linear restriction on the coefficients:**
   We may represent the hypothesis $H_0 : \beta_3 + \beta_4 = 1$ as $\mathbf{R}\underset{\sim}{\boldsymbol{\beta}} = \underset{\sim}{\mathbf{r}}$

   where $\mathbf{R} = \begin{bmatrix} 0 & 1 & -1 & \cdots & 0 \end{bmatrix}$ $\underset{\sim}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_2 \\ \beta_3 \\ \cdots \\ \beta_k \end{bmatrix}$, $\underset{\sim}{\mathbf{r}} = 1$

4. **Testing the significance of overall or complete regression Equation:**

   If we choose $\mathbf{R} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}_{(k-1)Xk}$ and $\underset{\sim}{\mathbf{r}} = \begin{bmatrix} 0 \\ 0 \\ \cdots \\ 0 \end{bmatrix}_{(k-1)Xk}$

   Then Eq. (8.22) is equivalent to the joint hypothesis

$$H_0 : \begin{bmatrix} \beta_2 \\ \beta_3 \\ \cdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \cdots \\ 0 \end{bmatrix} \text{ or } H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0$$

   that is, the set of explanatory variables $X_2, X_3, \cdots, X_k$ has no linear influence in the determination of Y. This is very important hypothesis. The test of this hypothesis is often referred to as a test of the overall relation.

5. $\mathbf{R} = [0 \ \mathbf{I}_s]$ and $\underset{\sim}{\mathbf{r}} = \underset{\sim}{\mathbf{0}}$

Here $\underset{\sim}{\mathbf{0}}$ is a null matrix of order $\mathrm{sX(k\text{-}s)}$ and $\underset{\sim}{\mathbf{r}}$ is an s-element column vector. This set up the hypothesis that the last s elements in $\underset{\sim}{\boldsymbol{\beta}}$ are jointly zero, i.e.,

$$\beta_{k-s+1} = \beta_{k-s+2} = \cdots = \beta_k = 0$$

For example, in an equation explaining the rate of inflation the explanatory variables might be grouped into two subsets, those measuring expectations of inflation and those measuring pressure of demand. The significance of either subset might be tested by using this for formulation with the numbering of the variables so arranged that those in the subset to be tested come at the end.

It is thus clear that a procedure for testing the general hypothesis $\mathbf{R}\underset{\sim}{\boldsymbol{\beta}} = \underset{\sim}{\mathbf{r}}$ will be extremely useful and powerful, since various specifications for $\mathbf{R}$ and $\underset{\sim}{\mathbf{r}}$ will cover a range of different hypothesis.

8.5 Test procedure for $\mathbf{R}\underset{\sim}{\boldsymbol{\beta}} = \underset{\sim}{\mathbf{r}}$

We have $\hat{\underset{\sim}{\boldsymbol{\beta}}} \sim N\left(\underset{\sim}{\boldsymbol{\beta}}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right)$

Since $\hat{\underset{\sim}{\boldsymbol{\beta}}}$ has multivariate normal distribution $\mathbf{R}\hat{\underset{\sim}{\boldsymbol{\beta}}}$ has multivariate normal distribution with

$$E\left(\mathbf{R}\hat{\underset{\sim}{\boldsymbol{\beta}}}\right) = \mathbf{R}\underset{\sim}{\boldsymbol{\beta}} \text{ and } \operatorname{var}\left(\mathbf{R}\hat{\underset{\sim}{\boldsymbol{\beta}}}\right) = \sigma^2 \mathbf{R} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}' \qquad (8.23)$$

i.e.,

$$\mathbf{R}\hat{\underset{\sim}{\boldsymbol{\beta}}} \sim N\left(\mathbf{R}\underset{\sim}{\boldsymbol{\beta}}, \sigma^2 \mathbf{R} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'\right)$$

$$\mathbf{R}\hat{\underset{\sim}{\boldsymbol{\beta}}} - \underset{\sim}{\mathbf{r}} \sim N\left(\underset{\sim}{\mathbf{0}}, \sigma^2 \mathbf{R} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'\right) \quad \left(\because \mathbf{R}\underset{\sim}{\boldsymbol{\beta}} = \underset{\sim}{\mathbf{r}}\right)$$

But we have a result if, $\underset{\sim}{\mathbf{z}} \sim N(\underset{\sim}{\mathbf{0}}, \boldsymbol{\Sigma})$, then $\underset{\sim}{\mathbf{z}}'\Sigma^{-1}\underset{\sim}{\mathbf{z}} \sim \chi^2_p$ where $p \ is \ \rho(\Sigma)$

Therefore

$$\frac{1}{\sigma^2}\left(\mathbf{R}\hat{\underset{\sim}{\boldsymbol{\beta}}} - \underset{\sim}{\mathbf{r}}\right)' \left[\mathbf{R} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'\right]^{-1} \left(\mathbf{R}\hat{\underset{\sim}{\boldsymbol{\beta}}} - \underset{\sim}{\mathbf{r}}\right) \sim \chi^2_q \qquad (8.24)$$

where $q = \rho\left(\mathbf{R} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'\right)$

$= \text{number of } (\text{independent}) \text{ linear hypothesis}$

We have already $\dfrac{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}}{\sigma^2} \sim \chi^2_{n-k}$ (8.25)

and is independently distributed with $\hat{\underset{\sim}{\boldsymbol{\beta}}}$. Therefore from Eqs. (8.24) and (8.25) by the definition of *F* distribution, we have

$$F = \frac{\left(\mathbf{R}\hat{\underset{\sim}{\boldsymbol{\beta}}} - \underset{\sim}{\mathbf{r}}\right)' \left[\mathbf{R} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'\right]^{-1} \left(\mathbf{R}\hat{\underset{\sim}{\boldsymbol{\beta}}} - \underset{\sim}{\mathbf{r}}\right) \Big/ q}{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}} \big/ (n-k)} \sim F_{q,n-k} \qquad (8.26)$$

which can be used as a test statistic for $H_0 : \mathbf{R}\underset{\sim}{\boldsymbol{\beta}} = \underset{\sim}{\mathbf{r}}$. If the value of F exceeds the critical F-value with $q, n-k$ d.f. at given l.o.s, we reject $H_0 : \mathbf{R}\underset{\sim}{\boldsymbol{\beta}} = \underset{\sim}{\mathbf{r}}$, otherwise we accept H_0

8.6.Prediction

Suppose that we have fitted a regression equation, and we now consider some specific vector of regressor values,

$$\underset{\sim}{\mathbf{c}}' = \begin{bmatrix} 1 & X_{2f} & \cdots & X_{kf} \end{bmatrix}$$

The X's may be hypothetical if an investigator is exploring possible effects of different scenarios, or they may be newly observed values. In either case we wish to predict the value of Y *conditional on* $\underset{\sim}{\mathbf{c}}$. Any such prediction is based on the assumption that the fitted model still holds in the prediction period. When a new value Y_f is also observed it is possible to test this stability assumption. An appealing **point prediction** is obtained by inserting the given X values into the regression equation, giving

$$\hat{Y}_f = \hat{\beta}_1 + \hat{\beta}_2 X_{2f} + \cdots + \hat{\beta}_k X_{kf} = \underset{\sim}{\mathbf{c}}'\hat{\beta} \qquad (8.27)$$

In the discussion of the Gauss-Markov theorem it was shown that $\mathbf{c}'\underset{\sim}{\hat{\boldsymbol{\beta}}}$ is the BLUE of $\mathbf{c}'\underset{\sim}{\boldsymbol{\beta}}$. In the present context $\mathbf{c}'\underset{\sim}{\boldsymbol{\beta}} = E\left(Y_f\right)$. Thus $\hat{Y}_f$ is an optimal predictor of $E\left(Y_f\right)$. Moreover, it was shown in Eq. (8.23) that $\mathrm{var}\left(\mathbf{R}\underset{\sim}{\hat{\boldsymbol{\beta}}}\right) = \sigma^2 \mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'$. Replacing $\mathbf{R}$ by $\underset{\sim}{\mathbf{c}}'$ gives

$$\mathrm{var}\left(\underset{\sim}{\mathbf{c}}'\underset{\sim}{\hat{\boldsymbol{\beta}}}\right) = \underset{\sim}{\mathbf{c}}'\,\mathrm{var}\left(\underset{\sim}{\hat{\boldsymbol{\beta}}}\right)\underset{\sim}{\mathbf{c}}$$

If we assume normality for the disturbance term, it follows that

$$\frac{\underset{\sim}{\mathbf{c}}'\underset{\sim}{\hat{\boldsymbol{\beta}}} - \underset{\sim}{\mathbf{c}}'\underset{\sim}{\boldsymbol{\beta}}}{\sqrt{\mathrm{var}\left(\underset{\sim}{\mathbf{c}}'\underset{\sim}{\hat{\boldsymbol{\beta}}}\right)}} \sim N(0,1)$$

when the unknown σ^2 in $\mathrm{var}\left(\underset{\sim}{\hat{\boldsymbol{\beta}}}\right)$ is replaced by $\hat{\sigma}^2$, the usual shift to the $t-$distribution occurs, giving

$$t = \frac{\hat{Y}_f - E\left(Y_f\right)}{\hat{\sigma}\sqrt{\underset{\sim}{\mathbf{c}}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\underset{\sim}{\mathbf{c}}}} \sim t(n-k) \qquad (8.28)$$

from which a 95 percent confidence interval for $E\left(Y_f\right)$ is

$$\hat{Y}_f \pm t_{0.025}(n-k)\,\hat{\sigma}\sqrt{\mathbf{c}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{c}} \qquad (8.29)$$

We have $\hat{Y}_f = \mathbf{c}'\underset{\sim}{\hat{\boldsymbol{\beta}}}$ as before, and now $Y_f = \underset{\sim}{\mathbf{c}}'\underset{\sim}{\boldsymbol{\beta}} + u_f$. The prediction error is thus

$$e_f = Y_f - \hat{Y}_f = u_f - \underset{\sim}{\mathbf{c}}'\left(\underset{\sim}{\hat{\boldsymbol{\beta}}} - \underset{\sim}{\boldsymbol{\beta}}\right)$$

The process of squaring both sides and taking expectations gives the variance of the prediction error as

$$\begin{aligned}\mathrm{var}\left(e_f\right) &= \sigma^2 + \underset{\sim}{\mathbf{c}}'\,\mathrm{var}\left(\underset{\sim}{\hat{\boldsymbol{\beta}}}\right)\underset{\sim}{\mathbf{c}} \\ &= \sigma^2\left(1 + \underset{\sim}{\mathbf{c}}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\underset{\sim}{\mathbf{c}}\right)\end{aligned}$$

From which we derive a *t* statistic

$$t = \frac{\hat{Y}_f - Y_f}{\hat{\sigma}\sqrt{1 + \underset{\sim}{\mathbf{c}}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\underset{\sim}{\mathbf{c}}}} \sim t(n-k) \qquad (8.30)$$

Thus a 95% confidence interval for Y_f is $\left(\hat{Y}_f \mp t_{0.025}(n-k)\hat{\sigma}\sqrt{1+\underset{\sim}{\mathbf{c}}'(\mathbf{X}'\mathbf{X})^{-1}\underset{\sim}{\mathbf{c}}}\right)$ (8.31)

Thus in brief

1. $\hat{Y}_f = \underset{\sim}{\mathbf{c}}'\hat{\underset{\sim}{\boldsymbol{\beta}}}$ is point predictor for $E(Y_f)$
2. $\left(\hat{Y}_f - t_{0.025}\sqrt{\underset{\sim}{\mathbf{c}}'(\mathbf{X}'\mathbf{X})^{-1}\underset{\sim}{\mathbf{c}}}, \hat{Y}_f + t_{0.025}\sqrt{\underset{\sim}{\mathbf{c}}'(\mathbf{X}'\mathbf{X})^{-1}\underset{\sim}{\mathbf{c}}}\right)$ is interval predictor for $E(Y_f)$, where $t_{0.025}$ is a two-tail interval value of t distribution with *n-k* d.f. at 5% l.o.s
3. $\left(\hat{Y}_f - t_{0.025}\sqrt{1+\underset{\sim}{\mathbf{c}}'(\mathbf{X}'\mathbf{X})^{-1}\underset{\sim}{\mathbf{c}}}, \hat{Y}_f + t_{0.025}\sqrt{1+\underset{\sim}{\mathbf{c}}'(\mathbf{X}'\mathbf{X})^{-1}\underset{\sim}{\mathbf{c}}}\right)$ is interval predictor for Y_f, where $t_{0.025}$ is a two-tail interval value of t distribution with *n-k* d.f. at 5% l.o.s

8.7 Self Assessment Questions

1. Derive the test procedure for the significance of a regression coefficient in the multiple regression (general linear) model
2. Derive the test for the significance of a general linear model completely.
3. Derive the test for testing the equality of two regression coefficients.
4. Derive the test for testing the significance of a subset of regression coefficients.
5. Under the assumption of normality of disturbances, the OLS estimator of $\boldsymbol{\beta}$, in GLM $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$ is distributed as a multivariate normal.
6. Define multiple correlation coefficient R and explain the test for significance of R.
7. Define the coefficient of determination R^2 and derive the test for significance of R^2.
8. Derive the test procedure for testing the significance of an individual regression coefficient in a general linear model.
9. Derive the test procedure for the meaningfulness of a general linear model.
10. Derive the test for the significance of a complete general linear model.
11. Construct confidence interval for the parameters of the general linear model.
12. For the general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, explain the test for testing $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ again the alternative $H_0 : \mathbf{R}\boldsymbol{\beta} \neq \mathbf{r}$ when $\mathbf{R}$ is a known matrix of order mXk and of rank m and $\mathbf{r}$ r is a known mX1 vector.
13. For the general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, Derive the test procedure for testing null hypothesis $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{0},$ where $\mathbf{R}$ is a matrix of linear restrictions on the parameters of $\boldsymbol{\beta}$.
14. Discuss the problem of prediction in GLM.
15. Construct the point predictor of the regressand in GLM.
16. Construct the interval predictor of the regressand in GLM.
17. Construct the point predictor of the mean of the regressand in GLM.
18. Construct the interval predictor of the mean of the regressand in GLM.
19. Prove that the estimator of the variance of the disturbance term is distributed as a Chi-square distribution.
20. Prove that the OLS residual vector $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ is distributed independently of $\hat{\boldsymbol{\beta}}$, the OLS estimator of $\boldsymbol{\beta}$.

8.8 References

1. Gujarati, D.N. (2005): *Basic Econometrics, 4th Ed., Tata McGraw-Hill.*
2. Johnston, J. (1984): *Econometric Methods, 3rd Ed., McGraw-Hill,* New York.
3. Montgomery, D.C., Peck, E.A. and Geoffrey Vining, G. (2003): *Introduction to Linear Regression Analysis, 3rd Ed., Wiley*
4. Draper, N.R., and H. Smith(1998): *Applied Regression Analysis, 3rd Ed., John Wiley & Sons,* New York.
5. G.S. Maddala (2001): *Introduction to Econometrics, 3rd Ed., John Wiley & Sons, Ltd.*
6. Johnston, J. and DiNardo J (1997): *Econometric Methods*, 4th Ed., McGraw Hill.
7. Hill, Carter, William Griffiths, and George Judge(2001): *Undergraduate Econometrics, John Willey & Sons,* New York.
8. Koutsoyiannis, A(1973): Theory of Econometrics, Harper & Row, New York.

Lesson 9

MULTIPLE REGRESSION ANALSIS: APPLICATIONS

9.0 Objective:

The objective of this lesson is to demonstrate the application of the multiple regression analysis technique which we have discussed in lessons 6-8 some practical illustrations.

Structure of the Lesson:

9.1 Introduction

In this lesson, we consider some practical illustration, where our multiple regression analysis, discussed in lesions 6-8, can be applicable. In section 9.2, consider child mortality data for demonstration of multiple regression analysis. In section 9.3, we estimate the famous Cobb-Douglas production function using multiple regression analysis technique lastly in section 9.4, we estimate the total cost function using polynomial regression model by applying multiple regression analysis technique.

9.2 Estimation of GLM - An Application to Child Mortality in Relation to per capita GNP and Female Literacy Rate

In the following table, we have given the cross-sectional data for 64 countries on child mortality (CM), Female literacy rate in percent (FLR) and per capita GNP in 1980.

Table 9.1: FERTILITY AND OTHER DATA FOR 64 COUNTRIES

Observation	CM	FLR	PGNP	Observation	CM	FLR	PGNP
1	128	37	1870	33	142	50	8640
2	204	22	130	34	104	62	350

3	202	16	310	35	287	31	230
4	197	65	570	36	41	66	1620
5	96	76	2050	37	312	11	190
6	209	26	200	38	77	88	2090
7	170	45	670	39	142	22	900
8	240	29	300	40	262	22	230
9	241	11	120	41	215	12	140
10	55	55	290	42	246	9	330
11	75	87	1180	43	191	31	1010
12	129	55	900	44	182	19	300
13	24	93	1730	45	37	88	1730
14	165	31	1150	46	103	35	780
15	94	77	1160	47	67	85	1300
16	96	80	1270	48	143	78	930
17	148	30	580	49	83	85	690
18	98	69	660	50	223	33	200
19	161	43	420	51	240	19	450
20	118	47	1080	52	312	21	280
21	269	17	290	53	12	79	4430
22	189	35	270	54	52	83	270
23	126	58	560	55	79	43	1340
24	12	81	4240	56	61	88	670
25	167	29	240	57	168	28	410
26	135	65	430	58	28	95	4370
27	107	87	3020	59	121	41	1310
28	72	63	1420	60	115	62	1470
29	128	49	420	61	186	45	300
30	27	63	19830	62	47	85	3630
31	152	84	420	63	178	45	220
32	224	23	530	64	142	67	560

Note: CM = Child mortality, the number of deaths of children under age 5 in a year per 1000 live births.
FLR = Female literacy rate, percent.
PGNP = per capita GNP in 1980.

Source: Chandan Mukherjee, Howard White, and Marc Whyte, Econometrics and Data Analysis for Developing Countries, Routledge, London, 1998, p. 456.

Using the above child mortality data estimate the regression equation of the child mortality (CM) on the female literacy rate (FLR) and per capita GNP (PGNP) and carry out the multiple regression analysis completely.

Solution:

From the given data we have

Sample size n = 64 and number of variables k = 3

$\sum Y$ = 9056 $\sum X_2$ = 3276 $\sum X_3$ = 89680

$\sum Y^2$ =1645102 $\sum X_2^2$ = 210304 $\sum X_3^2$ = 593717400

$\sum X_2 Y$ = 361686 $\sum X_3 Y$ =7370550 $\sum X_2 X_3$ =5789760

$$\bar{Y} = \frac{9056}{64} = 141.50$$

$$X'X = \begin{bmatrix} n & \sum X_2 & \sum X_3 \\ & \sum X_2^2 & \sum X_2 X_3 \\ & & \sum X_3^2 \end{bmatrix} = \begin{bmatrix} 64 & 3276 & 89680 \\ & 210304 & 5789760 \\ & & 593717400 \end{bmatrix}$$

$$\mathbf{X}'\underset{\sim}{\mathbf{y}} = \begin{bmatrix} \sum Y \\ \sum X_2 Y \\ \sum X_3 Y \end{bmatrix} = \begin{bmatrix} 9056 \\ 361686 \\ 7370550 \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.077114818 & -0.001203744 & 0.000000090 \\ -0.001203744 & 0.000025290 & -0.000000065 \\ 0.000000090 & -0.000000065 & 0.000000002 \end{bmatrix}$$

Estimation of Regression Model:

From Eq. (6.11), the OLS estimator of $\underset{\sim}{\boldsymbol{\beta}}$ is

$$\underset{\sim}{\hat{\boldsymbol{\beta}}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\underset{\sim}{\mathbf{y}}$$

$$= \begin{bmatrix} 0.077114818 & -0.001203744 & 0.000000090 \\ -0.001203744 & 0.000025290 & -0.000000065 \\ 0.000000090 & -0.000000065 & 0.000000002 \end{bmatrix} \begin{bmatrix} 9056 \\ 361686 \\ 7370550 \end{bmatrix} = \begin{bmatrix} 263.642 \\ -2.2316 \\ -0.005647 \end{bmatrix}$$

Thus the estimated regression model is

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$$

$$\text{i.e., } \widehat{\text{CM}} = 263.642 - 2.2316 \text{ FLR} - 0.005647 \text{ PGNP}$$

Estimation of σ^2 :

From Eq. (6.18) an unbiased of σ^2 is

$$\hat{\sigma}^2 = \frac{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}}{n-k} = \frac{RSS}{n-k} = \frac{TSS - ESS}{n-k}$$

$$\begin{aligned}
\text{TSS} &= \text{ Total Sum of Squares} \\
&= \underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{y}} - n\overline{Y}^2 \\
&= \sum Y^2 - n\overline{Y}^2 \\
&= 1645102 - 64*141.50^2 \\
&= 363678
\end{aligned}$$

$$\begin{aligned}
\text{ESS} &= \text{ Explained Sum of Squares} \\
&= \underset{\sim}{\hat{\boldsymbol{\beta}}}'\mathbf{X}'\underset{\sim}{\mathbf{y}} - n\overline{Y}^2 \\
&= \begin{bmatrix} 263.642 & -2.2316 & -0.005647 \end{bmatrix} \begin{bmatrix} 9056 \\ 361686 \\ 7370550 \end{bmatrix} - 64*141.50^2 \\
&= 257362.3731
\end{aligned}$$

$$\begin{aligned}
\therefore \hat{\sigma}^2 &= \frac{TSS - ESS}{n-k} \\
&= \frac{363678 - 257362.3731}{64-3} \\
&= 1742.8791
\end{aligned}$$

Coefficient determination R^2 and adjusted R^2 :

From Eq. (7.21) $R^2 = \dfrac{ESS}{TSS} = \dfrac{257362.3731}{363678} = 0.708$

$$\begin{aligned}
Adjustged\ R^2 = \overline{R}^2 &= 1 - \left[\frac{n-1}{n-k}\right]\left(1 - R^2\right) \\
&= 1 \text{-} \left[\frac{64\text{-}1}{64\text{-}3}\right](1 - 0.708) \\
&= 0.6981
\end{aligned}$$

Testing the Significance of R^2 or Estimated model:

From Eq. (8.21) we have test statistic for

$H_0 : \beta_2 = \beta_3 = 0 \quad \text{(OR)} \quad H_0 : R^2 = 0$ is

$$F = \frac{R^2/(k-1)}{\left(1-R^2\right)/(n-k)} = \frac{0.708/(3-1)}{(1-0.708)/(64-3)} = 73.8325$$

Here $F \sim F_{3-1,64-3} = F_{2,61}$

From F - table, F - critical values are:

$F_{2,61} \simeq F_{2,60} = 3.15$ (at 5% l.o.s)

$F_{2,61} \simeq F_{2,60} = 4.98$ (at 1% l.o.s)

Since, the calculated F -Value (73.8325) is greater than both 5% and 1% critical F -values (3.15, 4.98). We reject H_0.

Hence we may conclude the estimated regression model is well fitted or the coefficient of determination R^2 is highly significant i.e, $R^2 \neq 0$ at both 1% and 5% l.o.s.

Testing the significance Individual Regression Coefficients:

From Eq. (8.11) the $t-$test for $H_0 : \beta_i = 0$ is

$$t = \frac{\hat{\beta}_i}{SE\left(\hat{\beta}_i\right)} \sim t_{n-k}$$

we have for i=1,2,3.

$$t_i = \frac{\hat{\beta}_i}{SE\left(\hat{\beta}_i\right)} = \frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{a_{ii}}}$$, where $a_{ii}\left(i=1,2,3\right)$ i^{th} diagonal elements of $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$

Thus substitute

$$a_{11} = 0.077114818$$
$$a_{22} = 0.000025290$$
$$a_{33} = 0.000000002$$
$$\hat{\sigma} = 41.7478$$

in the above formula we get the t-ratios

$$t_1 = 22.7411$$
$$t_2 = -10.6294$$
$$t_3 = -3.0246$$

From student t table at α =5% l.o.s. $t_{n-k}\left(0.025\right) = t_{64-3}\left(0.025\right) = 2.00$

95% C.I. of β_i is

$$\left(\beta_i - \left|t_{n-k}\left(0.025\right)\right| SE\left(\hat{\beta}_i\right),\ \beta_i + \left|t_{n-k}\left(0.025\right)\right| SE\left(\hat{\beta}_i\right)\right)$$

95% C.I. of β_1 is

$$\left(\beta_1 - \left|t_{61}\left(0.025\right)\right| SE\left(\hat{\beta}_1\right),\ \beta_1 + \left|t_{61}\left(0.025\right)\right| SE\left(\hat{\beta}_1\right)\right)$$
$$\left(263.642 - \left|2.00\right| * 11.5932,\ 263.642 + \left|2.00\right| * 11.5932\right)$$
$$\left(240.4556,\ 286.8284\right)$$

Similarly 95% confidence interval of $\beta_2 : \left(-2.6515,\ -1.8117\right)$

and 95% confidence interval of $\beta_3 : \left(-0.0094, -0.002\right)$

The above results may be presented in the follows table

The estimated Model

Variable	coefficients	standard error	t-statistic	95% C.I.
Constant	263.642000	11.5932	22.7411**	(240.4556,286.8284)
FLR	-2.231600	0.2100	-10.6294**	(-2.6515,-1.8117)
PGNP	-0.005647	0.0019	-3.0246**	(-0.0094,-0.002)

R^2 =0.708 $\bar{R}^2$ =0.6981 critical t-values: $t_{61}\left(0.025\right)$ =2.00 (5% l.o.s)

$F = 73.8325$ =2.96 (1% l.o.s)

ANOVA Table for Multiple Regression Analysis:

Source of Variation	Sum of Squares	Degrees of freedom	Mean Sum of Squares	F -value
Due to regression	ESS =257362.3731	k-1 = 2	128681.2	73.8326**
Due to residuals	RSS =106315.6269	n-k = 61	1742.9	
Total	TSS=363678.0000	n-1 = 63		

critical F -value: $F_{2,61} \cong 3.15$ (at 5% l.o.s)

$F_{2,61} \cong 4.98$ (at 1% l.o.s)

Where ** indicates the calculated t-values and calculated F-values are highly significant. It means that for the estimated model all regression coefficients are individually highly significant as well as the coefficient of determination R^2 is also highly significant. i.e., The overall fitting of the model is also significant. **Thus, the estimated model is well fitted**.

9.3 Estimation of Log-Linear Model: An Application to the Cobb-Douglas Production Function

We demonstrate transformations in this section by taking up the multivariable extension of the two variable log–linear model discussed in Lesson 4. The specific example we discuss is the celebrated **Cobb–Douglas production function** of production theory.

The Cobb–Douglas production function, in its stochastic form, may be expressed as

$$Y_i = \beta_1 X_{2i}^{\beta_2} X_{3i}^{\beta_3} e^{ui} \tag{9.1}$$

where Y = output, X_2 = labor input, X_3 = capital input

u = stochastic disturbance term, e = base of natural logarithm

From Eq. (9.1) it is clear that the relationship between output and the two inputs is nonlinear. However, if we log-transform this model, we obtain:

$$\begin{aligned} \log Y_i &= \log \beta_1 + \beta_2 \log X_{2i} + \beta_3 \log X_{3i} + u_i \\ &= \beta_0 + \beta_2 \log X_{2i} + \beta_3 \log X_{3i} + u_i, \qquad \text{where } \beta_0 = \log \beta_1 \end{aligned} \tag{9.2}$$

Thus written, the model is linear in the parameters β_0, β_2, and β_3 and is therefore a linear regression model. Notice, though, it is nonlinear in the variables Y and X but linear in the logs of these variables. In short, (9.2) is a *log-log, double-log,* or *log-linear model,* the multiple regression counterpart of the two-variable log-linear model (4.3).

The properties of the Cobb–Douglas production function are quite well known:

1. β_2 is the (partial) elasticity of output with respect to the labor input, that is, it measures the percentage change in output for, say, a 1 percent change in the labor input, holding the capital input constant.
2. Likewise, β_3 is the (partial) elasticity of output with respect to the capital input, holding the labor input constant.

3. The sum $(\beta_2+\beta_3)$ gives information about the *returns to scale,* that is, the response of output to a proportionate change in the inputs. If this sum is 1, then there are *constant returns to scale,* that is, doubling the inputs will double the output, tripling the inputs will triple the output, and so on. If the sum is less than 1, there are *decreasing returns to scale*—doubling the inputs will less than double the output. Finally, if the sum is greater than 1, there are *increasing returns to scale*—doubling the inputs will more than double the output.

We may before proceeding further, note that whenever you have a log–linear regression model involving any number of variables the coefficient of each of the X variables measures the (partial) elasticity of the dependent variable Y with respect to that variable. Thus, if you have a *k*-variable log-linear model:

$$\log Y_i = \log \beta_1 + \beta_2 \log X_{2i} + \beta_3 \log X_{3i} + \ldots + \beta_k \log X_{ki} + u_i \qquad (9.3)$$

each of the (partial) regression coefficients, β_2 through β_k, is the (partial) elasticity of Y with respect to variables X_2 through X_k.

To illustrate the Cobb–Douglas production function, we consider the data shown in Table 9.2; these data are for the agricultural sector of Taiwan for 1958–1972.

TABLE 9.2: REAL GROSS PRODUCT, LABOR DAYS, AND REAL CAPITAL INPUT IN THE AGRICULTURAL SECTOR OF TAIWAN, 1958–1972

Year	Real gross product (millions of NT \$)* Y	Labor days (millions of days) X_2	Real capital input (millions of NT \$) X_3
1958	16607.70	275.5	17803.70
1959	17511.30	274.4	18096.80
1960	20171.20	269.7	18271.80
1961	20932.90	267.0	19167.30
1962	20406.00	267.8	19647.60
1963	20831.60	275.0	20803.50
1964	24806.30	283.0	22076.60
1965	26465.80	300.7	23445.20
1966	27403.00	307.5	24939.00
1967	28628.70	303.7	26713.70
1968	29904.50	304.7	29957.80
1969	27508.20	298.6	31585.90
1970	29035.50	295.5	33474.50
1971	29281.50	299.0	34821.80
1972	31535.80	288.1	41794.30

*Source: Thomas Pei-Fan Chen, "Economic Growth and Structural Change in Taiwan—1952–1972, A Production Function Approach," unpublished Ph.D. thesis, Dept. of Economics, Graduate Center, City University of New York, June 1976, Table II. *New Taiwan dollars.*

Assuming that the model (9.2) satisfies the assumptions of the classical linear regression model, we obtained the following regression by the OLS method.

The estimated Model

Variable	coefficients	standard error	t-statistic
Constant	-3.3380	2.4500	-1.36
log X_2	1.4988	0.5398	2.78*
Log X_3	0.4899	0.1020	4.80**

R^2 =0.889 $\overline{R}^2$ =0.871 critical t-values: $t_{12}\left(0.025\right)$ =2.179 (5% l.o.s)
=3.055 (1% l.o.s)

ANOVA Table for Multiple Regression Analysis:

Source of Variation	Sum of Squares	Degrees of freedom	Mean Sum of Squares	F -value
Due to regression	0.53804	2	0.26902	48.07**
Due to residuals	0.06716	12	0.00560	
Total	0.60520	14		

critical F -value: $F_{2,12} \cong$ 3.88 (at 5% l.o.s)
$F_{2,12} \cong$ 6.93 (at 1% l.o.s)

From the above results, we may notice that the overall log-linear or Cobb-Douglas model is well fitted though the constant term is not statistically significant. Here it may be noted the output elasticity of Labour is **just** significant where as the output elasticity of Capital is **highly** significant.

From the above analysis we see that in the Taiwanese agricultural sector for the period 1958–1972 the output elasticities of labor and capital were 1.4988 and 0.4899, respectively. In other words, over the period of study, holding the capital input constant, a 1 percent increase in the labor input led on the average to about a 1.5 percent increase in the output. Similarly, holding the labor input constant, a 1 percent increase in the capital input led on the average to about a 0.5 percent increase in the output. Adding the two output elasticities, we obtain 1.9887, which gives the value of the returns to scale parameter. As is evident, over the period of the study, the Taiwanese agricultural sector was characterized by increasing returns to scale.

From a purely statistical viewpoint, the estimated regression line fits the data quite well. The R^2 value of 0.8890 means that about 89 percent of the variation in the (log of) output is explained by the (logs of) labor and capital.

Note: Here the details of the computation of the results are not presented, as they are similar of those presented in the previous application presented in section 9.2.

9.4 Estimation of Polynomial Regression Model An Application to the Total Cost Function

We now consider a class of multiple regression models, the **polynomial regression models,** that have found extensive use in econometric research relating to cost and production functions. In introducing these models, we further extend the range of models to which the classical linear regression model can easily be applied.

The parabola is represented by the following equation:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2$$

which is called a *quadratic function*, or more generally, a *second-degree polynomial* in the variable X —the highest power of X represents the degree of the polynomial (if X^3 were added to the preceding function, it would be a third-degree polynomial, and so on).
The stochastic version of the above equation may be written as

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i \tag{9.4}$$

which is called a *second-degree polynomial* regression.

The general k^{th} *degree polynomial regression* may be written as

$$Y = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_k X_i^k + u_i \tag{9.5}$$

Notice that in these types of polynomial regressions there is only one explanatory variable on the right-hand side but it appears with various powers, thus making them multiple regression models. Incidentally, note that if X_i is assumed to be fixed or nonstochastic, the powered terms of X_i also become fixed or nonstochastic.

In short, polynomial regression models can also be estimated by the traditional OLS method.
As an example of the polynomial regression, consider the data on output and total cost of production of a commodity in the short run given in Table 9.3. Using this data we fit the following cubic or third-degree polynomial regression equation:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_i \tag{9.6}$$

where Y = total cost and X = output.

TABLE 9.3: TOTAL COST (Y) AND OUTPUT (X)

Output	Total cost, $
1	193
2	226
3	240
4	244
5	257
6	260
7	274
8	297
9	350
10	420

Source: Basic Econometrics-4th Edition, Author: Damodar N.Gujarati p.227
When the third-degree polynomial regression was fitted to the data of Table 7.4, we obtained the following results:

The estimated regression equation

$$\hat{Y}_i = 141.7667 + 63.4776X_i \text{-} 12.9615X_i^2 + 0.9396X_i^3$$

i.e., Estimated Total cost= 141.7667+63.4776output-12.9615 output^2 +0.9396 output^3

The estimated Model

Variable	**coefficients**	**standard error**	t**-statistic**
Constant	141.7670	6.3750	22.24**
X	63.4780	4.7790	13.28**
X^2	-12.9615	0.9857	-13.15**
X^3	0.9396	0.0591	15.90**

R^2 =0.998 $\overline{R}^2$ =0.998 critical t-values: $t_6(0.025)$=2.447 (5% l.o.s)

=3.707 (1% l.o.s)

ANOVA Table for Multiple Regression Analysis:

Source of Variation	**Sum of Squares**	**Degrees of freedom**	**Mean Sum of Squares**	F **-value**
Due to regression	38918	3	12973	1202.22**
Due to residuals	65	6	11	
Total	38983	9		

critical F -value: $F_{3,6} \cong$ 4.76 (at 5% l.o.s)

$F_{3,6} \cong$ 9.78 (at 1% l.o.s)

From the above results we may observe that the individual regression coefficients as well as the coefficients determination ***R*^{2}** are **highly** significant. Hence, we may conclude that the cubic model is the best fit for the estimation total cost. Since ***R*^{2}**=0.998, almost 99% of variation in the total cost is explained by the number of output units.

9.5 Self Assessment Questions

1. Explain how you estimate the Cobb-Douglas model given by

$$Y_i = \beta_1 X_{2i}^{\beta_2} X_{3i}^{\beta_3} e^{ui}$$

2. Explain the multiple regression analysis by means of an illustration.
3. Distinguish between traditional regression model and Cobb-Douglas model by means of illustrations.

9.6 References

1. Gujarati, D.N. (2005): *Basic Econometrics, 4th Ed., Tata McGraw-Hill.*
2. Johnston, J. (1984): *Econometric Methods, 3rd Ed., McGraw-Hill,* New York.
3. Montgomery, D.C., Peck, E.A. and Geoffrey Vining, G. (2003): *Introduction to Linear Regression Analysis, 3rd Ed., Wiley*
4. Draper, N.R., and H. Smith(1998): *Applied Regression Analysis, 3rd Ed., John Wiley & Sons,* New York.
5. G.S. Maddala (2001): *Introduction to Econometrics, 3rd Ed., John Wiley & Sons, Ltd.*
6. Johnston, J. and DiNardo J (1997): *Econometric Methods*, 4th Ed., McGraw Hill.
7. Hill, Carter, William Griffiths, and George Judge(2001): *Undergraduate Econometrics, John Willey & Sons,* New York.
8. *Koutsoyiannis, A(1973): Theory of Econometrics, Harper & Row,* New York.

Lesson 10

TESTS FOR THE CHOICE BETWEEN LINEAR AND LOG-LINEAR MODELS

10.0 Objective:

After studying this lesson, the student will learn how to choose a true model between the given linear and log-linear models using different tests, since a demonstration problem is also given.

Structure of the Lesson:

10.1 Introduction

Sometimes equations are estimated in log form to take care of the heteroscedasticity problem (which we will discuss in Lesson 16). In many cases the choice of the functional form is dictated by other considerations like convenience in interpretation and some economic reasoning. For instance, if we are estimating a production function, the linear form

$$X = \alpha + \beta_1 L + \beta_2 K \qquad (10.1)$$

where X is the output, L the labor, and K the capital, implies perfect substitutability among the inputs of production. On the other hand, the logarithmic form

$$\log X = \alpha + \beta_1 \log L + \beta_2 \log K \qquad (10.2)$$

implies a Cobb-Douglas production function with unit elasticity of substitution. Both these formulations are special cases of the CES (constant elasticity of substitution) production function.

For the estimation of demand functions the log form is often preferred because it is easy to interpret the coefficients as elasticities. For instance,

$$\log Q = \alpha + \beta_1 \log P + \beta_2 \log Y \qquad (10.3)$$

where Q is the quantity demanded, P the price, and Y the income, implies that β_1 is the price elasticity and β_2 is the income elasticity. A linear demand function implies that these elasticities depend on the particular point along the demand curve that we are at. In this case we have to consider some methods of choosing statistically between the two functional forms.

When comparing the linear with the log-linear forms, we cannot compare the R^2's because R^2 is the ratio of explained variance to the total variance and the variances of y and

$\log y$ are different. Comparing R^2's in this case is like comparing two individuals A and B, where A eats 65% of a carrot cake and B eats 70% of a strawberry cake. The comparison does not make sense because there are two different cakes.

The question of estimation in linear model versus log-linear model has received considerable attention during recent years. Several statistical tests have been suggested for testing the linear versus log-linear. In this lesson we have discussed only two of these tests, which are easy to apply.

10.2 MWD (MacKinnon, White, and Davidson) Test

The choice between a linear regression model (the regressand is a linear function of the regressors) or a log–linear regression model (the log of the regressand is a function of the logs of the regressors) is a perennial question in empirical analysis. We can use a test proposed by MacKinnon, White, and Davidson, which for brevity we call the **MWD test** to choose between the two models.

To illustrate this test, assume the following

H_0 : *Linear Model:* $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i \quad i = 1, 2, \ldots, n$ (10.4)

H_1 : *Log–Linear Model:* $\log Y_i = \alpha_1 + \alpha_2 \log X_{2i} + \alpha_3 \log X_{3i} + \ldots + \alpha_k \log X_{ki} + v_i$ (10.5)

where, as usual, H_0 and H_1 denote the null and alternative hypotheses.

The MWD test involves the following steps:

Step I: Estimate the linear model (10.4) using OLS method and compute the estimate of Y values given by

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \ldots + \hat{\beta}_k X_{ki}, \quad i = 1, 2, \ldots, n \tag{10.6}$$

Step: II: Estimate the log–linear model (10.5) by applying OLS method and obtain the estimates of $\log Y$ values given by

$$\widehat{\log Y_i} = \hat{\alpha}_1 + \hat{\alpha}_2 \log X_{2i} + \hat{\alpha}_3 \log X_{3i} + \ldots + \hat{\alpha}_k \log X_{ki}, \quad i = 1, 2, \ldots, n \tag{10.7}$$

Step III: Using the $\hat{Y}_i$ values (computed at step I) and $\widehat{\log Y_t}$ values (computed at step II), compute an artificial variable namely

$$Z_i = \log \hat{Y}_i - \widehat{\log Y_i}, \quad i = 1, 2, \ldots, n \tag{10.8}$$

Now regress Y on $X_2, X_3, \ldots, X_k$ and Z variables. Reject H_0 if the coefficient of Z is statistically significant by the usual t test, otherwise accept H_0 .

Step IV: Compute another artificial variable (as in Step III) namely

$$W_i = \hat{Y}_i - \exp\left(\widehat{\log Y_i}\right), \quad i = 1, 2, \ldots, n \tag{10.9}$$

Now regress $\log Y$ on $\log X_2, \log X_3, \ldots, \log X_k$ and W variables. Reject H_1 if the coefficient of W is statistically significant by the usual t test, otherwise accept H_1.

The logic of MWD test is quite simple. If the linear model is in fact the correct model, the constructed variable Z should not be statistically significant in Step III, for in that case the log values of the estimated Y values from the linear model and those estimated from the log–linear model should not be different. The same comment applies to the alternative hypothesis H_1.

An application (*The demand for roses*):
Table 10.1 gives quarterly data on these variables:

Y = quantity of roses sold, dozens

X_2 = average wholesale price of roses, \$/dozen

X_3 = average wholesale price of carnations, \$/dozen

1971–III to 1975–II in the Detroit Metropolitan area you are asked to consider the following demand functions:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$$

$$\log Y_t = \alpha_1 + \alpha_2 \log X_{2t} + \alpha_3 \log X_{3t} + v_t$$

Table 10.1: Quarterly data on the Sales of Roses

Year and quarter	Y	X_2	X_3	Year and quarter	Y	X_2	X_3
1971– III	11,484	2.26	3.49	1973 – III	8,038	2.60	3.13
– IV	9,348	2.54	2.85	– IV	7,476	2.89	3.20
1972 – I	8,429	3.07	4.06	1974 –I	5,911	3.77	3.65
– II	10,079	2.91	3.64	– II	7,950	3.64	3.60
– III	9,240	2.73	3.21	– III	6,134	2.82	2.94
– IV	8,862	2.77	3.66	– IV	5,868	2.96	3.12
1973 – I	6,216	3.59	3.76	1975 – I	3,160	4.24	3.58
– II	8,253	3.23	3.49	– II	5,872	3.69	3.53

Source: Basic Econometrics-4th Edition, Author: Damodar N.Gujarati p.236

Solution:
Here

H_0: the true model is linear i.e., $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$

versus

H_1: the true model is log-linear i.e., $\log Y_t = \alpha_1 + \alpha_2 \log X_{2t} + \alpha_3 \log X_{3t} + v_t$

Now, as per the above MWD test the computations are presented below
Step 1:

The estimated linear model:

$$\hat{Y}_t = 9734 - 3782.2 X_{2t} + 2815 X_{3t}$$
$$t = (3.37) \quad (-6.61) \qquad (2.97)$$
$$F = 21.84 \qquad R^2 = 0.77$$

From student- t table, at 5% l.o.s. the two-tailed critical value of t with 13 d.f. = 2.179.
Since calculated t-values of all the regression coefficients are greater than the critical t value, we may conclude that all coefficients are statistically significant. Since F-value is large, the coefficient of determination R^2 is also significant and thus the above linear model is well fitted to the given data.

Using the above estimated linear model we tabulate $\hat{Y}_t$ values.

t	Y_t	X_{2t}	X_{3t}	$\hat{Y}_t$
1	11484	2.26	3.49	11011.6
2	9348	2.54	2.85	8150.8
3	8429	3.07	4.06	9552.8
4	10079	2.91	3.64	8975.5
5	9240	2.73	3.21	8445.7
6	8862	2.77	3.66	9561.3
7	6216	3.59	3.76	6741.4
8	8253	3.23	3.49	7342.9
9	8038	2.60	3.13	8712.2
10	7476	2.89	3.20	7812.4
11	5911	3.77	3.65	5751.0
12	7950	3.64	3.60	6101.9
13	6134	2.82	2.94	7345.2
14	5868	2.96	3.12	7322.4
15	3160	4.24	3.58	3776.2
16	5872	3.69	3.53	5715.7

Step 2:

The estimated log-linear model:

$$\widehat{\log Y_t} = 9.228 - 1.7612 \log X_{2t} + 1.3403 \log X_{3t}$$
$$t = (16.23) \quad (-5.90) \qquad\qquad (2.54)$$
$$F = 17.50 \qquad R^2 = 0.73$$

Since calculated t-values of all regression coefficients are greater than critical t value (2.179), we may conclude that all coefficients are statistically significant at 5% l.o.s. Since F-value is large, R^2 is also significant thus the above log-linear model is well fitted to the given data.

Using the above log-linear model we tabulate below $\widehat{\log Y_t}$ values.

t	$\log Y_t$	$\log X_{2t}$	$\log X_{3t}$	$\widehat{\log Y_t}$
1	9.34871	0.81536	1.24990	9.46701
2	9.14292	0.93216	1.04732	8.98987
3	9.03943	1.12168	1.40118	9.13031
4	9.21821	1.06815	1.29198	9.07824
5	9.13130	1.00430	1.16627	9.02223
6	9.08953	1.01885	1.29746	9.17241
7	8.73488	1.27815	1.32442	8.75190

8	9.01833	1.17248	1.24990	8.83813
9	8.99194	0.95551	1.14103	9.07433
10	8.91945	1.06126	1.16315	8.91775
11	8.68457	1.32708	1.29473	8.62596
12	8.98093	1.29198	1.28093	8.66927
13	8.72160	1.03674	1.07841	8.84738
14	8.67727	1.08519	1.13783	8.84168
15	8.05833	1.44456	1.27536	8.39311
16	8.67795	1.30563	1.26130	8.61893

Step 3:

Using $\log \hat{Y}$, the last column of the table, obtained at Step 1 and $\widehat{\log Y}$, the last column of the table, obtained at Step 2 we compute an artificial variable

$Z_t = \log \hat{Y}_t - \widehat{\log Y_t}$ and tabulated below.

t	Y_t	X_{2t}	X_{3t}	Z_t
1	11484	2.26	3.49	-0.16030
2	9348	2.54	2.85	0.01601
3	8429	3.07	4.06	0.03428
4	10079	2.91	3.64	0.02401
5	9240	2.73	3.21	0.01919
6	8862	2.77	3.66	-0.00690
7	6216	3.59	3.76	0.06413
8	8253	3.23	3.49	0.06336
9	8038	2.60	3.13	-0.00190
10	7476	2.89	3.20	0.04572
11	5911	3.77	3.65	0.03117
12	7950	3.64	3.60	0.04708
13	6134	2.82	2.94	0.05442
14	5868	2.96	3.12	0.05702
15	3160	4.24	3.58	-0.15660
16	5872	3.69	3.53	0.03204

Now using the above table we regress the variable Y on X_2, X_3 and Z and the regression results are as follows

$$\hat{Y}_t = 9728 - 3783.1X_{2t} + 2817.7X_{3t} + 85.0\,Z_t$$
$$t = (3.22)\ (-6.33)\quad (2.84)\quad (0.02)$$
$$F = 13.4 \quad R^2 = 0.77$$

Since the t-value (0.02) is less than the critical t-value (2.179), the coefficient of Z (85) is not significant. **Therefore as per MWD Test, we accept H_0. In other words, we conclude that the true model is linear**

Step 4:

Using $\hat{Y}$, the last column of the table constructed at Step 1 and $\exp\left(\widehat{\log Y}\right)$, the last column of the table constructed at Step 2, we compute another artificial variable, $W_t = \hat{Y}_t - \exp\left(\widehat{\log Y_t}\right)$ and tabulated below.

t	$\log Y_t$	$\log X_{2t}$	$\log X_{3t}$	W_t
1	9.34871	0.81536	1.24990	-1914.60
2	9.14292	0.93216	1.04732	129.43
3	9.03943	1.12168	1.40118	321.87
4	9.21821	1.06815	1.29198	212.95
5	9.1313	1.00430	1.16627	160.49
6	9.08953	1.01885	1.29746	-66.50
7	8.73488	1.27815	1.32442	418.77
8	9.01833	1.17248	1.24990	450.82
9	8.99194	0.95551	1.14103	-16.15
10	8.91945	1.06126	1.16315	349.13
11	8.68457	1.32708	1.29473	176.47
12	8.98093	1.29198	1.28093	280.63
13	8.72160	1.03674	1.07841	389.07
14	8.67727	1.08519	1.13783	405.84
15	8.05833	1.44456	1.27536	-640.29
16	8.67795	1.30563	1.26130	180.23

Now using the above table we regress the variable $\log Y$ on $\log X_2$, $\log X_3$ and W and the regression results are as follows

$$\widehat{\log Y_t} = 9.1489 - 1.9705 \log X_{2t} + 1.5896 \log X_{3t} + 0.0001\ \mathrm{W_t}$$
$$t = (17.08) \quad (-6.42) \qquad (3.07) \qquad (1.66)$$
$$F = 14.17 \quad R^2 = 0.78$$

Since the t-value (1.66) is less than the critical t-value (2.179), the coefficient of W is not significant. **Therefore as per MWD Test, we accept H_1. In other words, we conclude that the true model is log- linear.**

In this particular example by applying MWD Test we get conclusion of both linear model and non-linear model are true models. <u>In between these, we choose linear model, since when compared with calculate t-value of W in log-linear model is larger than the calculate t–value of Z in linear model.</u> <u>Thus ultimately we choose linear model as the true model</u>. But, in general the MWD Test will conclude either linear model or non-linear model as the true model.

10.3 The BM (Bera and McAleer)Test

This is the test suggested by Bera and McAleer which for brevity we call the **BM test** to choose between the two models i) A linear regression model or ii) A log–linear regression model.
The BM test involves the following steps:

The null and alternative hypotheses H_0, and H_1 are same as in the MWD test.
Step I and Step II are same as in MWD test.
Step III: Using the $\hat{Y}_i$ values computed at Step I, run the following artificial regression

$$Z_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + \varepsilon_{1i}, \quad i = 1, 2, \ldots, n \qquad (10.10)$$

$$where \;\; Z_i = \log \hat{Y}_i \,, \quad i = 1, 2, \ldots, n$$

and obtain the residuals $\hat{\varepsilon}_{1i}, \quad i = 1, 2, \ldots, n$, from the regression equation (10.10).
Now, the test for H_0 is based on θ_1 in the artificial regression

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + \theta_1 \hat{\varepsilon}_{1i} + \omega_{1i} \quad i = 1, 2, \ldots, n \qquad (10.11)$$

We use the usual t-test to test this hypotheses. ***If*** $\theta_1 = 0$ ***is accepted, we accept*** H_0 ***that is we choose the linear model***

Step IV: Using the $\widehat{\log Y_i}$ values computed at Step II, run the following artificial regression

$$W_i = \alpha_1 + \alpha_2 \log X_{2i} + \alpha_3 \log X_{3i} + \ldots + \alpha_k \log X_{ki} + \varepsilon_{2i}, \quad i = 1, 2, \ldots, n \qquad (10.12)$$

$$\text{where } \; \mathrm{W}_i = \exp(\widehat{\log Y_i}) \,, \quad i = 1, 2, \ldots, n$$

and obtain the residuals $\hat{\varepsilon}_{2i}, \quad i = 1, 2, \ldots, n$, from the regression equation (10.12).
Now, the test for H_1 is based on θ_2 in the artificial regression

$$\log Y_i = \alpha_1 + \alpha_2 \log X_{2i} + \alpha_3 \log X_{3i} + \ldots + \alpha_k \log X_{ki} + \theta_2 \hat{\varepsilon}_{2i} + \omega_{2i} \quad i = 1, 2, \ldots, n \quad (10.13)$$

If $\theta_2 = 0$ ***is accepted, we accept*** H_1 ***that is we choose the log-linear model.***
Remark : A problem arises if both these hypotheses are rejected or both are accepted.
Notes:

1. The student is advised to apply the above BM test to the application given in MWD test.
2. The above two tests can also be applied for the choice between linear model and semi-log linear model (obtained from linear model by replacing Y variable with $\log Y$ variable).

10.4 Self Assessment Questions

1. Explain the MWD test for choosing between linear and log-linear models for the given data.
2. Explain the BM test for choosing between linear and log-linear models for the given data.
3. Explain the MWD test for choosing between linear and semi-log linear models for the given data.

4. Explain the BM test for choosing between linear and semi-log linear models for the given data.
5. Discuss the merits and demerits of the MWD test.
6. Discuss the merits and demerits of the BM test.
7. Distinguish between MWD test and BM test.

10.5 References

1. Gujarati, D.N. (2005): *Basic Econometrics, 4th Ed., Tata McGraw-Hill.*
2. Johnston, J. (1984): *Econometric Methods, 3rd Ed., McGraw-Hill,* New York.
3. Montgomery, D.C., Peck, E.A. and Geoffrey Vining, G. (2003): *Introduction to Linear Regression Analysis, 3rd Ed., Wiley*
4. Draper, N.R., and H. Smith(1998): *Applied Regression Analysis, 3rd Ed., John Wiley & Sons,* New York.
5. G.S. Maddala (2001): *Introduction to Econometrics, 3rd Ed., John Wiley & Sons, Ltd.*
6. Johnston, J. and DiNardo J (1997): *Econometric Methods*, 4th Ed., McGraw Hill.
7. Hill, Carter, William Griffiths, and George Judge(2001): *Undergraduate Econometrics, John Willey & Sons,* New York.
8. *Koutsoyiannis, A(1973): Theory of Econometrics, Harper & Row,* New York.

Lesson 11

ESTIMATION SUBJECT TO LINEAR RESTRICTIONS

11.0 Objective:

In this lesson, the student will learn how to obtain the OLS estimator of $\underset{\sim}{\boldsymbol{\beta}}$ in the GLM

$$\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}}$$

subject a set of linear restrictions on $\underset{\sim}{\boldsymbol{\beta}}$ namely

$$\mathbf{R}\underset{\sim}{\boldsymbol{\beta}} = \underset{\sim}{\mathbf{r}}$$

which is often called as restricted least squares estimator. Further, he/she will also learn an important application of restricted least squares estimator.

Structure of the Lesson:

11.1 Introduction

Economic theory often suggests that the coefficients of a relation should obey a linear restriction; for example, constant returns to scale imply that the exponents in a Cobb-Douglas production function should sum to unity, and the absence of the money illusion on the part of consumers implies that the sum of the money income and price elasticities in a demand function should be zero. These restrictions may be dealt with in two ways. One is to fit the function free of any restrictions and then test whether the estimated coefficients come sufficiently close to satisfying the restriction. The appropriate theory for the test has already been developed in the previous lessons.

An alternative way of dealing with the problem is to incorporate the restriction in the fitting process so that the estimated coefficients satisfy the restriction exactly. In some cases this is most simply done by working out directly the special form of the estimating equations for the problem in hand.

For instance, consider the Cobb–Douglas production function:

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

where Y = output, X_2 = labor input, and X_3 = capital input. Written in $\log$ form, the equation becomes

$$\log Y_i = \beta_0 + \beta_2 \log X_{2i} + \beta_3 \log X_{3i} + u$$

where $\beta_0 = \log \beta_1$.

Now if there are constant returns to scale (equi-proportional change in output for an equi-proportional change in the inputs), economic theory would suggest that

$$\beta_2 + \beta_3 = 1$$

which is an example of a linear restriction. There are two approaches to deal this problem.

1. The first approach is to fit the function free of any restrictions and then test whether the estimated coefficients come sufficiently close to satisfying the restrictions. Thus estimate the above log-linear model by applying OLS and test for

   $$H_0 : \beta_2 + \beta_3 = 1 \text{ i.e.,} H_0 : \underset{\sim}{\mathbf{c}}'\underset{\sim}{\boldsymbol{\beta}} = 1,$$

   where $c = \begin{pmatrix} 0 & 1 & 1 \end{pmatrix}'$ $\underset{\sim}{\boldsymbol{\beta}} = \begin{pmatrix} \beta_0 & \beta_2 & \beta_3 \end{pmatrix}'$

   by applying the usual t-test namely

   $$t = \frac{\underset{\sim}{\mathbf{c}}'\underset{\sim}{\boldsymbol{\beta}} - 1}{\hat{\sigma}\sqrt{\underset{\sim}{\mathbf{c}}'(\mathbf{X}'\mathbf{X})\underset{\sim}{\mathbf{c}}}} \sim t_{n-k}$$

2. An alternative approach is to incorporate the restrictions in the fitting process so that the estimated coefficients satisfy the restrictions exactly. Thus in the above example, we incorporate the restriction $\beta_2 + \beta_3 = 1$ in the above log-linear model and then estimate the resultant model. This estimator is called the restricted least squares estimator.

11.2 Restricted Least Squares Estimation

In Lesson 8, we have described the test procedure for the hypothesis that the elements of the population vector $\underset{\sim}{\boldsymbol{\beta}}$ obey the set of $q(\leq k)$ linear restrictions in the relation

$$H_0 : \mathbf{R}\underset{\sim}{\boldsymbol{\beta}} = \underset{\sim}{\mathbf{r}}$$

If H_0 is not rejected, one may wish to re-estimate the model, incorporating the restrictions in the estimation process. One important reason for such re-estimation is that it will improve the efficiency of the estimates. This produces an estimator $\underset{\sim}{\mathbf{b}}$ which then satisfies

$$\mathbf{R}\underset{\sim}{\mathbf{b}} = \underset{\sim}{\mathbf{r}} \tag{11.1}$$

First we describe the estimator $\underset{\sim}{\mathbf{b}}$ and next we look at an important application of this new estimator.

The assumed model, as before, is

$$\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}} \text{ with } \mathrm{E}(\underset{\sim}{\mathbf{u}}) = \underset{\sim}{\mathbf{0}} \text{ and } \mathrm{E}(\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}') = \sigma^2 \mathbf{I_n} \tag{11.2}$$

Now we should chose an estimator $\underset{\sim}{\mathbf{b}}$ of $\underset{\sim}{\boldsymbol{\beta}}$, which minimizes

$$\left(\underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\mathbf{b}}\right)'\left(\underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\mathbf{b}}\right)$$

subject to the restrictions $\mathbf{R}\underset{\sim}{\mathbf{b}} = \underset{\sim}{\mathbf{r}}$. For this purpose, we define

$$\varphi = \left(\underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\mathbf{b}}\right)'\left(\underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\mathbf{b}}\right) - 2\underset{\sim}{\lambda}'\left(\mathbf{R}\underset{\sim}{\mathbf{b}} - \underset{\sim}{\mathbf{r}}\right) \tag{11.3}$$

where $\underset{\sim}{\boldsymbol{\lambda}}$ denotes a column vector of q Lagrange multipliers. Taking the partial derivatives of φ gives

$$\frac{\partial\phi}{\partial\underset{\sim}{\mathbf{b}}} = -2\mathbf{X}'\underset{\sim}{\mathbf{y}} + 2\mathbf{X}'\mathbf{X}\underset{\sim}{\mathbf{b}} - 2\mathbf{R}'\underset{\sim}{\lambda}$$

and $$\frac{\partial\phi}{\partial\underset{\sim}{\boldsymbol{\lambda}}} = -2\left(\mathbf{R}\underset{\sim}{\mathbf{b}} - \underset{\sim}{\mathbf{r}}\right)$$

Setting these partial derivatives to zero gives the equations to be solved for $\underset{\sim}{\mathbf{b}}$ and $\underset{\sim}{\boldsymbol{\lambda}}$, namely

$$\mathbf{X}'\mathbf{X}\underset{\sim}{\mathbf{b}} - \mathbf{X}'\underset{\sim}{\mathbf{y}} - \mathbf{R}'\underset{\sim}{\lambda} = 0 \tag{11.4}$$

and $$\mathbf{R}\underset{\sim}{\mathbf{b}} - \underset{\sim}{\mathbf{r}} = \underset{\sim}{\mathbf{0}} \tag{11.5}$$

Pre-multiplying Eq. (11.4) by $\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ gives

$$\mathbf{R}\underset{\sim}{\mathbf{b}} - \mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\underset{\sim}{\lambda} = 0 \quad \left(\because \hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{y}} \text{ is the OLS estimator of } \underset{\sim}{\boldsymbol{\beta}}\right)$$

where $\underset{\sim}{\hat{\boldsymbol{\beta}}}$ is unrestricted least squares estimator and the above may be written as

$$\underset{\sim}{\lambda} = \left[\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right]^{-1}\left(\underset{\sim}{\mathbf{r}} - \mathbf{R}\underset{\sim}{\hat{\boldsymbol{\beta}}}\right) \quad \left(\because \mathbf{R}\underset{\sim}{\mathbf{b}} = \underset{\sim}{\mathbf{r}}\right)$$

Substituting this in Eq. (11.4) and simplifying then we get

$$\underset{\sim}{\mathbf{b}} = \underset{\sim}{\hat{\boldsymbol{\beta}}} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\left[\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right]^{-1}\left(\underset{\sim}{\mathbf{r}} - \mathbf{R}\underset{\sim}{\hat{\boldsymbol{\beta}}}\right) \tag{11.6}$$

Formula (11.6) defines the restricted least-squares estimator satisfying the set of q restrictions embodied in $\mathbf{R}\underset{\sim}{\mathbf{b}} = \underset{\sim}{\mathbf{r}}$.

To prove that $\underset{\sim}{\mathbf{b}}$ is unbiased estimator of $\underset{\sim}{\boldsymbol{\beta}}$:

We have $\hat{\boldsymbol{\beta}} = \underset{\sim}{\boldsymbol{\beta}} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{u}}$ and using this in Eq. (11.6) we get

$$\underset{\sim}{\mathbf{b}} = \underset{\sim}{\boldsymbol{\beta}} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{u}} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\left[\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right]^{-1}\left(\underset{\sim}{\mathbf{r}} - \mathbf{R}\underset{\sim}{\boldsymbol{\beta}} - \mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{u}}\right)$$

Since $\mathbf{R}\underset{\sim}{\boldsymbol{\beta}} = \underset{\sim}{\mathbf{r}}$, we may write $\underset{\sim}{\mathbf{b}}$ as

$$\underset{\sim}{\mathbf{b}} = \underset{\sim}{\boldsymbol{\beta}} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{u}} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\left[\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right]^{-1}\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{u}} \tag{11.7}$$

Taking expectation on both sides we get

$$E\left(\underset{\sim}{\mathbf{b}}\right) = \underset{\sim}{\boldsymbol{\beta}} \qquad \left(\because E\left(\underset{\sim}{\mathbf{u}}\right) = 0\right)$$

<u>To derive var($\underset{\sim}{\mathrm{b}}$):</u>

Eq. (11.7) becomes

$$\underset{\sim}{\mathbf{b}} - \underset{\sim}{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{u}} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\left[\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right]^{-1}\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{u}}$$

$$= \mathbf{A}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{u}} \qquad \text{where } \mathbf{A} = \mathbf{I} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\left[\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right]^{-1}\mathbf{R}$$

By definition

$$\operatorname{var}\left(\underset{\sim}{\mathbf{b}}\right)=E\left[\left(\underset{\sim}{\mathbf{b}}-\underset{\sim}{\boldsymbol{\beta}}\right)\left(\underset{\sim}{\mathbf{b}}-\underset{\sim}{\boldsymbol{\beta}}\right)'\right]$$

$$=\mathbf{A}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'E\left(\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}'\right)\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{A}'$$

$$\operatorname{var}\left(\underset{\sim}{\mathbf{b}}\right)=\sigma^2\mathbf{A}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{A}' \qquad \left(\because E\left(\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}'\right)=\sigma^2\mathbf{I_n}\right)$$

$$=\sigma^2\mathbf{A}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \qquad \text{(Explanation of simplification is given below)}$$

$$\Bigg(\mathrm{A}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathrm{A}'=\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}-\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\left[\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right]^{-1}\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]\left[\mathbf{I}\text{-}\mathbf{R}'\left[\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right]^{-1}\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]$$

$$=\left(\mathbf{X}'\mathbf{X}\right)^{-1}-\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\left[\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right]^{-1}\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

$$-\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\left[\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right]^{-1}\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

$$+\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\left[\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right]^{-1}\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\left[\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right]^{-1}\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

$$=\left(\mathbf{X}'\mathbf{X}\right)^{-1}-\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\left[\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right]^{-1}\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

$$=\mathbf{A}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \qquad \left(\because \text{last two terms will be cancelled}\right)\Bigg)$$

11.3 An alternative Expression of the test statistic for $\mathrm{H}_0:\mathbf{R}\underset{\sim}{\boldsymbol{\beta}}=\underset{\sim}{\mathbf{r}}$

Corresponding to the restricted least squares estimator $\underset{\sim}{\mathbf{b}}$, we may define the residual vector

$$\underset{\sim}{\mathbf{e}}_*=\underset{\sim}{\mathbf{y}}-\mathbf{Xb} \tag{11.8}$$

which may be written as

$$\underset{\sim}{\mathbf{e}}_*=\underset{\sim}{\mathbf{y}}-\mathbf{X}\hat{\underset{\sim}{\boldsymbol{\beta}}}-\mathbf{X}\left(\underset{\sim}{\mathbf{b}}-\hat{\underset{\sim}{\boldsymbol{\beta}}}\right)=\underset{\sim}{\mathbf{e}}-\mathbf{X}\left(\underset{\sim}{\mathbf{b}}-\hat{\underset{\sim}{\boldsymbol{\beta}}}\right)$$

where $\underset{\sim}{\mathbf{e}}=\underset{\sim}{\mathbf{y}}-\mathbf{X}\hat{\underset{\sim}{\boldsymbol{\beta}}}$ is the vector of OLS residuals vector. Now

$$\underset{\sim}{\mathbf{e}}_*'\underset{\sim}{\mathbf{e}}_*=\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}+\left(\underset{\sim}{\mathbf{b}}-\hat{\underset{\sim}{\boldsymbol{\beta}}}\right)'\left(\mathbf{X}'\mathbf{X}\right)\left(\underset{\sim}{\mathbf{b}}-\hat{\underset{\sim}{\boldsymbol{\beta}}}\right)$$

The cross product term vanishing since $\mathbf{X}'\underset{\sim}{\mathbf{e}}=0$. Thus

$$\underset{\sim}{\mathbf{e}}_*'\underset{\sim}{\mathbf{e}}_*-\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}=\left(\underset{\sim}{\mathbf{b}}-\hat{\underset{\sim}{\boldsymbol{\beta}}}\right)'\left(\mathbf{X}'\mathbf{X}\right)\left(\underset{\sim}{\mathbf{b}}-\hat{\underset{\sim}{\boldsymbol{\beta}}}\right) \tag{11.9}$$

Using Eq. (11.6) in Eq. (11.9) we get

$$\underset{\sim}{\mathbf{e}}_*'\underset{\sim}{\mathbf{e}}_* - \underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}} = \left(\underset{\sim}{\mathbf{r}} - \mathbf{R}\hat{\underset{\sim}{\boldsymbol{\beta}}}\right)'\left[\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right]^{-1}\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{X}\right)$$
$$\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\left[\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right]^{-1}\left(\underset{\sim}{\mathbf{r}} - \mathbf{R}\hat{\underset{\sim}{\boldsymbol{\beta}}}\right) \tag{11.10}$$
$$= \left(\underset{\sim}{\mathbf{r}} - \mathbf{R}\hat{\underset{\sim}{\boldsymbol{\beta}}}\right)'\left[\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right]^{-1}\left(\underset{\sim}{\mathbf{r}} - \mathbf{R}\hat{\underset{\sim}{\boldsymbol{\beta}}}\right)$$

But from Eq. (8.26) for testing the null hypothesis $H_0 : \mathbf{R}\underset{\sim}{\boldsymbol{\beta}} = \underset{\sim}{\mathbf{r}}$, we have the test statistic, namely

$$F = \frac{\left(\underset{\sim}{\mathbf{r}} - \mathbf{R}\hat{\underset{\sim}{\boldsymbol{\beta}}}\right)'\left[\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right]^{-1}\left(\underset{\sim}{\mathbf{r}} - \mathbf{R}\hat{\underset{\sim}{\boldsymbol{\beta}}}\right)\Big/q}{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}\big/(n-k)} \sim F_{q,n-k} \tag{11.11}$$

Using Eq. (11.10), then Eq. (11.11) becomes

$$\boxed{F = \frac{\left(RSS_{RR} - RSS_{UR}\right)/q}{RSS_{UR}\big/(n-k)} = \frac{\left(\underset{\sim}{\mathbf{e}}_*'\underset{\sim}{\mathbf{e}}_* - \underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}\right)/q}{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}\big/(n-k)} \sim F_{q,n-k}} \tag{11.12}$$

where RSS_{RR} = Residual sum of squares from Restricted Regression Model = $\underset{\sim}{\mathbf{e}}_*'\underset{\sim}{\mathbf{e}}_*$

RSS_{UR} = Residual sum of squares from Unrestricted Regression Model= $\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}$

Or equivalently from (11.9), it can be seen

$$F = \frac{\left(\underset{\sim}{\mathbf{b}} - \hat{\underset{\sim}{\boldsymbol{\beta}}}\right)'\left(\mathbf{X}'\mathbf{X}\right)\left(\underset{\sim}{\mathbf{b}} - \hat{\underset{\sim}{\boldsymbol{\beta}}}\right)\Big/q}{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}\big/(n-k)} \sim F_{q,n-k} \tag{11.13}$$

Thus we may use any one of the formulae Eqs. (11.11), (11.12) or (11.13) as test statistic for $H_0 : \mathbf{R}\underset{\sim}{\boldsymbol{\beta}} = \underset{\sim}{\mathbf{r}}$. But in between these formulae, formula Eq. (11.12) is simpler, which we will use in the later applications.

Note: In the above we have,

Unrestricted Regression Model: $\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}}$

Restricted Regression Model: $\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}}$ with the set of linear restrictions $\mathbf{R}\underset{\sim}{\boldsymbol{\beta}} = \underset{\sim}{\mathbf{r}}$

11.4 Self Assessment Questions

1. Explain the OLS estimation of the GLM $\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{u}, \quad \underset{\sim}{u} \sim (\underset{\sim}{0}, \sigma_u^2 I_n)$ subject to the linear restrictions $R\underset{\sim}{\beta} = \underset{\sim}{r}$.
2. For the GLM $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, explain the test for testing $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ again the alternative $H_0 : \mathbf{R}\boldsymbol{\beta} \neq \mathbf{r}$ when $\mathbf{R}$ is a known matrix of order mXk and of rank m and $\mathbf{r}$ is a known mX1 vector.
3. Derive the restricted least squares estimator of $\underset{\sim}{\beta}$ in the GLM $\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{u}$, $\underset{\sim}{u} \sim (\underset{\sim}{0}, \sigma_u^2 I_n)$ subject to $R\underset{\sim}{\beta} = \underset{\sim}{r}$.
4. Show that the restricted least squares estimator is unbiased.

5. Derive the variance of the restricted least squares estimator.
6. In the GLM $\mathbf{y} = \mathbf{X\beta} + \mathbf{u}$, derive the test statistic for $\mathrm{H}_0 : \mathbf{R}\underset{\sim}{\boldsymbol{\beta}} = \underset{\sim}{\mathbf{r}}$ in terms of the residual sum of squares of restricted and unrestricted regression models.

11.5 References

1. Gujarati, D.N. (2005): *Basic Econometrics, 4th Ed., Tata McGraw-Hill.*
2. Johnston, J. (1984): *Econometric Methods, 3rd Ed., McGraw-Hill,* New York.
3. Montgomery, D.C., Peck, E.A. and Geoffrey Vining, G. (2003): *Introduction to Linear Regression Analysis, 3rd Ed., Wiley*
4. Draper, N.R., and H. Smith(1998): *Applied Regression Analysis, 3rd Ed., John Wiley & Sons,* New York.
5. G.S. Maddala (2001): *Introduction to Econometrics, 3rd Ed., John Wiley & Sons, Ltd.*
6. Johnston, J. and DiNardo J (1997): *Econometric Methods*, 4th Ed., McGraw Hill.
7. Hill, Carter, William Griffiths, and George Judge(2001): *Undergraduate Econometrics, John Willey & Sons,* New York.
8. *Koutsoyiannis, A(1973): Theory of Econometrics, Harper & Row,* New York.

Lesson 12

TEST OF STRUCTUAL CHANGE IN REGRESSION MODELS

12.0 Objective:

After studying this lesson, the student will understand clearly the concept of structural change or parameter stability and some tests for testing the structural change between two regression equations. Chow test for equality of two regressions is also demonstrated with an example.

Structure of the Lesson:

12.1 Introduction

When we use a regression model involving time series data, it may happen that there is a **structural change** in the relationship between the regressand *Y* and the regressors. By structural change, we mean that the values of the parameters of the model do not remain the same through the entire time period. Sometime the structural change may be due to external forces (e.g., the oil embargoes imposed by the OPEC oil cartel in 1973 and 1979 or the Gulf War of 1990–1991), or due to policy changes (such as the switch from a fixed exchange-rate system to a flexible exchange-rate system around 1973) or action taken by Congress (e.g., the tax changes initiated by President Reagan in his two terms in office or changes in the minimum wage rate) or to a variety of other causes.

When we estimate a multiple regression equation and use it for predictions at future points of time we assume that the parameters are constant over the entire time period of estimation and prediction. To test this hypothesis of parameter constancy (or stability) some tests have been proposed.

12.2 Test for Structural Change between Two Regression Equations:

Suppose we have data on two variables

$Y =$ consumption expenditure

and $X =$ disposable income

The data cover two distinct sub periods, n_1 observations relating to war time years and n_2 observations relating to peace time years. Suppose we wish to investigate whether there is any change, or shift in the consumption function between the wartime and peace time periods. Such a change is referred to as a structural change or structural break. Let us denote the consumption functions by

$$Y = \alpha_1 + \beta_1 X + u_1 \qquad \text{wartime function} \tag{12.1}$$

$$Y = \alpha_2 + \beta_2 X + u_2 \qquad \text{peace time function} \tag{12.2}$$

This is the unrestricted form of the model, allowing intercepts and slopes to be different in the two periods. This model would be set up in the matrix form as follows.

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{n_1} \\ Y_{n_1+1} \\ Y_{n_1+2} \\ \vdots \\ Y_{n_1+n_2} \end{bmatrix} = \begin{bmatrix} 1 & X_1 & 0 & 0 \\ 1 & X_2 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n_1} & 0 & 0 \\ 0 & 0 & 1 & X_{n_1+1} \\ 0 & 0 & 1 & X_{n_1+2} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & X_{n_1+n_2} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n_1} \\ u_{n_1+1} \\ u_{n_1+2} \\ \vdots \\ u_{n_1+n_2} \end{bmatrix} \tag{12.3}$$

where the war time observations have been listed first and the peace time observations last. More compactly, Eq. (12.3) can be written as

$$\underset{\sim}{\mathbf{y}} = \begin{bmatrix} \underset{\sim}{\mathbf{y}}_1 \\ \underset{\sim}{\mathbf{y}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \underset{\sim}{\mathbf{u}}_1 \\ \underset{\sim}{\mathbf{u}}_2 \end{bmatrix} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}} \tag{12.4}$$

where the data matrix $\mathbf{X}$ is block-diagonal. As discussed above this is the unrestricted model. Applying OLS to equation Eq. (12.4) gives

$$\underset{\sim}{\hat{\boldsymbol{\beta}}} = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\beta}_1 \\ \hat{\alpha}_2 \\ \hat{\beta}_2 \end{bmatrix} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{y}}$$
$$= \begin{bmatrix} \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{X}_1'\underset{\sim}{\mathbf{y}}_1 \\ \mathbf{X}_2'\underset{\sim}{\mathbf{y}}_2 \end{bmatrix} = \begin{bmatrix} \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\underset{\sim}{\mathbf{y}}_1 \\ \left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\underset{\sim}{\mathbf{y}}_2 \end{bmatrix} \tag{12.5}$$

These estimates are seen to be identical with those obtained by applying OLS separately to Eqs. (12.1) and (12.2). Using Eq. (12.5), one can obtain the vector $\underset{\sim}{\mathbf{e}}$ of n_1+n_2 residuals, and $\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}$ gives the unrestricted residual sum of squares. Also the unrestricted residual sum of squares, $\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}$ for model Eq. (12.4) may be obtained as the sum of residual sum of squares obtained for models Eqs. (12.1) and (12.2).

Now let us set up the null hypothesis of no structural change between wartime function and peace time function. i.e., $H_0:\alpha_1=\alpha_2, \beta_1=\beta_2$

Now the model becomes

$$Y_i=\alpha+\beta X_i+u_i \quad \text{i=1,2,…,n}$$

where $\alpha=\alpha_1=\alpha_2, \beta=\beta_1=\beta_2$ and $n=n_1+n_2$.

The model is called as **restricted model** and it may also be written as

$$\begin{bmatrix} \underset{\sim}{\mathbf{y}}_1 \\ \underset{\sim}{\mathbf{y}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \underset{\sim}{\mathbf{u}} \tag{12.6}$$

The contrast with the unrestricted model in Eq. (12.4) is that $\mathbf{X}_1$ and $\mathbf{X}_2$ matrices are now stacked vertically, so that only two parameters are required to describe the relation.

Now we can test the given H_0 by using the following test statistic, (From Eq. (11.12) of Lesson 11), in which $q=2$ (number of restrictions) and $k=4$ (number of parameters in unrestricted model)

$$\boxed{F=\frac{\left(\underset{\sim}{\mathbf{e}}_*'\underset{\sim}{\mathbf{e}}_*-\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}\right)/2}{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}/(n-4)}=\frac{\left(RSS_{RR}-RSS_{UR}\right)/2}{RSS_{UR}/(n-4)}\sim F_{2,n-4}} \tag{12.7}$$

where $\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}$ $\left(RSS_{UR}\right)$ is residual sum of squares obtained from unrestricted model (12.4) and $\underset{\sim}{\mathbf{e}}_*'\underset{\sim}{\mathbf{e}}_*$ $\left(RSS_{RR}\right)$ is residual sum of squares obtained from restricted model (12.6).

If the calculated $F-$value from Eq. (12.7) is greater than the critical $F_{2,n-4}$ at a given l.o.s α, we reject $H_0:\alpha_1=\alpha_2$, $\beta_1=\beta_2$. In otherworld's, we conclude there is a structural change between wartime and peace time consumption functions. Otherwise we accept H_0 of no structural change.

12.3 Testing the structural change in intercept

For testing the structural change in the intercept i.e., $H_0:\alpha_1=\alpha_2=\alpha$, the restricted and unrestricted models may then be set up as follows:

unrestrited Model Restricted Model

$$\begin{bmatrix} \underset{\sim}{\mathbf{y}}_1 \\ \underset{\sim}{\mathbf{y}}_2 \end{bmatrix} = \begin{bmatrix} \underset{\sim}{\mathbf{i}}_1 & \underset{\sim}{\mathbf{0}} & \underset{\sim}{\mathbf{X}}_1 \\ \underset{\sim}{\mathbf{0}} & \underset{\sim}{\mathbf{i}}_2 & \underset{\sim}{\mathbf{X}}_2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{bmatrix} + \underset{\sim}{\mathbf{u}} \qquad \begin{bmatrix} \underset{\sim}{\mathbf{y}}_1 \\ \underset{\sim}{\mathbf{y}}_2 \end{bmatrix} = \begin{bmatrix} \underset{\sim}{\mathbf{i}}_1 \underset{\sim}{\mathbf{X}}_1 \\ \underset{\sim}{\mathbf{i}}_2 \underset{\sim}{\mathbf{X}}_2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \underset{\sim}{\mathbf{u}} \tag{12.8}$$

Now OLS may then be applied directly to each model in Eq. (12.8) and H_0 may be tested using the following test statistic $(q=1,\ k=3)$.

$$\boxed{F = \frac{\underset{\sim}{\mathbf{e}}'_* \underset{\sim}{\mathbf{e}}_* - \underset{\sim}{\mathbf{e}}' \underset{\sim}{\mathbf{e}}}{\underset{\sim}{\mathbf{e}}' \underset{\sim}{\mathbf{e}} / (n-3)} = \frac{RSS_{RR} - RSS_{UR}}{RSS_{UR}/(n-3)} \sim F_{1,n-3}} \tag{12.9}$$

All the above notations are as in the usual way.

If the calculated $F-$value from Eq. (12.7) is greater than the critical $F_{2,n-4}$ at a given l.o.s α , we reject $H_0 : \alpha_1 = \alpha_2$, $\beta_1 = \beta_2$. In otherworld's, we conclude there is a structural change between wartime and peace time consumption functions. Otherwise we accept H_0 of no structural change.

12.4 Testing the structural change in slope

For testing the structural change in the slope
$H_0 : \beta_1 = \beta_2 = \beta$, the restricted and unrestricted models may then be set up as follows:

Restricted Model unrestrited Model

$$\begin{bmatrix} \underset{\sim}{\mathbf{y}}_1 \\ \underset{\sim}{\mathbf{y}}_2 \end{bmatrix} = \begin{bmatrix} \underset{\sim}{\mathbf{i}}_1 & \underset{\sim}{\mathbf{0}} & \underset{\sim}{\mathbf{X}}_1 \\ \underset{\sim}{\mathbf{0}} & \underset{\sim}{\mathbf{i}}_2 & \underset{\sim}{\mathbf{X}}_2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{bmatrix} + \underset{\sim}{\mathbf{u}} \qquad \begin{bmatrix} \underset{\sim}{\mathbf{y}}_1 \\ \underset{\sim}{\mathbf{y}}_2 \end{bmatrix} = \begin{bmatrix} \underset{\sim}{\mathbf{i}}_1 & \underset{\sim}{\mathbf{X}}_1 & \underset{\sim}{\mathbf{0}} & \underset{\sim}{\mathbf{0}} \\ \underset{\sim}{\mathbf{0}} & \underset{\sim}{\mathbf{0}} & \underset{\sim}{\mathbf{i}}_2 & \underset{\sim}{\mathbf{X}}_2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \underset{\sim}{\mathbf{u}} \tag{12.10}$$

where $\underset{\sim}{\mathbf{i}}_1$ denoted a column vector of n_1 units, $\underset{\sim}{\mathbf{i}}_2$ is column vector of n_2 units, $\underset{\sim}{\mathbf{X}}_1$ is a column vector of n_1 observations on wartime income, and $\underset{\sim}{\mathbf{X}}_2$ a column vector n_2 observations on peace time income OLS may then be applied directly to each model in Eq. (12.10) and H_0 may be tested using the following formula $(q=1,\ k=4)$.

$$\boxed{F = \frac{\underset{\sim}{\mathbf{e}}'_* \underset{\sim}{\mathbf{e}}_* - \underset{\sim}{\mathbf{e}}' \underset{\sim}{\mathbf{e}}}{\underset{\sim}{\mathbf{e}}' \underset{\sim}{\mathbf{e}} / (n-4)} = \frac{RSS_{RR} - RSS_{UR}}{RSS_{UR}/(n-4)} \sim F_{1,n-4}} \tag{12.11}$$

where $\underset{\sim}{\mathbf{e}}'_* \underset{\sim}{\mathbf{e}}_*$ (RSS_{RR}) and $\underset{\sim}{\mathbf{e}}' \underset{\sim}{\mathbf{e}}$ (RSS_{UR}) are respectively the residuals sum of squares obtained from restricted and unrestricted models.

If the calculated $F-$value from Eq. (12.7) is greater than the critical $F_{2,n-4}$ at a given l.o.s α , we reject $H_0 : \alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$. In other words, we conclude there is a structural change between wartime and peace time consumption functions. Otherwise we accept H_0 of no structural change.

Note: In the above test it may be noted that the unrestricted model in testing $H_0:\alpha_1=\alpha_2$, is the restricted model in testing $H_0:\beta_1=\beta_2$.

12.5 Tests of structural change with *k* variables

Let us write the usual general linear model

$$\underset{\sim}{\mathbf{y}}=\mathbf{X}\underset{\sim}{\boldsymbol{\beta}}+\underset{\sim}{\mathbf{u}} \tag{12.12}$$

Suppose we have war time data, and peace time data on Y, $\mathrm{X}_2,\mathrm{X}_3,\ldots,\mathrm{X}_\mathrm{k}$ variables.
The general linear model for wartime data becomes

$$\underset{\sim}{\mathbf{y}}_1=\mathbf{X}_1\underset{\sim}{\boldsymbol{\beta}}_1+\underset{\sim}{\mathbf{u}}_1 \tag{12.13}$$

and for peace time data Eq. (12.12) may be written as

$$\underset{\sim}{\mathbf{y}}_2=\mathbf{X}_2\underset{\sim}{\boldsymbol{\beta}}_2+\underset{\sim}{\mathbf{u}}_2 \tag{12.14}$$

where $\underset{\sim}{\mathbf{y}}_1$ and $\underset{\sim}{\mathbf{y}}_2$ are respectively vectors of n_1 and n_2 observations on endogenous variable in wartime and peace time $\mathbf{X}_1$ and $\mathbf{X}_2$ are respective data matrices of $n_1\mathrm{xk}$ and $n_2\mathrm{xk}$ orders on explanatory variables in wartime and peace time. $\underset{\sim}{\mathbf{u}}_1$ and $\underset{\sim}{\mathbf{u}}_2$ are the respective disturbance vectors and $\underset{\sim}{\boldsymbol{\beta}}_1$, $\underset{\sim}{\boldsymbol{\beta}}_2$ are the respective vectors of unknown parameters in both the models.
The models (12.13) and (12.14) can be combined as

$$\begin{bmatrix}\underset{\sim}{\mathbf{y}}_1\\ \underset{\sim}{\mathbf{y}}_2\end{bmatrix}=\begin{bmatrix}\mathbf{X}_1 & 0\\ 0 & \mathbf{X}_2\end{bmatrix}\begin{bmatrix}\underset{\sim}{\boldsymbol{\beta}}_1\\ \underset{\sim}{\boldsymbol{\beta}}_2\end{bmatrix}+\begin{bmatrix}\underset{\sim}{\mathbf{u}}_1\\ \underset{\sim}{\mathbf{u}}_2\end{bmatrix} \tag{12.15}$$

The model (12.15) is called the **unrestricted model**.

Let us partition $\mathbf{X}_1$ and $\mathbf{X}_2$ by the first column of units and the remaining *k-1* columns of observations on the explanatory variables as follows:

$$\mathbf{X}_1=\begin{bmatrix}\underset{\sim}{\mathbf{i}}_1 & \mathbf{X}_1^*\end{bmatrix} \text{ and } \mathbf{X}_2=\begin{bmatrix}\underset{\sim}{\mathbf{i}}_2 & \mathbf{X}_1^*\end{bmatrix}$$

Now we consider the following 3 types of models:

Model-I (common regression for both periods):

$$\begin{bmatrix}\underset{\sim}{\mathbf{y}}_1\\ \underset{\sim}{\mathbf{y}}_2\end{bmatrix}=\begin{bmatrix}\underset{\sim}{\mathbf{i}}_1 & \mathbf{X}_1^*\\ \underset{\sim}{\mathbf{i}}_2 & \mathbf{X}_2^*\end{bmatrix}\underset{\sim}{\boldsymbol{\beta}}+\begin{bmatrix}\underset{\sim}{\mathbf{u}}_1\\ \underset{\sim}{\mathbf{u}}_2\end{bmatrix}$$

Model-II (differential intercepts and common vector of regression slopes):

$$\begin{bmatrix}\underset{\sim}{\mathbf{y}}_1\\ \underset{\sim}{\mathbf{y}}_2\end{bmatrix}=\begin{bmatrix}\underset{\sim}{\mathbf{i}}_1 & \underset{\sim}{\mathbf{0}} & \mathbf{X}_1^*\\ \underset{\sim}{\mathbf{0}} & \underset{\sim}{\mathbf{i}}_2 & \mathbf{X}_1^*\end{bmatrix}\begin{bmatrix}\alpha_1\\ \alpha_2\\ \underset{\sim}{\boldsymbol{\beta}}^*\end{bmatrix}+\begin{bmatrix}\underset{\sim}{\mathbf{u}}_1\\ \underset{\sim}{\mathbf{u}}_2\end{bmatrix}$$

Model-III (Differential intercepts and differential slopes):

$$\begin{bmatrix}\underset{\sim}{\mathbf{y}}_1\\ \underset{\sim}{\mathbf{y}}_2\end{bmatrix}=\begin{bmatrix}\underset{\sim}{\mathbf{i}}_1 & \underset{\sim}{\mathbf{0}} & \mathbf{X}_1^* & \underset{\sim}{\mathbf{0}}\\ \underset{\sim}{\mathbf{0}} & \underset{\sim}{\mathbf{i}}_2 & \underset{\sim}{\mathbf{0}} & \mathbf{X}_1^*\end{bmatrix}\begin{bmatrix}\alpha_1\\ \alpha_2\\ \underset{\sim}{\boldsymbol{\beta}}_1^*\\ \underset{\sim}{\boldsymbol{\beta}}_2^*\end{bmatrix}+\begin{bmatrix}\underset{\sim}{\mathbf{u}}_1\\ \underset{\sim}{\mathbf{u}}_2\end{bmatrix} \tag{12.16}$$

where we have partitioned the *k*-elements $\underset{\sim}{\boldsymbol{\beta}}$ vector as

$$\underset{\sim}{\boldsymbol{\beta}} = \begin{bmatrix} \alpha \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{bmatrix} = \begin{pmatrix} \alpha \\ \underset{\sim}{\boldsymbol{\beta}}^* \end{pmatrix} \tag{12.17}$$

Application of OLS to each model will yield a residual sum of squares (RSS) with an associated number of degrees of freedom as indicated by

Model I:	**RSS_1**	**n-k**
Model II:	**RSS_2**	**n-k-1**
Model III:	**RSS_3**	**n-2k**

where $n = n_1 + n_2$ indicates the total number of observations in the combined samples. The test statistics for various hypothesis are then as follows:

$H_0 : \alpha_1 = \alpha_2$ **Test of differential intercepts**

$$\boxed{F = \frac{\mathrm{RSS_1\text{-}RSS_2}}{\mathrm{RSS_2}/(n-k-1)} \sim F_{(1,n-k-1)}} \tag{12.18}$$

$H_0 : \underset{\sim}{\boldsymbol{\beta}}_1^* = \underset{\sim}{\boldsymbol{\beta}}_2^*$ **Test of differential slope vectors**

$$\boxed{F = \frac{(\mathrm{RSS_2\text{-}RSS_3})/(\mathrm{k\text{-}1})}{\mathrm{RSS_3}/(n-2k)} \sim F_{(k,n-2k)}} \tag{12.19}$$

$H_0 : \underset{\sim}{\boldsymbol{\beta}}_1 = \underset{\sim}{\boldsymbol{\beta}}_2$ **Test of differential regressions (intercepts and slopes)**

$$\boxed{F = \frac{(\mathrm{RSS_1\text{-}RSS_3})/\mathrm{k}}{\mathrm{RSS_3}/(n-2k)} \sim F_{(k,n-2k)}} \tag{12.20}$$

The degrees of freedom in the numerator are simply obtained as the difference in the degrees of freedom of the two residual sums of squares in the numerator. This is equal to the number of restrictions involved when going from the unrestricted to the restricted model.

In the above tests, if the computed F-value does not exceed the critical F-value taken from the F-table at a given l.o.s, then we accept H_0. Otherwise, reject H_0.

Note: In the similar manner, one can test the structural change in a subset of coefficients (i.e, the stability of a subset of coefficients). The principle of the test is the same as in all the test of structural change.

12.6 Chow Test for the equality of Two regression Equations :

Suppose we have data on the variables Y (dependent variable) and *k-1* explanatory variables $X_2, X_3, \ldots, X_k$ for two sub periods namely sub period–I and sub period–II. Further let

Let us suppose there are n_1 sets of observations in sub period–I and n_2 sets of observations in sub period–II.

Now the GLM for sub period–I may be written as

$$\underset{n_1\times 1}{\underset{\sim}{\mathbf{y}}_1} = \underset{n_1\times k}{\mathbf{X}_1}\ \underset{k\times 1}{\underset{\sim}{\boldsymbol{\beta}}_1} + \underset{n_1\times 1}{\underset{\sim}{\mathbf{u}}_1} \tag{12.21}$$

Similarly the GLM for sub periods–II may be written as

$$\underset{n_2\times 1}{\underset{\sim}{\mathbf{y}}_2} = \underset{n_2\times k}{\mathbf{X}_2}\ \underset{k\times 1}{\underset{\sim}{\boldsymbol{\beta}}_2} + \underset{n_2\times 1}{\underset{\sim}{\mathbf{u}}_2} \tag{12.22}$$

where $\underset{\sim}{\mathbf{y}}_1$ and $\underset{\sim}{\mathbf{y}}_2$ are respectively vectors of n_1 and n_2 observations on endogenous variable in wartime and peace time $\mathbf{X}_1$ and $\mathbf{X}_2$ are respective data matrices of $n_1\mathrm{xk}$ and $n_2\mathrm{xk}$ orders on explanatory variables in wartime and peace time. $\underset{\sim}{\mathbf{u}}_1$ and $\underset{\sim}{\mathbf{u}}_2$ are the respective disturbance vectors and $\underset{\sim}{\boldsymbol{\beta}}_1,\ \underset{\sim}{\boldsymbol{\beta}}_2$ are the respective vectors of unknown parameters in both the models.

Now our objective is to test the null hypothesis

$$H_0 : \underset{\sim}{\boldsymbol{\beta}}_1 = \underset{\sim}{\boldsymbol{\beta}}_2 \tag{12.23}$$

i.e., there is no structural change in the regressions of two sub periods

Or

H_0 : Both the regression equations are equal.

For this purpose, Chow-test is as follows:

This test assumes that

- $\underset{\sim}{\mathbf{u}}_1 \sim N\left(\underset{\sim}{\mathbf{0}}, \sigma^2\mathbf{I}_{n_1}\right)$ and $\underset{\sim}{\mathbf{u}}_2 \sim N\left(\underset{\sim}{\mathbf{0}}, \sigma^2\mathbf{I}_{n_2}\right)$. That is, the error terms in the two sub period regressions are normally distributed with same (homoscedastic) variance σ^2.
- The two error terms u_{1t} and u_{1t} are independently distributed.

Now, the mechanics of Chow-test are as follows:

1. Estimate the regression Eq. (12.21) of sub period–I, and compute $\underset{\sim}{\mathbf{e}}_1 = \underset{\sim}{\mathbf{y}}_1 - \mathbf{X}_1\hat{\underset{\sim}{\boldsymbol{\beta}}}_1$ and hence, obtain the residual sum of squares $\underset{\sim}{\mathbf{e}}_1'\underset{\sim}{\mathbf{e}}_1$ denoted by $\mathbf{RSS_1}$ with d.f. $n_1 - k$.
2. Similarly estimate the regression Eq. (12.22) of sub period–II, and compute $\underset{\sim}{\mathbf{e}}_2 = \underset{\sim}{\mathbf{y}}_2 - \mathbf{X}_2\hat{\underset{\sim}{\boldsymbol{\beta}}}_2$ and hence, obtain the residual sum of squares $\underset{\sim}{\mathbf{e}}_2'\underset{\sim}{\mathbf{e}}_2$ denoted by RSS_2 with d.f. $n_2 - k$.
3. Since the samples of period– I and period– II are independent, we can add $\mathbf{RSS_1}$ and $\mathbf{RSS_2}$ to obtain what may be called the ***unrestricted residual sum of squares*** $\left(\mathbf{RSS_{UR}}\right)$ given by

   $$\mathbf{RSS_{UR}} = \mathbf{RSS_1} + \mathbf{RSS_2} = \underset{\sim}{\mathbf{e}}_1'\underset{\sim}{\mathbf{e}}_1 + \underset{\sim}{\mathbf{e}}_2'\underset{\sim}{\mathbf{e}}_2 \tag{12.24}$$

   with $\left(n_1 + n_2 - 2k\right)$ d.f.
4. Under $H_0 : \underset{\sim}{\boldsymbol{\beta}}_1 = \underset{\sim}{\boldsymbol{\beta}}_2 = \underset{\sim}{\boldsymbol{\beta}}$ (say), the pooled regression of Eqs. (12.21) and (12.22) becomes

   $$\underset{n\times 1}{\underset{\sim}{\mathbf{y}}} = \underset{nxk}{\mathbf{X}}\ \underset{k\times 1}{\underset{\sim}{\boldsymbol{\beta}}} + \underset{n\times 1}{\underset{\sim}{\mathbf{u}}} \tag{12.25}$$

where $n = n_1 + n_2$, $\underset{\sim}{\mathbf{y}} = \begin{pmatrix} \underset{\sim}{\mathbf{y}}_1 \\ \underset{\sim}{\mathbf{y}}_2 \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$, $\underset{\sim}{\mathbf{u}} = \begin{pmatrix} \underset{\sim}{\mathbf{u}}_1 \\ \underset{\sim}{\mathbf{u}}_2 \end{pmatrix}$

Now, the regression model (12.25) is called as **restricted regression model** and by applying OLS to it yields restricted residual sum of squares denoted by $\mathbf{RSS}_{RR}$, given by

$$\mathbf{RSS_{RR}} = \underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}\ , \quad \text{where } \underset{\sim}{\mathbf{e}} = \underset{\sim}{\mathbf{y}} - \mathbf{X}\hat{\underset{\sim}{\boldsymbol{\beta}}} \quad \text{with (n-k) d.f.} \qquad (12.26)$$

5. Now, the idea behind the Chow test is that if in fact there is no structural change (that regression Eqs. (12.21) and (12.22) are essentially the same) then $\mathbf{RSS_{RR}}$ and $\mathbf{RSS_{UR}}$ should not be statistically different. Therefore if we form the F-ratio

$$\boxed{F = \frac{(\mathbf{RSS_{RR}} - \mathbf{RSS_{UR}})/k}{\mathbf{RSS_{UR}}/(n-2k)} \sim F_{k,n-2k}} \qquad (12.27)$$

then Chow has shown that under $H_0 : \underset{\sim}{\boldsymbol{\beta}}_1 = \underset{\sim}{\boldsymbol{\beta}}_2$, F ratio given above follows F distribution with $(k, n-2k)$ d.f.

6. If the above computed F **value does not exceed the critical** F **value** taken from the F table at a given level of significance, then **we accept**

   H_0 : No structural change between two regressions

   which means the two regressions are essentially the same,

   otherwise we reject H_0, that is the two regressions are different.

Note: In the above test, we may use the following alternative formula for computing the residual sum of squares,

RSS= Total sum of squares-Explained sum of squares = TSS-ESS (12.28)

There are some limitations about the Chow test that must be kept in mind:

1. The assumptions underlying the test must be fulfilled. For example, one should find out if the error variances in the regressions (12.21) and (12.22) are the same.
2. The Chow test will tell us only if the two regressions (12.21) and (12.22) are different, without telling us whether the difference is on account of the intercepts, or the slopes, or both. But in Lesson 13, on dummy variables, we will see how we can answer this question.

An application of Chow test:

The following table gives data on disposable personal income (Y) and personal savings(X) in billions of dollars for the United States for the sub period–I (1970-1981) and sub period-II (1982-1995). Using this data, t*est for the equality of the two regression equations of sub period–I (1970-1981) and sub period-II (1982-1995) using Chow test.*

Table 12.1: SAVINGS AND PERSONAL DISPOSABLE INCOME (BILLIONS OF DOLLARS), UNITED STATES, 1970–1995

Sub Period I			Sub Period II		
Observation	Savings (Y)	Income (X)	Observation	Savings (Y)	Income (X)
1970	61.0	727.1	1982	205.5	2347.3
1971	68.6	790.2	1983	167.0	2522.4
1972	63.6	855.3	1984	235.7	2810.0
1973	89.6	965.0	1985	206.2	3002.0

1974	97.6	1054.2	1986	196.5	3187.6
1975	104.4	1159.2	1987	168.4	3363.1
1976	96.4	1273.0	1988	189.1	3640.8
1977	92.5	1401.4	1989	187.8	3894.5
1978	112.6	1580.1	1990	208.7	4166.8
1979	130.1	1769.5	1991	246.4	4343.7
1980	161.8	1973.3	1992	272.6	4613.7
1981	199.1	2200.2	1993	214.4	4790.2
			1994	189.4	5021.7
			1995	249.3	5320.8

Source: Economic Report of the President, 1997, Table B-28, p. 332.

Solution:

Let us suppose the following three regressions:

for sub period-I: $Y_t = \alpha_1 + \beta_1 X_t + u_{1t}$ (12.29)

for sub period-II: $Y_t = \alpha_2 + \beta_2 X_t + u_{2t}$ (12.30)

for total period: $Y_t = \alpha + \beta X_t + u_t$ (12.31)

Now our object is to test the null hypothesis

$H_0 : \alpha_1 = \alpha_2 = \alpha$ and $\beta_1 = \beta_2 = \beta$ i.e. there is No Structural Change between the equations

i.e., The two regression equations (12.29) and (12.30) are **same**.

against the alternative hypothesis

H_1 : there is a Structural Change between the equations

i.e., The two regression equations (12.29) and (12.30) are **not the same**.

From Eq. (2.42) of Lesson 2, we have the relation

$$r^2 = 1 - \frac{RSS}{TSS} \quad \text{which implies} \quad RSS = \left(1 - r^2\right) TSS$$

In the following application of Chow test, we use this formula for computation of residual sum of squares (RSS) in each of the regressions (12.29) to (12.31). The advantage of this formula is it does not require the estimation of the parameters (and hence the residuals).

The various steps of Chow test are as follows:

1. Estimation of $\mathbf{RSS_1}$ based on regression Eq. (12.29) based on the sample of sub period-I: we have

n_1 =12 $\quad$ $k = 2$

$\sum X$ =15748.5 $\quad$ $\sum Y$ =1277.3

$\bar{X}$ =106.4417 $\quad$ $\bar{Y}$ =1312.3750

$\sum X^2$ =154199.23 $\quad$ $\sum Y^2$ =23218025.97

$\sum XY$ =1881149.97

Coefficient of determination (square of correlation coefficient) is

$$r_1^2 = \frac{\left(\frac{1}{n_1}\sum XY - \bar{X}\bar{Y}\right)^2}{\left(\frac{1}{n_1}\sum X^2 - \bar{X}^2\right)\left(\frac{1}{n_1}\sum Y^2 - \bar{Y}^2\right)} = 0.9021$$

Total sum of squares is given by

$$TSS_1 = \sum Y^2 - n\bar{Y}^2 = 18241.2892$$

Therefore, $RSS_1 = \left(1 - r_1^2\right) TSS_1 = 1785.0321$

2. Estimation of $\mathbf{RSS}_2$ based on regression Eq. (12.30) based on the sample of sub period-II:

we have

n_2 =14 $\quad k = 2$

$\sum X$ =53024.6 $\quad \sum Y$ =2937

$\bar{X}$ =3787.4714 $\quad \bar{Y}$ =209.7857

$\sum X^2$ =212664792.5 $\quad \sum Y^2$ =628760.26

$\sum XY$ =11299709.93

$r_2^2 = 0.2072$

$TSS_2 = 12619.6171$

$RSS_2 = \left(1 - r_2^2\right) TSS_2 = 10005.2207$

3. Estimation of $\mathbf{RSS}_{RR}$ based on the regression Eq. (12.31) using the pooled sample of the samples of sub period-I and sub period-II:

we have

n = 26 $\quad k = 2$

$\sum X$ = 68773.10 $\quad \sum Y$ = 4214.30

$\bar{X}$ = 2645.1192 $\quad \bar{Y}$ = 162.0885

$\sum X^2$ = 235882818.47 $\quad \sum Y^2$ = 782959.49

$\sum XY$ =13180859.9

$r^2 = 0.7672$

$TSS = 99870.0865$

Therefore, $\mathbf{RSS}_{RR} = \left(1 - r^2\right) TSS = 23248.2982$

From Eq. (12.24) $\mathbf{RSS}_{\mathbf{UR}} = 11790.2528$

Substituting $\mathbf{RSS}_{RR}$ and $\mathbf{RSS}_{\mathbf{UR}}$ in Eq. (12.27) we get

$F = \boldsymbol{10.69}$

At 1% of l.o.s. the critical F value at (2, 22) d.f. is 7.72

Since the above calculated F value is greater than the critical F value at 1% l.o.s. we reject H_0 .

Hence, ***we may conclude that there is a high significant difference between the two regression equations of sub period-I and sub period-II. In other words, we conclude that*** **there is a structural change between the two regression equations.**

Remark: A drawback of the above Chow test is that we could not tell whether the structural difference in the two regressions was because of differences in the intercept terms or the slope coefficients or both.

12.7 Test of Structural Change when n_1<k

A special problem arises if one of the sub periods has fewer observations than the number of parameters to be estimated in the model. Suppose, we may have a sample of $n_1(>\mathrm{k})$ observations on the variables $\mathrm{Y},\mathrm{X}_2,\mathrm{X}_3,\ldots,\mathrm{X}_\mathrm{k}$. An additional sample of $n_2(<\mathrm{k})$ observations on these variables become available and the question is whether they may be considered to come from the same population or the regression for the two samples have no structural change. The appropriate test is as follows.

1. To the first n_1 observations fit the OLS regression

   $$\underset{\sim}{\mathbf{y}}_1 = \mathbf{X}_1\underset{\sim}{\mathbf{b}}_1 + \underset{\sim}{\mathbf{e}}_1 \tag{12.32}$$

   where $\mathbf{X}_1$ is data matrix of n_1 observations on the set of $\mathrm{X}_2,\mathrm{X}_3,\ldots,\mathrm{X}_\mathrm{k}$ variables and compute the residual sum of squares $\underset{\sim}{\mathbf{e}}_1'\underset{\sim}{\mathbf{e}}_1$.
2. Pool the n_1+n_2 sample observations to give $\underset{\sim}{\mathbf{y}}$ and $\mathbf{X}$ and fit the least-squares regression

   $$\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\mathbf{b}} + \underset{\sim}{\mathbf{e}} \tag{12.33}$$

   and again compute the residual sum of squares $\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}$.
3. The test of the null hypothesis that the n_2 additional observations obey the same relation as the first sample is given by

   $$\boxed{F = \frac{(\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}} - \underset{\sim}{\mathbf{e}}_1'\underset{\sim}{\mathbf{e}}_1)/n_2}{\underset{\sim}{\mathbf{e}}_1'\underset{\sim}{\mathbf{e}}_1/(n_1-k)}} \tag{12.34}$$

   which is distributed as F with $(n_2, n_1\text{-k})$ degrees of freedom.
4. Compute the F statistic defined in Eq. (12.34) and reject the hypothesis of a common structure if F exceeds a preselected critical value of $F_{(n_2,n_1-k)}$.

12.8 Tests of Structural change (*k* variables, *p* periods)

Now let us consider the tests of structural change for the relationships of *k* explanatory variables between *p* periods. The tests need not be applied only across periods. We might examine the stability of a relation across countries, industries, social groups, or whatever.
The usual hierarchy of three models may be set up as follows:

i. $$\begin{bmatrix} \underset{\sim}{\mathbf{y}}_1 \\ \underset{\sim}{\mathbf{y}}_2 \\ \vdots \\ \underset{\sim}{\mathbf{y}}_p \end{bmatrix} = \begin{bmatrix} \underset{\sim}{\mathbf{i}}_1 & \mathbf{X}_1^* \\ \underset{\sim}{\mathbf{i}}_2 & \mathbf{X}_2^* \\ \vdots & \vdots \\ \underset{\sim}{\mathbf{i}}_p & \mathbf{X}_p^* \end{bmatrix} \begin{bmatrix} \alpha \\ \underset{\sim}{\boldsymbol{\beta}}^* \end{bmatrix} + \underset{\sim}{\mathbf{u}} \tag{12.35}$$

Common intercept, common slope vector in all p classes.

ii. $$\begin{bmatrix} \underset{\sim}{\mathbf{y}}_1 \\ \underset{\sim}{\mathbf{y}}_2 \\ \vdots \\ \underset{\sim}{\mathbf{y}}_p \end{bmatrix} = \begin{bmatrix} \underset{\sim}{\mathbf{i}}_1 & \underset{\sim}{\mathbf{0}} & \cdots & \underset{\sim}{\mathbf{0}} & \mathbf{X}_1^* \\ \underset{\sim}{\mathbf{0}} & \underset{\sim}{\mathbf{i}}_2 & \cdots & \underset{\sim}{\mathbf{0}} & \mathbf{X}_2^* \\ \vdots & \vdots & & \vdots & \vdots \\ \underset{\sim}{\mathbf{0}} & \underset{\sim}{\mathbf{0}} & \cdots & \underset{\sim}{\mathbf{i}}_p & \mathbf{X}_p^* \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \\ \underset{\sim}{\boldsymbol{\beta}}^* \end{bmatrix} + \underset{\sim}{\mathbf{u}} \tag{12.36}$$

Differential intercepts, common slope vector

iii. $$\begin{bmatrix} \underset{\sim}{\mathbf{y}}_1 \\ \underset{\sim}{\mathbf{y}}_2 \\ \vdots \\ \underset{\sim}{\mathbf{y}}_p \end{bmatrix} = \begin{bmatrix} \underset{\sim}{\mathbf{i}}_1 & \underset{\sim}{\mathbf{0}} & \cdots & \underset{\sim}{\mathbf{0}} & \mathbf{X}_1^* & 0 & \cdots & 0 \\ \underset{\sim}{\mathbf{0}} & \underset{\sim}{\mathbf{i}}_2 & \cdots & \underset{\sim}{\mathbf{0}} & 0 & \mathbf{X}_2^* & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \underset{\sim}{\mathbf{0}} & \underset{\sim}{\mathbf{0}} & \cdots & \underset{\sim}{\mathbf{i}}_p & 0 & 0 & \cdots & \mathbf{X}_p^* \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \\ \underset{\sim}{\boldsymbol{\beta}}_1^* \\ \underset{\sim}{\boldsymbol{\beta}}_2^* \\ \vdots \\ \underset{\sim}{\boldsymbol{\beta}}_p^* \end{bmatrix} + \underset{\sim}{\mathbf{u}} \tag{12.37}$$

Differential intercepts, differential slope vectors. Here $\underset{\sim}{\mathbf{i}}_i$ is the column vector of n_i units $\left(i=1,2,\ldots,p\right)$ and $\mathbf{X}_i^*$ is the $n_i \text{x} \left(\text{k-1}\right)$ matrix of observations on the explanatory variables in class i $\left(i=1,2,\ldots,p\right)$.

Application of OLS to each model will yield a residual sum of squares (**RSS**) with an associated number of d.f. as indicated by

Model I:	**RSS_1**	**n-k**
Model II:	**RSS_2**	**n-k-p+1**
Model III:	**RSS_3**	**n-pk**

where $n = n_1 + n_2 + \ldots + n_p$ indicates the total number of observations in the combined samples.

The test statistics for various hypotheses are as follows:

Test of differential intercepts:

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_p$$

$$\boxed{F = \frac{\left(\text{RSS}_1\text{-RSS}_2\right)/\left(p\text{-}1\right)}{\text{RSS}_2/\left(n-k-p+1\right)} \sim F_{(p-1,n-k-p+1)}} \tag{12.38}$$

Test of differential slope vectors:

$$H_0 : \underset{\sim}{\boldsymbol{\beta}}_1^* = \underset{\sim}{\boldsymbol{\beta}}_2^* = \ldots = \underset{\sim}{\boldsymbol{\beta}}_p^*$$

$$\boxed{F = \frac{\left(\mathrm{RSS_2\text{-}RSS_3}\right)/\left[(\mathrm{k\text{-}1})(p-1)\right]}{\mathrm{RSS_3}/(n-pk)}} \qquad (12.39)$$

Test of differential regressions (intercepts and slopes):

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_p \text{ and } \underset{\sim}{\boldsymbol{\beta}}_1^* = \underset{\sim}{\boldsymbol{\beta}}_2^* = \ldots = \underset{\sim}{\boldsymbol{\beta}}_p^*$$

$$\boxed{F = \frac{\left(\mathrm{RSS_1\text{-}RSS_3}\right)/\left[(p-1)k\right]}{\mathrm{RSS_3}/(n-pk)}} \qquad (12.40)$$

The degrees of freedom in the numerator are simply obtained as the difference in the degrees of freedom of the two residual sums of squares in the numerator. This is equal to the number of restrictions involved in going from the unrestricted to the restricted model;
In the above tests, if the computed F-value does not exceed the critical F-value taken from the F-table at a given l.o.s, then we accept H_0. Otherwise, reject H_0.

Note: In the similar manner, one can test the structural change in a subset of coefficients (i.e, the stability of a subset of coefficients). The principle of the test is the same as in all the test of structural change.

12.9 Self Assessment Questions

1. Explain Chow-test for comparison of two regression equations.
2. Describe a test for testing the equality of two regression equations.
3. Describe a test for testing the equality of slopes in two regression equations.
4. Describe a test for testing the equality of intercepts in two regression equations.
5. Describe a method of testing the equality of two regression equations.

12.10 References

1. Gujarati, D.N. (2005): *Basic Econometrics, 4th Ed., Tata McGraw-Hill.*
2. Johnston, J. (1984): *Econometric Methods, 3rd Ed., McGraw-Hill,* New York.
3. Montgomery, D.C., Peck, E.A. and Geoffrey Vining, G. (2003): *Introduction to Linear Regression Analysis, 3rd Ed., Wiley*
4. Draper, N.R., and H. Smith(1998): *Applied Regression Analysis, 3rd Ed., John Wiley & Sons,* New York.
5. G.S. Maddala (2001): *Introduction to Econometrics, 3rd Ed., John Wiley & Sons, Ltd.*
6. Johnston, J. and DiNardo J (1997): *Econometric Methods*, 4th Ed., McGraw Hill.
7. Hill, Carter, William Griffiths, and George Judge(2001): *Undergraduate Econometrics, John Willey & Sons,* New York.
8. *Koutsoyiannis, A(1973): Theory of Econometrics, Harper & Row,* New York.

Lesson 13

DUMMY VARIABLES

13.0 Objective:

After studying this lesson, the student will understand clearly what does it mean by a *dummy variable*, how the dummy variables can be used for comparing two or more regression equations, and what is its advantage when compared with Chow test for equality of two regressions.

Structure of the Lesson:

13.1 Introduction (The Nature of Dummy Variables)

In regression analysis it frequently happens that the dependent variable is influenced, not only by variables which can be readily quantified on some well known defined scale (e.g. income, output, prices, costs, height, and temperature) but also by variables which are essentially qualitative in nature (e.g. sex, race, color, religion, nationality, wars, earthquakes, strikes, political upheavals, and changes in government economic policy). For example, holding all other factors constant, female college teachers are found to found to earn less than their male counter parts, and non-whites are found to earn less than whites. This may result from sex or race discrimination, but whatever the reason, qualitative variables such as sex and race do influence the dependent variable and clearly should be included among the explanatory variables.

Since such qualitative variables usually indicate the presence or absence of a "quality" or an attribute, such as male or female, black or white, or Catholic or non-Catholic, one method of "quantifying" such attributes is by constructing artificial variables which take on values 1 or 0, 0 indicating the absence of an attribute, and 1 indicating the presence of that attribute variable which assume such 0 and 1 values are called dummy variables. Alternative names are indicator variables, binary variables, categorical variables, qualitative variables, and dichotomous variables.

Thus dummy variables are specially constructed variables which may be used to represent various factors such as

1. Temporal effects (wars, earthquakes, strikes, etc.)
2. Spatial effects (regional differences, nationality, etc.)
3. Qualitative variables (sex, race, color, etc.)
4. Broad grouping of quantitative variables (grouping of age, income etc.)

Under the heading of temporal effects we sometimes postulate that a behavioral relation shifts between one period and another; for example the consumption function might be expected to show a down ward shift in wartime compared with its peace time position, or a wage-determination equation might shift with a change of political regime or many relations may be expected to show seasonal shifts, if we are dealing with quarterly or monthly data. Spatially we sometimes expect shift in economic functions between one region of a country and another as a consequence of regional differences in economic structure and prospects. Then qualitative variables such as sex, marital status, social or occupational class will often play an important role in determining economic behavior and must be incorporated in the estimation process. Finally, we may sometimes have fully ordinal variables such as income and age but a broad grouping may be sufficient for the purpose in hand. All of these cases may be handled by the specification of appropriate dummy variables.

Dummy variables are a data-classifying device in that they divide a sample into various subgroups based on qualities or attributes (gender, marital status, race, religion, etc.) and *implicitly* allow one to run individual regressions for each subgroup. If there are differences in the response of the regressand to the variation in the qualitative variables in the various subgroups, they will be reflected in the differences in the intercepts or slope coefficients, or both, of the various subgroup regressions.

Dummy variables can be incorporated in regression models just as easily as quantitative variables. As a matter of fact, a regression model may contain regressors that are exclusively dummy, or qualitative, in nature. Such models are called ANOVA models.

Although a versatile tool, the dummy variable technique needs to be handled carefully. *First,* if the regression contains a constant term, the number of dummy variables must be one less than the number of classifications of each qualitative variable. *Second,* the coefficient attached to the dummy variables must *always* be interpreted in relation to the base, or reference, group—that is, the group that receives the value of zero. The base chosen will depend on the purpose of research at hand. *Finally,* if a model has several qualitative variables with several classes, introduction of dummy variables can consume a large number of degrees of freedom. Therefore, one should always weigh the number of dummy variables to be introduced against the total number of observations available for analysis.

Among its various applications, this lesson considered but a few. These included (1) comparing two (or more) regressions, and (2) deseasonalizing time series data.

13.2 Regression on one quantitative variable and one dummy variable with two classes or categories

Suppose we have data on two variables

Y = Consumption expenditure

and X =Disposable income

The data cover two distinct sub periods, n_1 observations relating to wartime years and n_2 observations relating to peace time years. Suppose we wish to investigate whether there is any change, or shift, in the consumption function between the wartime and peace time periods. Such a change is referred to as a structural change or structural break. Let us denote the consumption functions by

$Y = \alpha_1 + \beta X + u$ wartime function

$Y = \alpha_2 + \beta X + u$ peace time function (13.1)

Here we assumed the slope coefficient β is common in both periods. Now, to see whether there is any structural change or not (that is to test $H_0 : \alpha_1 = \alpha_2$), we have two alternative procedures.

1. Write restricted (with $\alpha_1 = \alpha_2$) and unrestricted (with $\alpha_1 \neq \alpha_2$) models and obtain the residual sum of squares **(RSS)** in both the models and use the following formula
$$F = \frac{\text{restricted RSS - unrestricted RSS}}{\text{unrestricted RSS}/(n-2)} \sim F_{1,n-2}$$
2. By incorporating a dummy variable(s), appropriately, in the model.

The first procedure, we have already discussed elaborately in Lesson 12. Now, we discuss the second procedure in this lesson.

Now in order to test the null hypothesis

$H_0 : \alpha_1 = \alpha_2$

we can combine the two consumption functions given in Eq. (13.1) by incorporating a dummy variable namely D and now, the regression equation of Y (consumption expenditure) on quantitative variable X (disposable income) and a dummy variable D which has two categories may be written as

$Y = \alpha_1 + (\alpha_2 - \alpha_1) D + \beta X + u$

D = 1 if observation belong to peace time year

= 0 if observation belong to war time year

After pooling the data of two periods, the model of the above regression equation becomes

$$Y_t = \alpha_1 + (\alpha_2 - \alpha_1) D_t + \beta X_t + u_t \quad t = 1, 2, \ldots, n \ (n = n_1 + n_2) \qquad (13.2)$$

where

Y_t = consumption expenditure in year 't'

X_t = disposable income in year 't'

u_t = disturbance term in year 't'

D_t =1 if t is peace time year

=0 if t is wartime year

What is the meaning of the above model. Assuming, as usual, that $E(u_t)=0$, we see that

Mean consumption function in wartime: $E(Y_t/X_t, D_t = 0) = \alpha_1 + \beta X_t$

Mean consumption function in peace time: $E(Y_t/X_t, D_t = 1) = \alpha_2 + \beta X_t$ (13.3)

Thus, the model (13.2) postulates that the wartime consumption function and peacetime consumption function have the same slope (β) but different intercepts (α_1 and α_2). In other words, it is assumed that the level of consumption expenditure in peace time and wartime are different, but the marginal propensity to consume (or rate of consumption expenditure) is same in both periods. If β is not common to both periods nothing to be gained by using dummy variables; One merely fits separate regressions to wartime data and peace time data.

The model (13.2) may be written in matrix form as

$$\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\gamma}} + \underset{\sim}{\mathbf{u}} \tag{13.4}$$

where

$$\underset{\sim}{\mathbf{y}} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & 0 & X_1 \\ 1 & 0 & X_1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & X_{n_1} \\ 1 & 1 & X_{n_1+1} \\ \vdots & \vdots & \vdots \\ 1 & 1 & X_{n_1+n_2} \end{pmatrix}, \ \underset{\sim}{\boldsymbol{\gamma}} = \begin{pmatrix} \alpha_1 \\ \alpha_2 - \alpha_1 \\ \beta \end{pmatrix}, \ \underset{\sim}{\mathbf{u}} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n_1} \\ u_{n_1+1} \\ \vdots \\ u_{n_1+n_2} \end{pmatrix}$$

From the model (13.2), it is clear that, α_1 gives the common intercept for both periods and $\alpha_2 - \alpha_1$ gives the additional intercept for peace time. Now testing the significance of D in equation (13.2) is, in effect, testing the hypothesis

$$H_0 : \alpha_2 - \alpha_1 = 0 \text{ i.e., } \alpha_2 = \alpha_1 \tag{13.5}$$

which is testing whether the peace time and the war time intercepts are significantly different or not. **Thus, using a dummy variable in the equation, we can test the structural change in wartime and peace time consumption function**. Alternatively H_0 may be interpreted as there is no significant difference between the levels of the consumption expenditure pertaining to wartime period and peace time periods, In other words, there is no effect of war on the level of the consumption expenditure if we accept H_0.

The statistical significance of the estimated $\widehat{\alpha_2 - \alpha_1}$ may be tested based on the traditional **t-test** by running the regression model (13.4) for the given data. If **t-test** shows that $\widehat{\alpha_2 - \alpha_1}$ is statistically significant, we reject H_0: the levels of mean consumption expenditure are same for both periods. Thus we conclude that the intercepts of the two equations are different.

Remarks:

1. To distinguish the two categories of the qualitative variable, we have introduced only one dummy variable. Hence, one dummy variable suffices to distinguish two

categories. **The general rule is this:** If a qualitative variable has ***m*** categories, introduce only ***m -1*** dummy variables.

2. The group, category or classification that is assigned the value '0' is often referred to as the base, control, comparison or omitted category. It is the base in the sense that comparisons are made with that category.

13.3 Regression on one quantitative variable and one qualitative variable with more than two classes

Suppose that based on the cross-sectional data we want to regress the annual expenditure on health care by an individual on the income and education of the individual. Since the variable education is qualitative in nature, suppose we consider three mutually exclusive levels of education:

1. Less than high school
2. High school
3. College

Now, unlike the previous case, we have more than two categories of the qualitative variable 'education'. Therefore, following the rule that the number of dummies be less one than the number of categories of the variable, we should introduce only two dummies to take care of the 3 levels of 'education'. Assuming that the 3 educational groups have a common slope but different intercepts in the regression of 'annual expenditure on health care' on 'annual income', we can use the following model:

$$Y_i = \alpha_0 + (\alpha_1 - \alpha_0)D_{1i} + (\alpha_2 - \alpha_0)D_{2i} + \beta X_i + u_i, \qquad i = 1,2,\ldots,n \tag{13.6}$$

where i =sample unit
$n=n_1+n_2$
Y = annual expenditure on health care
X =annual income
D_1 =1 if high school education
=0 otherwise
D_2 = 1 if college education
=0 otherwise

Note that in the preceding assignment of the dummy variables we are arbitrarily treating the "less than high school education" category as base category.
Therefore, the intercept α_0 will reflect the intercept for this category. The differential intercepts $\alpha_1 - \alpha_0$ and $\alpha_2 - \alpha_0$ tell by how much the intercepts of the other two categories differ from the intercept of the base category.

From model (13.6), it is clear that

1. The intercept of the group of individuals 'less than high school education' is α_0
2. The intercept of the group of individuals of 'high school education' is α_1
3. The intercept of the group of individuals of 'college education' is α_2

Alter running the regression (13.6), one can easily find out whether the differential intercepts $\alpha_1 - \alpha_0$ and $\alpha_2 - \alpha_0$ are individually statistically significant, that is, different from the base group. We may also test $H_0: \alpha_1 - \alpha_0 = 0 \text{ and } \alpha_2 - \alpha_0 = 0$ simultaneously using ANOVA technique.

If $\alpha_1 - \alpha_0$ is **not statistically significant**, then α_0 is the common intercept for both the categories of individuals of 'less than high school education' as well as 'high school' education'. If $\alpha_1 - \alpha_0$ is **statistically significant**, then α_1 is the intercept of the category of individuals 'high school education'.

13.4 Regression on one Quantitative variable and two Qualitative Variables (two dummy variables)

The technique of dummy variable can be easily extended to handle more than one qualitative variable. Suppose we have data on two variables namely

Y =annual salary of a college teacher

X = years of teaching experience

The data covers both male and female teachers as well as color (black, red or white) of the teacher.

Now we can write the regression model with one quantitative variable and two qualitative variables (sex & color) as

$$Y_i = \alpha_0 + \alpha_1 S_i + \gamma_1 C_{1i} + \gamma_2 C_{2i} + \beta X_i + u_i \qquad (13.7)$$

where i = i^{th} sample unit

Y = annual salary

X =year of teaching experience

$$S = \begin{cases} 1 & \text{if male} \\ 0 & \text{otherwise} \end{cases}$$

$$C_1 = \begin{cases} 1 & \text{if red} \\ 0 & \text{otherwise} \end{cases}$$

$$C_2 = \begin{cases} 1 & \text{if white} \\ 0 & \text{otherwise} \end{cases}$$

Notice that the first qualitative variable sex has two categories and hence needs one dummy variable(s) where as the second qualitative variable color has three categories and hence needs two dummy variables, C_1 and C_2. Note also that the omitted or base category now is "black female teacher", Once again, the model (13.7) assumes common slope for all categories and differ only in the intercept coefficients. OLS estimation of the model (13.7) enable us to test a variety of hypothesis. Thus if α_1 is statistically significant, it will mean that sex has an impact on the teacher's salary. Similarly if $\gamma_1\ (\gamma_2)$ is statistically significant it means that the mean salary of the red (white) teacher is significantly different from that of a black teacher.

Category	**intercept**
Black female teacher	α_0
Black male teacher	$\alpha_0+\alpha_1$
Red female teacher	$\alpha_0+\gamma_1$
Red male teacher	$\alpha_0+\alpha_1+\gamma_1$
White female teacher	$\alpha_0+\gamma_2$
While male teacher	$\alpha_0+\alpha_1+\gamma$

13.5 A generalization (two or more sets of dummy variables)

Following the preceding discussion, we can extend our model to include more than one quantitative variable and more than two qualitative variables. **The only precaution to be taken is that the number of dummies for each qualitative variable should be one less than the number of categories of that variable.** An example is given in the following.

Suppose that we have cross-sectional budget data for a number of quarters and we postulate that the consumption (q) of some commodity is given by

q =f(seasonal Dummies, Social Class factors other Economic variables)

If there are 4 seasons and 3 social classes then one way to set up this relation is

$$q=\alpha_0+\alpha_1Q_1+\alpha_2Q_2+\alpha_3Q_3+\beta_1S_1+\beta_2S_2+\gamma_2X_2+\ldots+\gamma_kX_k+u \qquad (13.8)$$

where

$$Q_i=\begin{cases}1 & \text{if observation relates to Quarter '}i\text{', } i\text{=1,2,3}\\ 0 & \text{otherwise}\end{cases}$$

$$S_j=\begin{cases}1 & \text{if observation relates to Social class '}j\text{', } j\text{=1,2}\\ 0 & \text{otherwise}\end{cases}$$

and $X_2, X_3, \ldots, X_k$ are the set of *k-1* economic variables such as income and relative prices. Here the base or omitted category is IV-Quarter and Social class III

Category		**Intercept**
IV Quarter & Social class	III:	α_0
IV Quarter & Social class	I :	$\alpha_0+\beta_1$
IV Quarter & Social class	II :	$\alpha_0+\beta_1$
I Quarter & Social class	III:	$\alpha_0+\alpha_1$
I Quarter & Social class	I :	$\alpha_0+\alpha_1+\beta_1$
I Quarter & Social class	II :	$\alpha_0+\alpha_1+\beta_2$
II Quarter & Social class	III:	$\alpha_0+\alpha_2$
II Quarter & Social class	I :	$\alpha_0+\alpha_2+\beta_1$
II Quarter & Social class	II :	$\alpha_0+\alpha_2+\beta_2$
III Quarter & Social class	III:	$\alpha_0+\alpha_3$
III Quarter & Social class	I :	$\alpha_0+\alpha_3+\beta_1$
III Quarter & Social class	II :	$\alpha_0+\alpha_3+\beta_2$

Another common application of this type occurs in the estimation of production function. Suppose we have data on output (Y) and inputs (X) for m firms over n years and we wish to estimate a production function. In doing so we might allow specifically for "year effects" and "firm effects" by fitting

$$Y_{ij} = \mu + \alpha_1 T_1 + \alpha_2 T_2 + \ldots + \alpha_{n-1} T_{n-1} + \beta_1 F_1 + \beta_2 F_2 + \ldots + \beta_{m-1} F_{m-1} + \gamma_2 X_{2ij} + \gamma_3 X_{3ij} + \ldots + \gamma_k X_{kij} + u_{ij}$$
$$i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n \tag{13.9}$$

where $F_i = \begin{cases} 1 \text{ if observation belongs to i}^{\text{th}} \text{ firm (i=1,2,…,m-1)} \\ 0 \qquad \text{otherwise} \end{cases}$

$T_j = \begin{cases} 1 \text{ if observation belongs to j}^{\text{th}} \text{ firm (j=1,2,…,n-1)} \\ 0 \qquad \text{otherwise} \end{cases}$

13.6 The Use of Dummy Variables for Testing the Equality of Two Regressions- An Alternative to the Chow Test

In the last lesson, we have discussed the Chow test to examine the structural stability of a regression model. The example we discussed there related to the relationship between savings and income in the United States over the period 1970–1995. We divided the sample period into two, 1970–1981 and 1982–1995, and showed on the basis of the Chow test that there was a difference in the regression of savings on income between the two periods. However, we could not tell whether the difference in the two regressions was because of differences in the intercept terms or the slope coefficients or both. Very often this knowledge itself is very useful.

For explaining the use of dummy variables for testing the equality of two regressions, let us reproduce Eqs. (12.29) and (12.30) here.

for sub period-I (*1970–1981*): $Y_t = \alpha_1 + \beta_1 X_t + u_{1t}$ (13.10)

for sub period-II (*1982–1995*): $Y_t = \alpha_2 + \beta_2 X_t + u_{2t}$ (13.11)

We see that there are four possibilities, which are given below.

1. Both the intercept and the slope coefficients are the same in the two regressions. This is the case of **coincident regressions** ($\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$).
2. Only the intercepts in the two regressions are different but the slopes are the same. This is the case of **parallel regressions** ($\alpha_1 \neq \alpha_2$ and $\beta_1 = \beta_2$).
3. The intercepts in the two regressions are the same, but the slopes are different. This is the situation of **concurrent regressions** ($\alpha_1 = \alpha_2$ and $\beta_1 \neq \beta_2$).
4. Both the intercepts and slopes in the two regressions are different. This is the case of **dissimilar regressions** ($\alpha_1 \neq \alpha_2$ and $\beta_1 \neq \beta_2$).

The Chow test procedure discussed in the last lesson, as noted earlier, tells us only if two (or more) regressions are different without telling us what is the source of the difference (which of (2) to (4)). But, this problem can be solved by using dummy variables appropriately. In other words, the source of difference, if any, can be pinned down by considering the following pooled regression.

$$Y_t = \alpha_1 + (\alpha_2 - \alpha_1) D_t + \beta_1 X_t + (\beta_2 - \beta_1)(D_t X_t) + u_t \tag{13.12}$$

where Y = savings
X = income
t = time
D = 1 for observations in sub period-II (1982–1995)
= 0, otherwise (i.e., for observations in sub period-I (1970–1981))

Now, estimating the above pooled regression equation (13.12) is equivalent to estimating the two individual regression equations (13.10) and (13.11). Further, by testing the significance of the coefficients of the variables D_t and $D_t X_t$, we can decide which one of the above three possibilities ((2) to (4)) is the source of difference between the two regression equations (13.10) and (13.11).

An application on the use of dummy variables:

For demonstration of the above procedure let us reconsider the example, which is used for the demonstration of the Chow test for the equality of the two regression equations in Section 12.6 of Lesson 12. Let us pool all the observations (26 in all) of Table 12.1 and present the data in the flowing table along with the dummy variables D_t and $D_t X_t$.

t=Year-1969	Saving (Y_t)	Income (X_t)	D_t	$D_t X_t$
1	61.0	727.1	0	0
2	68.6	790.2	0	0
3	63.6	855.3	0	0
4	89.6	965.0	0	0
5	97.6	1054.2	0	0
6	104.4	1159.2	0	0
7	96.4	1273.0	0	0
8	92.5	1401.4	0	0
9	112.6	1580.1	0	0
10	130.1	1769.5	0	0
11	161.8	1973.3	0	0
12	199.1	2200.2	0	0
13	205.5	2347.3	1	2347.3
14	167.0	2522.4	1	2522.4
15	235.7	2810.0	1	2810.0
16	206.2	3002.0	1	3002.0
17	196.5	3187.6	1	3187.6
18	168.4	3363.1	1	3363.1
19	189.1	3640.8	1	3640.8
20	187.8	3894.5	1	3894.5
21	208.7	4166.8	1	4166.8
22	246.4	4343.7	1	4343.7

23	272.6	4613.7	1	4613.7
24	214.4	4790.2	1	4790.2
25	189.4	5021.7	1	5021.7
26	249.3	5320.8	1	5320.8

We estimate the multiple regression equation (13.12) with the above data using Minitab statistical package and presented the output below:

```
Predictor        Coef        StDev           t          P
Constant         1.02        20.16        0.05      0.960
D              152.48        33.08        4.61      0.000
X             0.08033      0.01450        5.54      0.000
DX           -0.06547      0.01598       -4.10      0.000

S = 23.15       R-Sq = 88.2%     R-Sq(adj) = 86.6%

The estimated pooled regression equation:
```

$$Y_t = 1.02 + 152.48 D_t + 0.08033 X_t - 0.06547 \left(D_t X_t \right)$$

```
Analysis of Variance

Source        DF          SS          MS          F         P
Regression     3       88080       29360      54.78     0.000
Error         22       11790         536
Total         25       99870
```

From t-tables, we have two-tailed critical t-value with 22 (26-4) d.f. at 1% l.o.s. is 2.819.

From the above estimated regression equation, we may notice that the t-values (4.61 and -4.1) of the regression coefficients of the dummy variables D and DX are in magnitude greater than the critical t-value (2.819), and hence we may conclude that the regression coefficients of both D and DX are statistically highly significant.

Thus the regression coefficients $\left(\alpha_2 - \alpha_1\right)$ and $\left(\beta_2 - \beta_1\right)$ of regression equations (13.12) are significantly different from zero.

Therefore, $\alpha_1 \neq \alpha_2$ and $\beta_1 \neq \beta_2$. i.e., the intercepts and slopes in the two regression equations (13.10) and (13.11) are different.

Thus we conclude the two regression equations (13.10) and (13.11) are dissimilar, which is due to the 4th possible of the difference (listed in the beginning of the section).

Thus, by introducing dummy variables D and DX in the regression equation (13.12), we are able to identify the source of the structural change between two regression equations (13.10) and (13.11).

In other words, using dummy variables in a regression equation in an appropriate manner, we are not only testing the structural change between two regression equations,

but also we are identifying the source of the structural change out of four possibilities (mentioned in the beginning of this section).

Here, it may be noted using Chow test, one can test whether there is a structural change or not between two given regression equations, but it cannot identify the source of structural change. Hence, using dummy variables in regression analysis is better alternative method instead of applying Chow test.

Note: one can easily extend the use of dummy variables for the case of more than two regression equations (or equivalently for more than two categories or classifications) as given below.

For instance, with quarterly data we may specify

$$\begin{aligned} Y_i = \alpha_1 + (\alpha_2 - \alpha_1)D_2 + (\alpha_3 - \alpha_1)D_3 + (\alpha_4 - \alpha_1)D_4 \\ + \beta_1 X + (\beta_2 - \beta_1)D_2 X + (\beta_3 - \beta_1)D_3 X + (\beta_4 - \beta_1)D_4 X + u \end{aligned} \quad (13.13)$$

where

$$D_i = \begin{cases} 1 \text{ if an observation in Quarter i (i=2,3,4)} \\ 0 \qquad \text{otherwise} \end{cases}$$

Eq. (13.15) allows intercepts and regression slopes to vary across all four class.

Testing the significance of individual regression coefficients of dummy variables D_2, D_3, D_4, D_2X, D_3X and D_4X in the above regression model (13.13) is equivalent to testing the homogeneity of intercepts and homogeneity of slopes between the regression equation of four classes. Thus, just by testing the significance of the individual regression coefficients of dummy variables in a single regression equation, we are able to compare 4 regression equations in various aspects.

13.7 The Use of Dummy Variables in Seasonal Analysis

Many economic time series based on monthly or quarterly data exhibit seasonal patterns (regular oscillatory movements). Examples are sales of department stores at Christmas and other major holiday times, demand for money (or cash balances) by households at holiday times, demand for ice cream and soft drinks during summer, prices of crops right after harvesting season, demand for air travel, etc. Often it is desirable to remove the seasonal factor, or *component,* from a time series so that one can concentrate on the other components, such as the trend. The process of removing the seasonal component from a time series is known as deseasonalization or seasonal adjustment, and the time series thus obtained is called the **deseasonalized** or **seasonally adjusted** time series. Important economic time series, such as the unemployment rate, the consumer price index (CPI), the producer's price index (PPI), and the index of industrial production, are usually published in seasonally adjusted form. There are several methods of deseasonalizing a time series, but we will consider only one of these methods, namely, the *method of dummy variables.*

Deseasonalization of time series data:

Suppose we have *4n* quarterly observations on a variable Y, such as unemployment, imports or food prices. Such variables are likely to display a pronounced seasonal movement, and for purposes of economic intelligence and policy it is important to produce a

"deseasonalized" series, from which one can better assess whether unemployment, say, is really increasing or decreasing. There are several methods of deseasonalizing series in practice, but here we are only concerned with applications of dummy variables.
Let Y_{ij} denoted the observation on Y in the jth quarter of ith year (j=1, 2, 3, 4; i=1,2,…,n). This series contain seasonal components apart from trend and/or cyclical components. Deseasonalizing the series means to eliminate the seasonal component from the series. This can be achieved using dummy variables in the regression analysis frame work as follows:

Suppose D_1, D_2, D_3 and D_4 are four dummy variables to represent the four seasonal (quarters) effects respectively defined as

$$D_{jt} = \begin{cases} 1 \text{ if t occur in quarter j, j=1, 2, 3, 4} \\ 0 \qquad\qquad\qquad\qquad \text{otherwise} \end{cases} \tag{13.14}$$

Similarly to represent the trend and/or cyclical effects in the series, we have to incorporate the following polynomial in time $'t'$ of sufficiently higher order $'p'$.

$$\alpha_1 t + \alpha_2 t^2 + \ldots + \alpha_p t^p \tag{13.15}$$

With the above explanation we may write the value of Y in the time period $'t'$ as

$$Y_t = \alpha_1 t + \alpha_2 t^2 + \ldots + \alpha_p t^p + \beta_1 D_{1t} + \beta_2 D_{2t} + \beta_3 D_{3t} + \beta_4 D_{4t} + u_t \tag{13.16}$$

where u is random effect
For *4n* quarterly observations on the variables Y, we can write the model in compact form from Eq. (13.16) as follows:

$$\underset{\sim}{\mathbf{y}} = \mathbf{P}\underset{\sim}{\boldsymbol{\alpha}} + \mathbf{D}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}} \tag{13.17}$$

where

$$\underset{\sim}{\mathbf{y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{4n} \end{bmatrix}_{4n\text{x}1}, \mathbf{P} = \begin{bmatrix} 1 & 1^2 & \cdots & 1^p \\ 2 & 2^2 & \cdots & 2^p \\ 3 & 3^2 & \cdots & 3^p \\ 4 & 4^2 & \cdots & 4^p \\ \vdots & \vdots & & \vdots \\ 4n & (4n)^2 & \cdots & (4n)^p \end{bmatrix}_{4n\text{x}p}, \underset{\sim}{\boldsymbol{\alpha}} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix}_{p\text{x}1},$$

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 \end{bmatrix}_{4n\text{x}4}, \underset{\sim}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}, \; and \; \underset{\sim}{\mathbf{u}} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{4n} \end{bmatrix}_{4n\text{x}1}$$

Now, the deseasonalized series would be now be defined as

$$\underset{\sim}{\mathbf{y}}_{ds} = \underset{\sim}{\mathbf{y}} - \mathbf{D}\underset{\sim}{\mathbf{b}} \tag{13.18}$$

where $\underset{\sim}{\mathbf{b}} = (\mathbf{D}'\mathbf{N}\mathbf{D})^{-1}\mathbf{D}'\mathbf{N}\underset{\sim}{\mathbf{y}}, \quad \mathbf{N} = \mathbf{I} - \mathbf{P}(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'$ (13.19)

Substituting Eq. (13.19) in Eq. (13.18) we get

$$\underset{\sim}{\mathbf{y}}_{ds} = \mathbf{T}\underset{\sim}{\mathbf{y}}, \text{ where } \mathbf{T} = \mathbf{I} - \mathbf{D}\left(\mathbf{D}'\mathbf{N}\mathbf{D}\right)^{-1}\mathbf{D}'\mathbf{N} \qquad (13.20)$$

Thus the deseasonalized series of $\underset{\sim}{\mathbf{y}}$ can be expressed as a linear transformation of $\underset{\sim}{\mathbf{y}}$. Here, it may be noted $\mathbf{T}$ is idempotent matrix, but not symmetric.

13.8 Self Assessment Questions

1. Explain dummy variable regression models and explain how this is a better approach than Chow-test for comparison of two regression equations.
2. What are dummy variables? Bring out its applicability while testing the equality of two regression equations each with one exogenous variable.
3. Discuss the use of dummy variables with a suitable example.
4. Define Dummy variables-explain how you would use them to examine the presence or absence of structural changes in the frame work of linear models.
5. Explain the use of dummy variables in seasonal adjustment of time series data.

13.9 References

1. Gujarati, D.N. (2005): *Basic Econometrics, 4th Ed., Tata McGraw-Hill.*
2. Johnston, J. (1984): *Econometric Methods, 3rd Ed., McGraw-Hill,* New York.
3. Montgomery, D.C., Peck, E.A. and Geoffrey Vining, G. (2003): *Introduction to Linear Regression Analysis, 3rd Ed., Wiley*
4. Draper, N.R., and H. Smith(1998): *Applied Regression Analysis, 3rd Ed., John Wiley & Sons,* New York.
5. G.S. Maddala (2001): *Introduction to Econometrics, 3rd Ed., John Wiley & Sons, Ltd.*
6. Johnston, J. and DiNardo J (1997): *Econometric Methods*, 4th Ed., McGraw Hill.
7. Hill, Carter, William Griffiths, and George Judge(2001): *Undergraduate Econometrics, John Wiley & Sons,* New York.
8. *Koutsoyiannis, A(1973): Theory of Econometrics, Harper & Row,* New York.

Lesson 14

MULTICOLLINEARITY

14.0 Objective:

After studying this lesson, the student will have some clarity on the concept, nature and consequences of **multicollinearity**. From this lesson, the student will know how to detect multicollinearity using some tests and what the remedial measures are for multicollinearity problem.

Structure of the Lesson:

14.1 Introduction

Very often the data we use in multiple regression analysis cannot give decisive answers to the question we pose. This is because the standard errors are very high or equivalently the t-ratios are very low. The confidence intervals for the parameters of interest are thus very wide. This sort of situation occurs when the explanatory variables display little variation and/or high intercorrelations. The situation where the explanatory variables are highly intercorrelated is referred to as multicollinearity. When the explanatory variables are highly intercorrelated, it becomes difficult to disentangle the separate effects of each of the explanatory variables on the explained variable. The practical questions we need to ask are how high these intercorrelations have to be to cause problems in our inference about the individual parameters and what we can do about this problem.

The term "multicollinearity" was first introduced in 1934 by Ragnar Frisch in his book on confluence analysis and referred to a situation where the variables dealt with are subject to two or more relations. In his analysis, there was no dichotomy of explained and explanatory variables. It was assumed that all variables were subject to error and given the sample variances and covariances, the problem was to estimate the different linear relationships among the true variables. We will, however, be discussing the multicollinearity problem as it is commonly discussed in multiple regression analysis, namely, the problem of high intercorrelations among the explanatory variables.

14.2 The Nature of Multicollinearity

Originally multicollinearity meant the existence of a “perfect,” or exact, linear relationship among some or all explanatory variables of a regression model. For the *k*-variable regression involving explanatory variable $X_1, X_2, \ldots, X_k$ (where X_1 = 1 for all observations to allow for the intercept term), an exact linear relationship is said to exist if the following condition is satisfied:

$$\lambda_1 X_1 + \lambda_2 X_2 + \ldots + \lambda_k X_k = 0 \qquad (14.1)$$

where $\lambda_1, \lambda_2, \ldots, \lambda_k$ are constants such that not all of them are zero simultaneously.

Today, however, the term multicollinearity is used in a broader sense to include the case of perfect multicollinearity, as shown by Eq. (14.1), as well as the case where the X variables are intercorrelated but not perfectly so, as follows:

$$\lambda_1 X_1 + \lambda_2 X_2 + \ldots + \lambda_k X_k + v_i = 0 \qquad (14.2)$$

where v_i is a stochastic error term.

To see the difference between *perfect* and *less than perfect* multicollinearity, assume, for example, that $\lambda_2 \neq 0$. Then, Eq. (14.1) can be written as

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \ldots - \frac{\lambda_k}{\lambda_2} X_{ki} \qquad (14.3)$$

which shows how X_2 is exactly linearly related to other variables or how it can be derived from a linear combination of other X variables. In this situation, the coefficient of correlation between the variable X_2 and the linear combination on the right side of Eq. (14.3) is bound to be unity.

Similarly, if $\lambda_2 \neq 0$ Eq. (14.2) can be written as

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \ldots - \frac{\lambda_k}{\lambda_2} X_{ki} - \frac{\lambda_k}{\lambda_2} v_i \qquad (14.4)$$

which shows that X_2 is not an exact linear combination of other $X's$ because it is also determined by the stochastic error term v_i.

The preceding algebraic approach to multicollinearity can be portrayed following by the figure. In this figure the circles Y, X_2 and X_3 represent, respectively, the variations in Y (the dependent variable) and X_2 and X_3 (the explanatory variables). The degree of collinearity can be measured by the extent of the overlap (shaded area) of the X_2 and X_3 circles. There is no overlap between X_2 and X_3, and hence no collinearity. There is a “low” to “high” degree of collinearity—the greater the overlap between X_2 and X_3 (i.e., the larger the shaded area), the higher the degree of collinearity. In the extreme, if X_2 and X_3 were to overlap completely (or if X_2 were completely inside X_3, or vice versa), collinearity would be perfect.

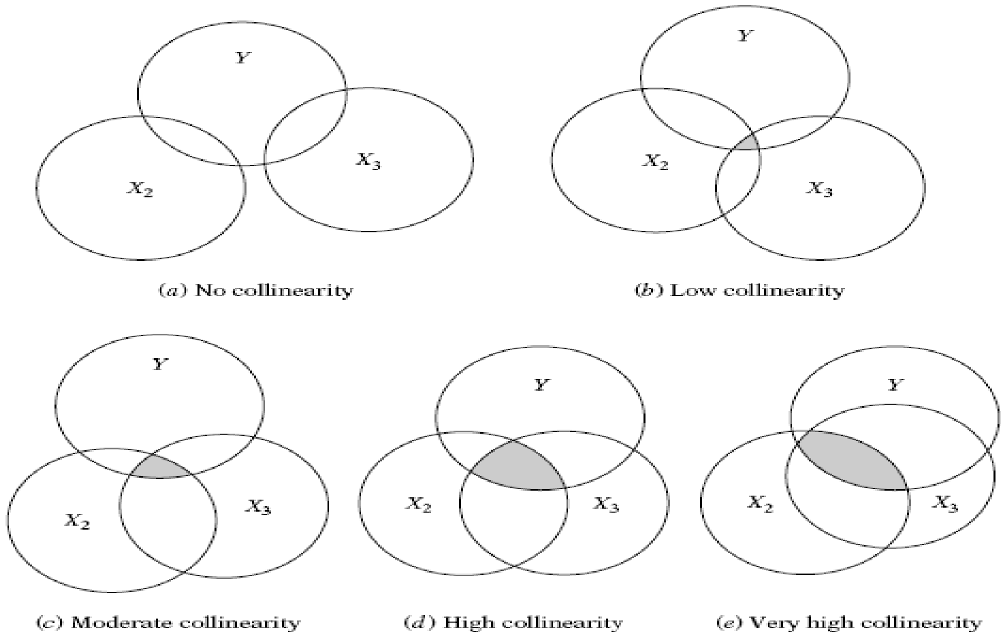


Figure 14.1

Why does the classical linear regression model assume that there is no multicollinearity among the $X's$? The reasoning is this: If multicollinearity is perfect in the sense of Eq. (14.1), the regression coefficients of the X variables are indeterminate and their standard errors are infinite. If multicollinearity is less than perfect, as in Eq. (14.2), the regression coefficients, although determinate, possess large standard errors (in relation to the coefficients themselves), which means the coefficients cannot be estimated with great precision or accuracy**.** The explanation of these statements is given in the following sections.

There are several sources of multicollinearity. As Montgomery and Peck note, multicollinearity may be due to the following factors:

1. The data collection method employed, for example, sampling over a limited range of the values taken by the regressors in the population.
2. Constraints on the model or in the population being sampled. For example, in the regression of electricity consumption (Y) on income (X_2) and house size (X_3) there is a physical constraint in the population in that families with higher incomes generally have larger homes than families with lower incomes.
3. Model specification, for example, adding polynomial terms to a regression model, especially when the range of the X variable is small.
4. An over defined model. This happens when the model has more explanatory variables than the number of observations. This could happen in medical research where there may be a small number of patients about whom information is collected on a large number of variables.

An additional reason for multicollinearity, especially in time series data, may be that the regressors included in the model share a *common trend,* that is, they all increase or decrease over time. Thus, in the regression of consumption expenditure on income, wealth, and population, the regressors income, wealth, and population may all be growing over time at more or less the same rate, leading to collinearity among these variables.

14.3 Consequences of Multicollinearity

Recall that if the assumptions of the classical model are satisfied, the OLS estimators of the regression estimators are BLUEs. It can be shown as explained below that even if multicollinearity is very high, as in the case of *near multicollinearity*, the OLS estimators still retain the property of BLUEs.

First, it is true that even in the case of near multicollinearity the OLS estimators are unbiased. But unbiasedness is a multisampling or repeated sampling property. What it means is that, keeping the values of the X variables fixed, if one obtains repeated samples and computes the OLS estimators for each of these samples, the average of the sample values will converge to the true population values of the estimators as the number of samples increases. But this says nothing about the properties of estimators in any given sample.

Second, it is also true that collinearity does not destroy the property of minimum variance: In the class of all linear unbiased estimators, the OLS estimators have minimum variance; that is, they are efficient. But this does not mean that the variance of an OLS estimator will necessarily be small.

Third, *multicollinearity is essentially a sample phenomenon* in the sense that even if the X variables are not linearly related in the population, they may be so related in the particular sample at hand: When we postulate the theoretical (population) regression function, we believe that all the X variables included in the model have a separate or independent influence on the dependent variable Y. But it may happen that in any given sample some or all of the X variables are so highly collinear that we cannot isolate their individual influence on Y. So although the theory says that all the $X's$ are important, our sample may not be “rich” enough to accommodate all X variables in the analysis.

As an illustration, consider the consumption–income example we know that, besides income, the wealth of the consumer is also an important determinant of consumption expenditure. Thus, we may write

$$\text{Consumption}_i = \beta_1 + \beta_2 \text{ Income}_i + \beta_3 \text{ Wealth}_i + u_i$$

Now it may happen that when we obtain data on income and wealth, the two variables may be highly, if not perfectly, correlated: Wealthier people generally tend to have higher incomes. Thus, although in theory income and wealth are independent variables to explain the behavior of consumption expenditure, in practice (i.e., in the sample) it may be difficult to disentangle the separate influences of income and wealth on consumption expenditure.

Ideally, to assess the individual effects of wealth and income on consumption expenditure we need a sufficient number of sample observations of wealthy individuals with low income, and high-income individuals with low wealth. Although this may be possible in cross-sectional studies (by increasing the sample size), it is very difficult to achieve in aggregate time series work.

For all these reasons, the fact that the OLS estimators are BLUEs despite multicollinearity is of little consolation in practice. We must see what happens or is likely to happen in any given sample as explained below.

The presence of multicollinearity has a number of potentially serious effects on the OLS estimators of the regression coefficients. In cases of near or high multicollinearity, one is likely to encounter the following consequences:

1. Even though, the OLS estimators are BLUEs, they have large variances and covariances, making precise estimation difficult as demonstrated below.

Suppose there are only two explanatory (independent) variables and let us suppose the variables $y,\ x_2$ and x_3 are in deviation form then the model is

$$y_i = \beta_2 x_{2i} + \beta_2 x_{3i} + u_i \qquad i = 1, 2, \ldots, n \tag{14.5}$$

with $E(u_i) = 0$

$E(u_i u_j) = 0$

$\operatorname{var}(u_i) = \sigma^2$

This model in compact form is

$$\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}} \tag{14.6}$$

where

$$\underset{\sim}{\mathbf{y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix},\ \mathbf{X} = \begin{bmatrix} x_{21} & x_{21} \\ x_{22} & x_{32} \\ \vdots & \vdots \\ x_{2n} & x_{3n} \end{bmatrix}, \underset{\sim}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix}, \underset{\sim}{\mathbf{u}} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

We have the variance-covariance matrix of $\hat{\underset{\sim}{\boldsymbol{\beta}}}$, the OLS estimator of $\underset{\sim}{\boldsymbol{\beta}}$, is

$$\operatorname{var}\left(\hat{\underset{\sim}{\boldsymbol{\beta}}}\right) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}, \qquad \text{where } \mathbf{X}'\mathbf{X} = \begin{bmatrix} \sum x_2^2 & \sum x_2 x_3 \\ \sum x_2 x_3 & \sum x_3^2 \end{bmatrix} \tag{14.7}$$

That is
$$\operatorname{var}\left(\hat{\underset{\sim}{\boldsymbol{\beta}}}\right) = \frac{\sigma^2}{\left(\sum x_2^2\right)\left(\sum x_3^2\right) - \left(\sum x_2 x_3\right)^2} \begin{bmatrix} \sum x_3^2 & -\sum x_2 x_3 \\ -\sum x_2 x_3 & \sum x_2^2 \end{bmatrix} \tag{14.8}$$

Now, from Eq. (14.8), we have

$$\text{var}\left(\hat{\beta}_2\right) = \sigma^2 \frac{\sum x_3^2}{\left(\sum x_2^2\right)\left(\sum x_3^2\right) - \left(\sum x_2 x_3\right)^2}$$

$$= \frac{\sigma^2}{\left(\sum x_2^2\right)\left(1 - \frac{\left(\sum x_2 x_3\right)^2}{\left(\sum x_2^2\right)\left(\sum x_3^2\right)}\right)}$$

$$= \frac{\sigma^2}{\left(\sum x_2^2\right)\left(1 - r_{23}^2\right)} \qquad (14.9)$$

where $r_{23} = \frac{\sum x_2 x_3}{\sqrt{\sum x_2^2}\sqrt{\sum x_3^2}}$ is the simple correlation between x_2 and x_3.

Similarly, $$\text{var}\left(\hat{\beta}_3\right) = \frac{\sigma^2}{\left(\sum x_3^2\right)\left(1 - r_{23}^2\right)} \qquad (14.10)$$

$$\text{cov}\left(\hat{\beta}_2, \hat{\beta}_3\right) = -\frac{\sigma^2\ r_{23}^2}{\left(\sum x_2 x_3\right)\left(1 - r_{23}^2\right)} \qquad (14.11)$$

If there is strong multicollinearity between x_2 and x_3, then the correlation r_{23} will be large and approaches to unity. As a consequence from equations (14.9), (14.10) and (14.11), we see that

$$r_{12}^2 \to 1 \ \Rightarrow\ \text{var}\left(\hat{\beta}_2\right) \to \infty,\ \text{var}\left(\hat{\beta}_3\right) \to \infty, \text{ and } \text{cov}\left(\hat{\beta}_2, \hat{\beta}_3\right) \to \pm\infty$$

Therefore, strong multicollinearity between x_2 and x_3 results in **large variances and covariance** of the OLS estimator $\hat{\beta}_2$ and $\hat{\beta}_3$.

When there are k-1(more than two) explanatory variables, multicollinearity produces similar effect. In this case

$$\text{var}\left(\hat{\beta}_j\right) = \frac{\sigma^2}{\sum x_j^2\left(1 - R_j^2\right)}, \qquad \text{j=2,3,...,k} \qquad (14.12)$$

where R_j^2 is the coefficient of determination from the regression of X_j on the reaming k-2 explanatory variables.

From Eq. (14.12), we may observe that

$$R_j^2 \to 1 \ \Rightarrow\ \text{var}\left(\hat{\beta}_j\right) \to \infty$$

Thus multicollinearity among the explanatory variables produce larger variances and covariances of the OLS estimators.

2. **Wider Confidence Intervals:** Because of consequence 1, the confidence intervals of the regression coefficients tend to be much wider.

3. **"Insignificant" *t* Ratios:** Recall that to test the null hypothesis $H_0: \ \beta_i = 0$, we use the *t* ratio $\dfrac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$, and compare the estimated *t* value with the critical *t* value from the student *t*-table. But, as we have seen, in case of high collinearity, the estimated standard errors $SE(\hat{\beta}_i)$ increase dramatically, thereby making the *t* values smaller. Therefore, in such cases, one will increasingly accept the null hypothesis $H_0: \ \beta_i = 0$ that the relevant true population value is zero. Hence, the probability of accepting a false hypothesis (i.e., type II error) increases. Thus, the t ratio of one or more regression coefficients tend to the statistically insignificant.

4. **A High R^2 but Few Significant *t* Ratios:** Although the t ratio of one or more coefficients is statistically insignificant R^2, the overall measure of goodness of fit, can be very high.
   Consider the *k*-variable linear regression model:
   $$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i \tag{14.13}$$
   In cases of high collinearity, it is possible to find, as we have just noted, that one or more of the partial slope coefficients are individually statistically insignificant on the basis of the *t* test. Yet the R^2 in such situations may be so high, say, in excess of 0.9, that on the basis of the F test one can convincingly reject the hypothesis that $\beta_2 = \beta_3 = \cdots = \beta_k = 0$. Indeed, this is one of the signals of multicollinearity—insignificant *t* values but a high overall R^2 (and a significant F value).

5. The OLS estimators and their standard errors can be sensitive to small changes in the data.

14.4 Detection of Multicollinearity

Having studied the nature and consequences of multicollinearity, the natural question is: How does one know that collinearity is present in any given situation, especially in models involving more than two explanatory variables? Here it is useful to bear in mind the following limits.

1. Multicollinearity is a question of degree and not of kind. The meaningful distinction is not between the presence and the absence of multicollinearity, but between its various degrees.
2. Since multicollinearity refers to the condition of the explanatory variables that are assumed to be non-stochastic, it is a feature of the sample and not of the population.

Therefore, we do not "test for multicollinearity" but can, if we wish, measure its degree in any particular sample.

Since multicollinearity is essentially a sample phenomenon, arising out of the largely non-experimental data collected in most social sciences, we do not have one unique method of detecting it or measuring its strength. What we have are some rules of thumb, some informal and some formal, but rules of thumb all the same. We now consider some of these rules.

1. **High R^2 but few significant *t* ratios:** As noted, this is the "classic" symptom of multicollinearity. If R^2 is high, say, in excess of 0.8, the F test in most cases will reject the hypothesis that the partial slope coefficients are simultaneously equal to zero, but the individual t tests will show that none or very few of the partial slope coefficients are statistically different from zero.

2. **High pair-wise correlations among regressors:** Another suggested rule of thumb is that if the pair-wise or zero-order correlation coefficient between two regressors is high, say, in excess of 0.8, then multicollinearity is a serious problem. The problem with this criterion is that, although high zero-order correlations may suggest collinearity, it is not necessary that they be high to have collinearity in any specific case. To put the matter somewhat technically, *high zero-order correlations are a sufficient but not a necessary condition for the existence of multicollinearity because it can exist even though the zero-order or simple correlations are comparatively low* (say, less than 0.50).
   Therefore, in models involving more than two explanatory variables, the simple or zero-order correlation will not provide reliable guidance to the presence of multicollinearity. Of course, if there are only two explanatory variables, the zero-order correlations will suffice.

3. **Variance Inflation factor (VIF):**
   In the multiple regression model

$$\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}} \text{ with } E\left(\underset{\sim}{\mathbf{u}}\right) = 0, \ \operatorname{var}\left(\underset{\sim}{\mathbf{u}}\right) = \sigma^2 \mathbf{I_n} \qquad (14.14)$$

we have $\hat{\underset{\sim}{\boldsymbol{\beta}}} = \left(\mathbf{X'X}\right)^{-1}\mathbf{X'}\underset{\sim}{\mathbf{y}}$ is the OLS estimator of $\underset{\sim}{\boldsymbol{\beta}}$.

Now, the variance of the ith component of $\hat{\underset{\sim}{\boldsymbol{\beta}}}$ is

$\operatorname{var}\left(\hat{\beta}_i\right) = \sigma^2 a_{ii}$, where a_{ii} is the i^{th} diagonal element of $\left(\mathbf{X'X}\right)^{-1}$

$$= \frac{\sigma^2}{\sum x_i^2}\frac{1}{\left(1-R_i^2\right)} \qquad (14.15)$$

where $\sum x_i^2 = \sum_{j=1}^{n}\left(X_{ij} - \bar{X}_i^2\right)$

$R_i^2 = \text{coefficient of determination of } \mathrm{X_i} \text{ on the remaining k-2}$
explnatory variables

But, Eq. (14.15) can be written as

$$\operatorname{var}\left(\hat{\beta}_i\right) = \frac{\sigma^2}{\sum x_i^2} VIF_i \qquad (14.16)$$

where $VIF_i = \dfrac{1}{\left(1-R_i^2\right)}$ (14.17)

But, we know that $0 \le R_i^2 \le 1$. Therefore, from Eq. (14.17), we may note that if R_i^2 increases toward unity, VIF_i also increases and in the limit it can be infinity (that is $R_i^2 \to 1 \Rightarrow VIF_i \to \infty$).

Some authors therefore use the VIF as an indicator of multicollinearity. The larger the value of VIF_i, the more “troublesome” or collinear the variable X_i. **As a rule of thumb,** if the VIF of a variable exceeds 10, which will happen if R_i^2 exceeds 0.90, that variable is said be highly collinear.

Note: VIF as a measure of collinearity is not free of criticism. As Eq. (14.16) shows, $\operatorname{var}\left(\hat{\beta}_i\right)$ depends on three factors: σ^2, $\sum X_i^2$, and VIF_i. A high VIF can be counterbalanced by a low σ^2 or a high $\sum X_i^2$. To put it differently, a high VIF is neither necessary nor sufficient to get high variances and high standard errors. Therefore, high multicollinearity, as measured by a high VIF, may not necessarily cause high standard errors. In all this discussion, the terms *high* and *low* are used in a relative sense.

4. Eigen values and condition index:

Since $\mathbf{X}'\mathbf{X}$ is a symmetric positive definite matrix, we know that all the eigen values of $\mathbf{X}'\mathbf{X}$ are real and positive. Let us denote the eigen values by $\lambda_1, \lambda_2, \ldots, \lambda_k$. Further, let us denote $\lambda_{\max}$ and $\lambda_{\min}$ as the maximum and minimum of $\lambda_1, \lambda_2, \ldots, \lambda_k$. Belsley, Kulh, and Welsch suggest a statistic, based on $\lambda_{\max}$ and $\lambda_{\min}$, called the **condition index** number of the $\mathbf{X}$ matrix, defined by

$$\boxed{CI(\mathbf{X}) = \frac{\sqrt{\text{Maximum eigenvalue}}}{\sqrt{\text{Minimum eigenvalue}}} = \frac{\sqrt{\lambda_{\max}}}{\sqrt{\lambda_{\min}}}} \tag{14.18}$$

Various applications with experimental and actual data sets suggest that **condition index** $CI(\mathbf{X})$ in the range of 20 to 30 are probably indicative of serious collinearity problems. Thus, **as a rule of thumb if the condition index is between 10 and 30**, there is moderate to strong multicollinearity and if it exceeds 30 there is severe multicollinearity. Some authors believe that the condition index is the best available multicollinearity diagnostic.

14.5 Remedial Measures

Rule-of-Thumb Procedures

One can try the following rules of thumb to address the problem of multicollinearity, the success depending on the severity of the collinearity problem.

1. **A priori information.**

   Suppose we consider the model

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \tag{14.19}$$

where Y = consumption, X_2 = income, and X_3 = wealth. As noted before, income and wealth variables tend to be highly collinear. But suppose a priori we believe that $\beta_3 = 0.10\beta_2$; that is, the rate of change of consumption with respect to wealth is one-tenth the corresponding rate with respect to income. We can then run the following regression:

$$
\begin{aligned}
Y_i &= \beta_1 + \beta_2 X_{2i} + 0.10\beta_3 X_{3i} + u_i \\
&= \beta_1 + \beta_2 X_{2i} + u_i
\end{aligned}
\qquad (14.20)
$$

where $X_i = X_{2i} + 0.1X_{3i}$. Once we obtain $\hat{\beta}_2$, we can estimate $\hat{\beta}_3$ from the postulated relationship between β_2 and β_3.

How does one obtain a priori information? It could come from previous empirical work in which the collinearity problem happens to be less serious or from the relevant theory underlying the field of study.

2. Combining cross-sectional and time series data.

A variant of the extraneous or a priori information technique is the combination of cross-sectional and time-series data, known as *pooling the data.* Suppose we want to study the demand for automobiles in the United States and assume we have time series data on the number of cars sold, average price of the car, and consumer income. Suppose also that

$$\log Y_t = \beta_1 + \beta_2 \log P_t + \beta_3 \log I_t + u_t \qquad (14.21)$$

where Y = number of cars sold, P = average price, I = income, and t = time. Our objective is to estimate the price elasticity β_2 and income elasticity β_3.

In time series data the price and income variables generally tend to be highly collinear. Therefore, if we run the preceding regression, we shall be faced with the usual multicollinearity problem. A way out of this has been suggested by Tobin. He says that if we have cross-sectional data (for example, data generated by consumer panels, or budget studies conducted by various private and governmental agencies), we can obtain a fairly reliable estimate of the income elasticity β_3 because in such data, which are at a point in time, the prices do not vary much. Let the cross-sectionally estimated income elasticity be $\hat{\beta}_3$. Using this estimate, we may write the preceding time series regression as

$$Y_t^* = \beta_1 + \beta_2 \log P_t + u_t \qquad (14.22)$$

where $Y_t^* = \ln Y_t - \hat{\beta}_3 \ln I_t$, that is, Y^* represents that value of Y after removing from it the effect of income. We can now obtain an estimate of the price elasticity β_2 from the preceding regression.

Although it is an appealing technique, pooling the time series and cross-sectional data in the manner just suggested may create problems of interpretation, because we are assuming implicitly that the cross-sectionally estimated income elasticity is the same thing as that which would be obtained from a pure time series analysis. Nonetheless, the technique has been used in many applications and is worthy of consideration in situations where the cross-sectional estimates do not vary substantially from one cross section to another.

3. Dropping a variable(s) and specification bias.
When faced with severe multicollinearity, one of the "simplest" things to do is to drop one of the collinear variables. But in dropping a variable from the model we may be committing a **specification bias** or **specification error.** Specification bias arises from incorrect specification of the model used in the analysis. Hence the remedy may be worse than the disease in some situations because, whereas multicollinearity may prevent precise estimation of the parameters of the model, omitting a variable may seriously mislead us as to the true values of the parameters. Recall that OLS estimators are BLUE despite nearcollinearity.

4. Transformation of variables.
Suppose we have time series data on consumption expenditure, income, and wealth. One reason for high multicollinearity between income and wealth in such data is that over time both the variables tend to move in the same direction. One way of minimizing this dependence is to proceed as follows.
If the relation

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \tag{14.23}$$

holds at time *t*, it must also hold at time *t* − 1 because the origin of time is arbitrary anyway. Therefore, we have

$$Y_{t-1} = \beta_1 + \beta_2 X_{2,t-1} + \beta_3 X_{3,t-1} + u_{t-1} \tag{14.24}$$

If we subtract Eq. (14.24) from Eq. (14.23), we obtain

$$\mathrm{Y_t} - Y_{t-1} = \beta_2\left(X_{2t} - X_{2,t-1}\right) + \beta_3\left(X_{3t} - X_{3,t-1}\right) + v_t, \text{ where } v_t = u_t - u_{t-1} \tag{14.25}$$

Eq. (14.25) is known as the **first difference form** because we run the regression, not on the original variables, but on the differences of successive values of the variables.
The first difference regression model often reduces the severity of multicollinearity because, although the levels of X_2 and X_3 may be highly correlated, there is no a priori reason to believe that their differences will also be highly correlated.

Another commonly used transformation in practice is the **ratio transformation.** Consider the model:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \tag{14.26}$$

where Y is consumption expenditure in real dollars, X_2 is GDP, and X_3 is total population.

Since GDP and population grow over time, they are likely to be correlated. One "solution" to this problem is to express the model on a per capita basis, that is, by dividing Eq. (14.23) by X_3, to obtain:

$$\frac{Y_t}{X_{3t}} = \beta_1\left(\frac{1}{X_{3t}}\right) + \beta_2\left(\frac{X_{2t}}{X_{3t}}\right) + \beta_3 + \left(\frac{u_t}{X_{3t}}\right) \tag{14.27}$$

Such a transformation may reduce multicollinearity in the original variables.

But the first-difference or ratio transformations are not without problems. For instance, the error term v_t in Eq. (14.25) may not satisfy one of the assumptions of the classical linear regression model, namely, that the disturbances are serially uncorrelated. Therefore, the remedy may be worse than the disease.

Hence, one should be careful in using the first difference or ratio method of transforming the data to resolve the problem of multicollinearity.

5. Additional or new data.

Since multicollinearity is a sample feature, it is possible that in another sample involving the same variables collinearity may not be so serious as in the first sample. Sometimes simply increasing the size of the sample (if possible) may reduce the collinearity problem.

For example, in the three-variable model from Eq. (14.9), we can see that as the sample size increases, $\sum x_2^2$ will generally increase. Therefore, for any given r_{23}, the variance of $\hat{\beta}_2$ will decrease, thus decreasing the standard error, which will enable us to estimate β_2 more precisely. Obtaining additional or "better" data is not always that easy.

6. Ridge Regression

One of the solutions often suggested for the multicollinearity problem is to use what is known as ridge regression first introduced by Hoerl and Kennard. Simply stated, the idea is if $\mathbf{X}'\mathbf{X}$ is close to singularity, then add a constant λ to the variances of the explanatory variables or equivalently to the diagonal elements of $\mathbf{X}'\mathbf{X}$, before solving the normal equations. The simple ridge estimator is

$$\boxed{\hat{\underset{\sim}{\boldsymbol{\beta}}}_R = \left(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{y}}} \tag{14.28}$$

There are several interpretations of ridge estimator. One is to obtain the least squares estimator of $\underset{\sim}{\boldsymbol{\beta}}$ subject to the condition $\Sigma\beta_i^2 = \underset{\sim}{\boldsymbol{\beta}}'\underset{\sim}{\boldsymbol{\beta}} = c$. Therefore to yield the ridge estimator of $\underset{\sim}{\boldsymbol{\beta}}$, we minimize the quantity

$$\left(\mathbf{y} - \mathbf{X}\underset{\sim}{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\underset{\sim}{\boldsymbol{\beta}}\right) + \lambda\left(\underset{\sim}{\boldsymbol{\beta}}'\underset{\sim}{\boldsymbol{\beta}} - c\right) \qquad \text{(where } \lambda \text{ is the Lagrangian multiplier)}$$

Differentiating it with respect to $\underset{\sim}{\boldsymbol{\beta}}$, we get

$$-2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + 2\lambda\underset{\sim}{\boldsymbol{\beta}} = 0 \ \text{ or } \ \left(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}\right)\underset{\sim}{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

Solving this equation for $\underset{\sim}{\boldsymbol{\beta}}$ gives the ridge estimator $\hat{\underset{\sim}{\boldsymbol{\beta}}}_R$ given in Eq. (14.28).

Since, we have

$$\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}} \ \text{ with } \ E\left(\underset{\sim}{\mathbf{u}}\right) = 0, \ \operatorname{var}\left(\underset{\sim}{\mathbf{u}}\right) = \sigma^2\mathbf{I_n}$$

It follows directly from Eq. (14.28) that

$$\begin{aligned}\hat{\underset{\sim}{\boldsymbol{\beta}}}_R &= \left(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}'(\mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}}) \\ &= \left(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \left(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{u}}\end{aligned}$$

and the mean vector and variance-covariance matrix of $\hat{\underset{\sim}{\boldsymbol{\beta}}}_R$ are given by

$$E\left(\hat{\underset{\sim}{\boldsymbol{\beta}}}_R\right) = \left(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{X}\underset{\sim}{\boldsymbol{\beta}}$$

$$\operatorname{var}\left(\hat{\underset{\sim}{\boldsymbol{\beta}}}_R\right) = E\left\{\left[\hat{\underset{\sim}{\boldsymbol{\beta}}}_R - E\left(\hat{\underset{\sim}{\boldsymbol{\beta}}}_R\right)\right]\left[\hat{\underset{\sim}{\boldsymbol{\beta}}}_R - E\left(\hat{\underset{\sim}{\boldsymbol{\beta}}}_R\right)\right]'\right\}$$

$$=E\left\{\left[\left(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{u}}\right]\left[\left(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{u}}\right]'\right\}$$

$$=E\left\{\left[\left(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{u}}\right]\left[\underset{\sim}{\mathbf{u}}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\right]\right\}$$

$$= \sigma^2\left(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}\right)^{-1}$$

The ridge estimator $\hat{\underset{\sim}{\boldsymbol{\beta}}}_R$ is thus biased, but it may be shown that the variances of the elements of $\hat{\underset{\sim}{\boldsymbol{\beta}}}_R$ are less than those of the OLS estimator. This raises the possibility that a ridge estimator may have a smaller mean-square error (MSE) than the OLS estimator. Hoerl and Kennard show that there always exists a constant $\lambda > 0$ such that

$$\sum_{i=1}^{k} MSE\left(\tilde{\beta}_i\right) < \sum_{i=1}^{k} MSE\left(\hat{\beta}_i\right)$$

where $\tilde{\beta}_i$ are the estimators of β_i from the ridge regression and $\hat{\beta}_i$ are the least squares estimators and k is the number of regressors. The main difficulty centre on the selection of a numerical value for the arbitrary scalar λ. Unfortunately, λ is a function of the regression parameters β_i and error variance σ^2, which are unknown . Hoerl and Kennard suggest trying different values of λ and picking the value of λ so that “the system will stabilize” or the “coefficients do not have unreasonable values.” Thus subjective arguments are used. Some others have suggested obtaining initial estimates of β_i and σ^2 and then using the estimated λ. This procedure can be iterated and we get the iterated ridge estimator.

The ridge technique essentially consists of an arbitrary numerical adjustment to the sample data, and one does not really know how to interpret the resultant estimators.

One other problem about ridge regression is the fact that t is not invariant to units of measurement of the explanatory variables and to linear transformations of variables. If we have two explanatory variables X_1 and X_2; and we measure X_1 in tens and X_2 in thousands, it does not make sense to add the same value of λ to the variances of both. This problem can be avoided by normalizing deviated variable by dividing it by its standard deviation. Even if X_1 and X_2 are measured in the same units, in some cases there are different linear transformation of X_1 and X_2 that are equally sensible.

7. Principal Components Regression

In the multiple regression equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u \tag{14.29}$$

if the explanatory variables $X_1, X_2, \ldots, X_k$ are highly collinear that is in the case of multicollinearity problem as a remedial measure we can use principal component regression analysis as explained below.

In case of multicollinearity problem it is advantageous to disregard some of the variables in order to reduce the problem. An alternative way to reduce the dimensionality is to use principal components. In the general principal component analysis the principal components with largest variances are used in order to explain as much of the total variation of the data on $X_1, X_2, \ldots, X_k$ where as in the context of multiple regression, it is sensible to take those principal components having the largest correlations with the dependent variable because the purpose in a regression is to explain the dependent variable. Thus, here we have to include the principal components in the regression analysis, according to the magnitude of their correlations with the dependent variable. In other words, the principal component with highest correlation with Y should be included first, the principal component with next highest correlation with Y should be included next and so on.

If the principal components have a natural intuitive meaning (i.e., a good interpretation). It is better to leave regression equation expressed in terms of the principal components. Otherwise, it is more convenient to transform back to the original variables.
The regression equation (14.29) in deviation form is

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \left(u - \overline{u}\right) \tag{14.30}$$

Eq. (14.30) can be written as

$$y = \underset{\sim}{\mathbf{x}}'\underset{\sim}{\boldsymbol{\beta}} + \varepsilon \qquad \text{where } \varepsilon = u - \overline{u} \tag{14.31}$$

$$\underset{\sim}{\mathbf{x}} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}, \quad \underset{\sim}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

Let us denote the variance-covariance matrix of the variables $x_1, x_2, \ldots, x_k$ by $\mathbf{V}_{kxk}$. Since $\mathbf{V}$ is positive definite matrix all eigen roots of $\mathbf{V}$ are positive, and we denote them as $\lambda_1, \lambda_2, \ldots, \lambda_k$. Now, let us denote the eigen vectors of $\mathbf{V}$ corresponding to these eigen roots as $\underset{\sim}{\mathbf{w}}_1, \quad \underset{\sim}{\mathbf{w}}_2, \quad \cdots, \quad \underset{\sim}{\mathbf{w}}_k$ and if we denote

$$\boldsymbol{\Omega} = \left(\underset{\sim}{\mathbf{w}}_1 \quad \underset{\sim}{\mathbf{w}}_2 \quad \cdots \quad \underset{\sim}{\mathbf{w}}_k\right)_{kxk}$$

then we have the relation

$$\boldsymbol{\Omega}'\mathbf{V}\boldsymbol{\Omega} = diag(\lambda_1, \lambda_2, \ldots, \lambda_k), \quad \text{where } \boldsymbol{\Omega} \text{ is orthogonal matrix} \tag{14.32}$$

Now suppose $z_1, z_2, \ldots, z_k$ are the principal components (obtained from the original variables $x_1, x_2, \ldots, x_k$, which are in deviation form) , then by the definition of principal components we may write

$$z_i = w_{i1}x_1 + w_{i2}x_2 + \ldots + w_{ik}x_k = \underset{\sim}{\mathbf{w}}_i'\underset{\sim}{\mathbf{x}}, \qquad \text{i=1,2,\ldots,k}$$

The above $z_1, z_2, \ldots, z_k$ principal components can be written in a vector form as

$$\underset{\sim}{\mathbf{z}} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_k \end{bmatrix} = \begin{bmatrix} \underset{\sim}{\mathbf{w}}_1'\underset{\sim}{\mathbf{x}} \\ \underset{\sim}{\mathbf{w}}_2'\underset{\sim}{\mathbf{x}} \\ \vdots \\ \underset{\sim}{\mathbf{w}}_k'\underset{\sim}{\mathbf{x}} \end{bmatrix} = \mathbf{\Omega}'\underset{\sim}{\mathbf{x}} \qquad (14.33)$$

Since $\mathbf{\Omega}$ is an orthogonal matrix, we have $\mathbf{\Omega}'\mathbf{\Omega} = \mathbf{I}_k$.

Now, Eq. (14.33) may be rewritten as

$$\underset{\sim}{\mathbf{x}} = (\mathbf{\Omega}')^{-1}\underset{\sim}{\mathbf{z}} = (\mathbf{\Omega}^{-1})^{-1}\underset{\sim}{\mathbf{z}} = \mathbf{\Omega}\underset{\sim}{\mathbf{z}} \qquad (\because \mathbf{\Omega}' = \mathbf{\Omega}^{-1}) \qquad (14.34)$$

Using Eq. (14.34) in Eq. (14.31) we get

$$\begin{aligned} y &= \underset{\sim}{\mathbf{z}}'\mathbf{\Omega}'\underset{\sim}{\boldsymbol{\beta}} + \varepsilon \\ &= \underset{\sim}{\mathbf{z}}'\underset{\sim}{\boldsymbol{\alpha}} + \varepsilon \end{aligned} \qquad (14.35)$$

$$\text{where} \quad \underset{\sim}{\boldsymbol{\alpha}} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{bmatrix}_{k\times 1} = \mathbf{\Omega}'\underset{\sim}{\boldsymbol{\beta}} \qquad (14.36)$$

Eq. (14.35) can be written as

$$y = \alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 + \ldots + \alpha_k z_k + \varepsilon \qquad (14.37)$$

Now, the regression equation (14.37) is with the new set of k explanatory variables $z_1, z_2, \ldots, z_k$ (principal components), those can be obtained using Eq. (14.33) from the original *k* explanatory variables $x_1, x_2, \cdots, x_k$, which are in deviation form.

The difference between the regression Eq. (14.30) and the regression Eq. (14.37) is in the former regression equation the explanatory variables $x_1, x_2, \cdots, x_k$ are highly collinear where as in the later regression equation, the explanatory variables $z_1, z_2, \ldots, z_k$ are uncorrelated.

The estimates of $\alpha_1, \alpha_2, \ldots, \alpha_k$ may be obtained by applying OLS method to regression Eq. (14.37) in the usual way. Then, the OLS estimates of the original parameters $\beta_1, \beta_2, \ldots, \beta_k$ may be obtained from the following equation which is based on Eq. (14.36)

$$\hat{\underset{\sim}{\boldsymbol{\beta}}} = \mathbf{\Omega}\hat{\underset{\sim}{\boldsymbol{\alpha}}} \qquad (14.38)$$

The variance-covariance matrix $\mathbf{V}$ can be computed based on the sample observations made on the variables $x_1, x_2, \ldots, x_k$. Using this matrix $\mathbf{V}$, one can compute the orthogonal matrix $\mathbf{\Omega}$, as the matrix of the eigen vectors of $\mathbf{V}$ corresponding to the computed eigen roots $\lambda_1, \lambda_2, \ldots, \lambda_k$.

Notes:

1. In the equation (14.37), the OLS estimators $\hat{\alpha}_i$'s are unaltered if some of the principal components z_j's are deleted from the equation.
2. One can think of choosing only those principal components that have highest correlation with y and discard the rest, but the same sort of procedure can be used with the

original set of variables $x_1, x_2, \cdots, x_k$ by first choosing the variable with the highest correlation with y, then the one with the highest partial correlation, and so on. This is what "step wise regression program" do.

3. The linear combinations z's often do not have economic meaning. This is one of the most important drawbacks of the above method.
4. Changing the units of measurement of the x's will change the principal components. This problem can be avoided if all variables are standardized to have unit variance

14.6 Self Assessment Questions

1. Explain the problem of multicollinearity and explain the estimation procedure of the model in the presence of multicollinearity
2. Explain the problem of multicollinearity and their consequences.
3. What are the sources of multicollinearity.
4. Explain the problem of multicollinearity with a suitable example. Also discuss the implications and tools for handling this problem.
5. Describe a test procedure for detection of multicollinearity and suggest some remedial measures.
6. Describe variance inflation factor as a test for detection of multicollinearity.
7. Describe condition index based on eigen values as a test for detection of multicollinearity.
8. Explain the Ridge regression method in detail and give the reasons for the popularity of this method.
9. Explain the Ridge regression method in detail and also explain its importance.
10. Explain the principle component regression method in detail as a remedial measure.

14.7 References

1. Gujarati, D.N. (2005): *Basic Econometrics, 4th Ed., Tata McGraw-Hill.*
2. Johnston, J. (1984): *Econometric Methods, 3rd Ed., McGraw-Hill,* New York.
3. Montgomery, D.C., Peck, E.A. and Geoffrey Vining, G. (2003): *Introduction to Linear Regression Analysis, 3rd Ed., Wiley*
4. Draper, N.R., and H. Smith(1998): *Applied Regression Analysis, 3rd Ed., John Wiley & Sons,* New York.
5. G.S. Maddala (2001): *Introduction to Econometrics, 3rd Ed., John Wiley & Sons, Ltd.*
6. Johnston, J. and DiNardo J (1997): *Econometric Methods*, 4th Ed., McGraw Hill.
7. Hill, Carter, William Griffiths, and George Judge(2001): *Undergraduate Econometrics, John Willey & Sons,* New York.
8. *Koutsoyiannis, A(1973): Theory of Econometrics, Harper & Row,* New York.

Lesson 15

GENERALIZED LEAST SQUARES ESTIMATOR

15.0 Objective:

In the lessons studied so far, we consider the regression models with spherical disturbances, which means the disturbances are uncorrelated and having common variance. Now, in this lesson our objective is to study the regression model with nonspherical disturbances, which are serially (auto) correlated or (and) having heterogonous variances.

Structure of the Lesson:

15.1 Introduction

The assumptions usually made concerning the linear regression model $\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}}$ are

$$E\left(\underset{\sim}{\mathbf{u}}\right) = \underset{\sim}{\mathbf{0}} \tag{15.1}$$

$$and\ Var\left(\underset{\sim}{\mathbf{u}}\right) = \sigma^2\mathbf{I} \tag{15.2}$$

Eq. (15.2) is described as the assumption of ***spherical disturbances***. It involves the double assumption that the disturbances (error terms) are homoscedastic as well as non-autocorrelated (serially uncorrelated). Sometimes this assumption may not be fulfilled; and in place of the assumption (15.2) we have to make the following assumption

$$\boxed{Var\left(\underset{\sim}{\mathbf{u}}\right) = \sigma^2\boldsymbol{\Omega}} \tag{15.3}$$

The assumption (15.3) is described as ***nonspherical disturbances***, since it allows the heteroscedastic disturbances and autocorrelated disturbances.

In the following sections, we will discuss the following

1. The sources of nonspherical disturbances.
2. To study the properties of OLS estimators in the presence of nonspherical disturbances.
3. To develop appropriate estimation procedure for the general linear model fulfilling the assumption (15.3).

15.2 The Sources of Nonspherical Disturbances

If the sample observations relate to households or firms in a cross-section study, the assumption of a common disturbance variance at all observation points is often implausible. For example, if Y refers to family expenditure and X to family income, the variance about the Engle curve is likely to increase with the size of X. Similarly if Y denotes profits and X is some measure of firm size, the same property is to be expected. The specification of the disturbance variance-covariance matrix of disturbances would then be

$$Var\left(\underset{\sim}{\mathbf{u}}\right) = \mathbf{\Omega} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \cdots & \cdots & & \cdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

which is the standard case of nonspherical disturbances and it still assumes that the disturbances are pair wise uncorrelated.

Another possibility of nonspherical disturbances in cross-section studies will arise when we are dealing with grouped data. Suppose the model is

$$Y_t = \alpha + \beta X_t + u_t \qquad \text{t=1,2,. . ., n}$$

where the u_t's are homoscedastic with zero covariances. However, suppose we only have access to data which have been averaged within m groups, where the ith group contains n_i observations. The form of the appropriate model to the data is now

$$\overline{Y}_i = \alpha + \beta \overline{X}_i + \overline{u}_i \qquad i = 1,\ . . .,\ m$$

and clearly

$$\text{var}\left(\overline{u}_i\right) = \frac{\sigma^2}{n_i} \qquad \text{where } \sigma^2 = \text{var}\left(u_i\right) \qquad i = 1,\ . . .,\ m$$

Thus

$$Var\left(\underset{\sim}{\mathbf{u}}\right) = \sigma^2\mathbf{\Omega} = \sigma^2 \begin{bmatrix} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & \cdots & 0 \\ \cdots & \cdots & & \cdots \\ 0 & 0 & \cdots & \frac{1}{n_m} \end{bmatrix}$$

Another possibility of nonspherical disturbances will arise when we are dealing with regression models using time series data, which occur relatively often in economics, business, and some fields of engineering. The assumption of uncorrelated and independent disturbances in time series data is often not appropriate. Usually, the disturbances in time series data exhibit serial correlation, that is $\mathrm{cov}\left(u_i, u_{i+j}\right) \neq 0$, when $\mathrm{j} \neq 0$. Such disturbances are said to be autocorrelated, which is a special case of nonspherical disturbances. There are several sources of autocorrelation. Perhaps the primary cause of autocorrelation in regression problems involving time series data is failure to include one or more important regressors in the model.

15.3 Properties of OLS Estimators under Nonspherical Disturbances

Our assumed model is

$$\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}}$$

where $\mathbf{X}$ is taken as a nonstochastic matrix with full column rank,

$E\left(\underset{\sim}{\mathbf{u}}\right) = \underset{\sim}{\mathbf{0}}$ and $Var\left(\underset{\sim}{\mathbf{u}}\right) = \sigma^2\boldsymbol{\Omega}$ $\left(\text{or } \mathbf{V}\right)$

The OLS estimator of $\boldsymbol{\beta}$ may be expressed as usual as

$$\underset{\sim}{\hat{\boldsymbol{\beta}}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{y}} = \underset{\sim}{\boldsymbol{\beta}} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{u}}$$

Thus $E\left(\underset{\sim}{\hat{\boldsymbol{\beta}}}\right) = \underset{\sim}{\boldsymbol{\beta}}$ so that OLS estimator $\underset{\sim}{\hat{\boldsymbol{\beta}}}$ is still unbiased. The variance-covariance matrix of $\underset{\sim}{\hat{\boldsymbol{\beta}}}$ is given by

$$\begin{aligned} \mathrm{var}\left(\underset{\sim}{\hat{\boldsymbol{\beta}}}\right) &= E\left(\underset{\sim}{\hat{\boldsymbol{\beta}}} - \underset{\sim}{\boldsymbol{\beta}}\right)\left(\underset{\sim}{\hat{\boldsymbol{\beta}}} - \underset{\sim}{\boldsymbol{\beta}}\right)' = \mathrm{E}\left\{\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right\} \\ &= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \end{aligned} \qquad (15.4\ )$$

Thus the conventional formula $\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ no longer measures the variances of the OLS estimator, and any application of it is potentially misleading. More importantly, even if one could use Eq. (15.4) to estimate the sampling variances the substitution of these numbers in the conventional t formulas and confidence interval formulas is strictly invalid since the assumptions used in deriving those inference procedures no longer apply. For the same reason the optimal minimum variance property of OLS no longer holds.

Thus, even though the OLS estimator $\underset{\sim}{\hat{\boldsymbol{\beta}}}$ is unbiased estimator of $\underset{\sim}{\boldsymbol{\beta}}$, it is not the BLUE of $\underset{\sim}{\boldsymbol{\beta}}$. Hence, there is a need to the development of a more appropriate estimator.

15.4 The Generalized Least Squares (Aitken) Estimator

Our assumed model is

$$\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}} \qquad (15.\ 5)$$

where $\mathbf{X}$ is taken as a nonstochastic matrix with full column rank,

$E\left(\underset{\sim}{\mathbf{u}}\right) = \underset{\sim}{\mathbf{0}}$, and $Var\left(\underset{\sim}{\mathbf{u}}\right) = \sigma^2\boldsymbol{\Omega}$ $\left(\text{or } \mathbf{V}\right)$

Since, the nonspherical variance-covariance matrix $\mathbf{\Omega}$ $(\text{or } \mathbf{V})$ is symmetric positive definite matrix, there exists a nonsingular matrix $\mathbf{T}$ such that

$$\mathbf{T\Omega T'} = \mathbf{I} \tag{15.6}$$

Now pre-multiply Eq. (15.5) with this non-singular matrix $\mathbf{T}$ to obtain

$$\underset{\sim}{\mathbf{y}}^* = \mathbf{X}^*\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}}^* \tag{15.7}$$

where $\underset{\sim}{\mathbf{y}}^* = \mathbf{T}\underset{\sim}{\mathbf{y}}$, $\mathbf{X}^* = \mathbf{TX}$, and $\underset{\sim}{\mathbf{u}}^* = \mathbf{T}\underset{\sim}{\mathbf{u}}$ (15.8)

with i) $E\left(\underset{\sim}{\mathbf{u}}^*\right) = \underset{\sim}{\mathbf{0}}$,

ii) $E\left(\underset{\sim}{\mathbf{u}}^*\underset{\sim}{\mathbf{u}}^{*\prime}\right) = E\left(\mathbf{T}\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}'\mathbf{T}'\right) = \sigma^2\mathbf{T\Omega T'} = \sigma^2\mathbf{I}$ (from Eq. (15.6)) (15.9)

Thus the model as given in Eq. (15.5) with nonspherical disturbance variance-covariance matrix is transformed into the traditional general linear model with spherical disturbance variance-covariance matrix (as given in Eq. (15.2)). As a consequence, we can apply OLS method to the model (15.7) to obtain OLS estimator of $\boldsymbol{\beta}$ and is given by

$$\underset{\sim}{\mathbf{b}} = \left(\mathbf{X}^{*\prime}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*\prime}\underset{\sim}{\mathbf{y}}^* = \left(\mathbf{X'T'TX}\right)^{-1}\mathbf{X'T'T}\underset{\sim}{\mathbf{y}} \tag{15.10}$$

and $\operatorname{var}\left(\underset{\sim}{\mathbf{b}}\right) = \sigma^2\left(\mathbf{X}^{*\prime}\mathbf{X}^*\right)^{-1} = \sigma^2\left(\mathbf{X'T'TX}\right)^{-1}$ (15.11)

Since $\mathbf{T}$ is nonsingular matrix from Eq. (15.6)

$$\left(\mathbf{T}'\right)^{-1}\mathbf{\Omega}^{-1}\mathbf{T}^{-1} = \mathbf{I} \Rightarrow \mathbf{T'T} = \mathbf{\Omega}^{-1} \tag{15.12}$$

Using Eq. (15.12) in Eqs. (15.10) and (15.11) we get

$$\boxed{\begin{aligned} &\underset{\sim}{\mathbf{b}} = \left(\mathbf{X'\Omega^{-1}X}\right)^{-1}\mathbf{X'\Omega^{-1}}\underset{\sim}{\mathbf{y}} && (15.13)\\ \text{and}\quad &\operatorname{var}\left(\underset{\sim}{\mathbf{b}}\right) = \sigma^2\left(\mathbf{X'\Omega^{-1}X}\right)^{-1} && (15.14)\end{aligned}}$$

The estimator $\underset{\sim}{\mathbf{b}}$ given in Eq. (15.13) is called the generalized least squares (GLS) estimator or Aitken's estimator and the variance-covariance matrix of $\underset{\sim}{\mathbf{b}}$ is given in Eq. (15.14). Since the model (15.7) satisfies the assumptions given in Eq. (15.9) which required for an application of OLS, it immediately follows that $\underset{\sim}{\mathbf{b}}$ is the BLUE of $\underset{\sim}{\boldsymbol{\beta}}$ in the model (15.5).

It may be noted that the above formulae are only operational if the elements of $\mathbf{\Omega}$ are known.

Note: If we take $Var\left(\underset{\sim}{\mathbf{u}}\right) = \mathbf{V}$ in the model (15.5) then

$$\underset{\sim}{\mathbf{b}} = \left(\mathbf{X'V^{-1}X}\right)^{-1}\mathbf{X'V^{-1}}\underset{\sim}{\mathbf{y}} \text{ and } \operatorname{var}\left(\underset{\sim}{\mathbf{b}}\right) = \left(\mathbf{X'V^{-1}X}\right)^{-1}.$$

This procedure of transforming the original variables in such a way that the transformed variables satisfy the assumptions of the classical linear regression model and then applying OLS to them is known as the method of generalized least squares (GLS). *In short, GLS is OLS on the transformed variables that satisfy the standard least-squares assumptions.* The estimator thus obtained is known as **GLS estimator or Aitken Estimator,** and this estimator is BLUE.

15.5 Derivation of an unbiased estimator of σ^2

Applying OLS to the transformed model (15.7), we get an unbiased estimator of σ^2 given by

$$\hat{\sigma}^2 = \frac{\underset{\sim}{\mathbf{e}}^{*\prime}\underset{\sim}{\mathbf{e}}^{*}}{(n-k)} \tag{15.15}$$

where

$$\underset{\sim}{\mathbf{e}}^{*} = \underset{\sim}{\mathbf{y}}^{*} - \mathbf{X}^{*}\underset{\sim}{\mathbf{b}} = \mathbf{T}\underset{\sim}{\mathbf{y}} - \mathbf{TX}\underset{\sim}{\mathbf{b}} = \mathbf{T}\left(\underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\mathbf{b}}\right)$$

Then

$$\hat{\sigma}^2 = \frac{\left(\underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\mathbf{b}}\right)'\mathbf{T}'\mathbf{T}\left(\underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\mathbf{b}}\right)}{(n-k)}$$

$$= \frac{\left(\underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\mathbf{b}}\right)'\mathbf{\Omega}^{-1}\left(\underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\mathbf{b}}\right)}{(n-k)} \qquad \left(\text{using Eq. (15.12)}\right)$$

$$= \frac{\underset{\sim}{\mathbf{e}}'\mathbf{\Omega}^{-1}\underset{\sim}{\mathbf{e}}}{(n-k)}, \quad \text{where} \quad \underset{\sim}{\mathbf{e}} = \underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\mathbf{b}}$$

$$= \frac{\underset{\sim}{\mathbf{y}}'\mathbf{\Omega}^{-1}\underset{\sim}{\mathbf{y}} - 2\underset{\sim}{\mathbf{b}}'\mathbf{X}'\mathbf{\Omega}^{-1}\underset{\sim}{\mathbf{y}} + \underset{\sim}{\mathbf{b}}'\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\underset{\sim}{\mathbf{b}}}{(n-k)}$$

$$= \frac{\underset{\sim}{\mathbf{y}}'\mathbf{\Omega}^{-1}\underset{\sim}{\mathbf{y}} - \underset{\sim}{\mathbf{b}}'\mathbf{X}'\mathbf{\Omega}^{-1}\underset{\sim}{\mathbf{y}}}{(n-k)} \quad (\text{since from Eq.(15.13) } \mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\underset{\sim}{\mathbf{b}} = \mathbf{X}'\mathbf{\Omega}^{-1}\underset{\sim}{\mathbf{y}}) \tag{15.16}$$

is an unbiased estimator of σ^2

15.6 To show that GLS (Aitken) estimator is also ML estimator

Let us consider the model

$$\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}} \tag{15.17}$$

with

1. $\mathbf{X}$ is a nonstochastic and full rank matrix
2. $E\left(\underset{\sim}{\mathbf{u}}\right) = \underset{\sim}{\mathbf{0}}$
3. $Var\left(\underset{\sim}{\mathbf{u}}\right) = \sigma^2\mathbf{\Omega}$
4. $\underset{\sim}{\mathbf{u}}$ is normally distributed

Since $\underset{\sim}{\mathbf{u}} \sim N\left(\underset{\sim}{\mathbf{0}}, \sigma^2\mathbf{\Omega}\right)$, the likelihood function is

$$p\left(\underset{\sim}{\mathbf{u}}\right) = \frac{1}{(2\pi)^{n/2}\left|\sigma^2\mathbf{\Omega}\right|^{1/2}} e^{-\frac{1}{2\sigma^2}\underset{\sim}{\mathbf{u}}'\mathbf{\Omega}^{-1}\underset{\sim}{\mathbf{u}}}$$

The likelihood in terms $\underset{\sim}{\mathbf{y}}$ is

$$L = p\left(\underset{\sim}{\mathbf{y}}/\mathbf{X}\right) = \frac{1}{\left(\sigma^2 2\pi\right)^{n/2} |\mathbf{\Omega}|^{1/2}} e^{-\frac{1}{2\sigma^2}\left(\underset{\sim}{\mathbf{y}}-\mathbf{X}\underset{\sim}{\boldsymbol{\beta}}\right)' \mathbf{\Omega}^{-1}\left(\underset{\sim}{\mathbf{y}}-\mathbf{X}\underset{\sim}{\boldsymbol{\beta}}\right)} J\left(\underset{\sim}{\mathbf{y}}\right)$$

where Jacobian transformation $J\left(\underset{\sim}{\mathbf{y}}\right) = \text{mod}\left|\frac{\partial \underset{\sim}{u}}{\partial \underset{\sim}{\mathbf{y}}}\right| = \text{mod}\left|\mathbf{I_n}\right| = 1$

Now, the log likelihood is

$$\log L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2}\log|\mathbf{\Omega}| - \frac{1}{2\sigma^2}\left(\underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\boldsymbol{\beta}}\right)' \mathbf{\Omega}^{-1}\left(\underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\boldsymbol{\beta}}\right) \qquad (15.18)$$

Maximizing $\log L$ with respect to $\underset{\sim}{\boldsymbol{\beta}}$ implies minimizing the weighted sum of squares

$$\left(\underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\boldsymbol{\beta}}\right)' \mathbf{\Omega}^{-1}\left(\underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\boldsymbol{\beta}}\right) = \underset{\sim}{\mathbf{y}}'\mathbf{\Omega}^{-1}\underset{\sim}{\mathbf{y}} - 2\underset{\sim}{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{\Omega}^{-1}\underset{\sim}{\mathbf{y}} + \underset{\sim}{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\underset{\sim}{\boldsymbol{\beta}} \qquad (15.19)$$

with respect to $\underset{\sim}{\boldsymbol{\beta}}$. This is equivalent to differentiate Eq. (15.19) with respect to $\underset{\sim}{\boldsymbol{\beta}}$ and setting equal to zero and is

$$\frac{\partial\left(\underset{\sim}{\mathbf{y}}'\mathbf{\Omega}^{-1}\underset{\sim}{\mathbf{y}} - 2\underset{\sim}{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{\Omega}^{-1}\underset{\sim}{\mathbf{y}} + \underset{\sim}{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\underset{\sim}{\boldsymbol{\beta}}\right)}{\partial \underset{\sim}{\boldsymbol{\beta}}} = 0$$

$$\Rightarrow \quad -2\mathbf{X}'\mathbf{\Omega}^{-1}\underset{\sim}{\mathbf{y}} + 2\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\underset{\sim}{\boldsymbol{\beta}} = 0$$

$$\Rightarrow \quad \underset{\sim}{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\underset{\sim}{\mathbf{y}}$$

$$\Rightarrow \quad \underset{\sim}{\boldsymbol{\beta}}* = \left(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\underset{\sim}{\mathbf{y}} \qquad (15.20)$$

Thus $\underset{\sim}{\boldsymbol{\beta}}*$ the MLE of $\underset{\sim}{\boldsymbol{\beta}}$ is same as $\underset{\sim}{\mathbf{b}}$, the GLS estimator $\underset{\sim}{\boldsymbol{\beta}}$.

On the assumption of normality for the disturbance term all the inference procedures carry through for this model. Thus the test of

$$H_0 : \mathbf{R\boldsymbol{\beta}} = r$$

is based on

$$F = \frac{\left(r - \mathbf{R}\underset{\sim}{\mathbf{b}}\right)'\left[\mathbf{R}\left(\mathbf{X}'\Omega\mathbf{X}\right)^{-1}\mathbf{R}'\right]^{-1}\left(r - \mathbf{R}\underset{\sim}{\mathbf{b}}\right)'/q}{\hat{\sigma}^2} \qquad (15.21)$$

where $\hat{\sigma}^2$ is as given in Eq. (15.16), having the $F(q, n-k)$ distribution under the null hypothesis, where $\underset{\sim}{\mathbf{b}}$ is the GLS estimator defined in Eq. (15.13).

The above formulae are only operational if the elements of Ω are known. In some exceptional cases this may be so, but in most practical cases it is not. We must therefore proceed to the development of operational procedures for such cases, but there is, in fact, no single procedure which is generally applicable. One must look for the procedure which is best suited to the features of each specific problem in turn.

15.7 Self Assessment Questions

1. Explain GLS method of estimation for GLM model.
2. Explain the generalized least squares (GLS) estimates and discuss the method of obtaining them. Also list out their properties.
3. Derive Aitken estimators of a general linear model.
4. Show that GLS estimator is BLUE.
5. Show that Aitken estimator is BLUE.
6. State and prove Aitken theorem for a generalized linear model $\underset{n X 1}{\mathbf{y}} = \underset{n X k}{\mathbf{X}} \underset{k X 1}{\boldsymbol{\beta}} + \underset{n X 1}{\mathbf{u}}$ with $E(\mathbf{u}) = 0$ and $E(\mathbf{uu}') = \sigma^2 \mathbf{\Omega}$.

15.8 References

1. Gujarati, D.N. (2005): *Basic Econometrics, 4th Ed., Tata McGraw-Hill.*

2. Johnston, J. (1984): *Econometric Methods, 3rd Ed., McGraw-Hill,* New York.

3. Montgomery, D.C., Peck, E.A. and Geoffrey Vining, G. (2003): *Introduction to Linear Regression Analysis, 3rd Ed., Wiley*

4. Draper, N.R., and H. Smith(1998): *Applied Regression Analysis, 3rd Ed., John Wiley & Sons,* New York.

5. G.S. Maddala (2001): *Introduction to Econometrics, 3rd Ed., John Wiley & Sons, Ltd.*

6. Johnston, J. and DiNardo J (1997): *Econometric Methods*, 4th Ed., McGraw Hill.

7. Hill, Carter, William Griffiths, and George Judge(2001): *Undergraduate Econometrics, John Willey & Sons,* New York.

8. Koutsoyiannis, A(1973): Theory of Econometrics, Harper & Row, New York.

Lesson 16

HETEROSCEDASTICITY: NATURE AND CONSEQUENCES

16.0 Objective:

One of the assumptions made in traditional multiple linear regression model regarding the disturbances is that they have common variance (homoscedasticity). This lesson relaxes this assumption, which means the disturbances are having heterogeneous variances and such disturbances are called heteroscedastic disturbances. The objective of this lesson is to discuss the nature, sources and consequences of heteroscedastic disturbances.

Structure of the Lesson:

16.1. Introduction

An important assumption of the traditional multiple linear regression model is that the variance of each disturbance term u_i, conditional on the chosen values of the explanatory variables, is some constant number equal to σ_u^2. This is the assumption of **homoscedasticity,** or *equal* (homo) *spread* (scedasticity), that is, *equal variance*. Symbolically,

$$Var\left(u_i\right) = \sigma_u^2 \qquad i = 1, 2, \ldots, n$$

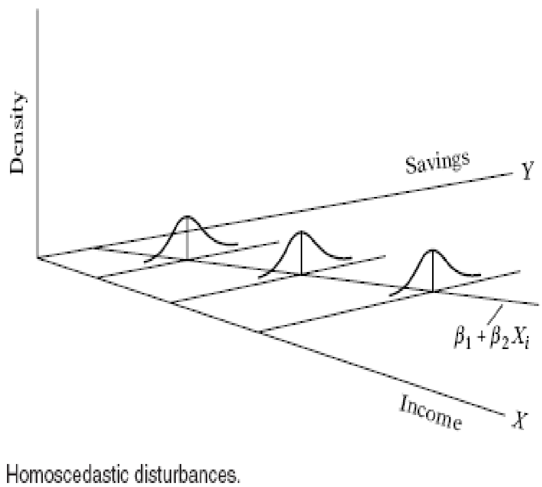
If we relax the assumption of **homoscedasticity** that is if the disturbance terms $u_i s$ do not have the equal variance, then we say that disturbances are ***heteroscedastic disturbances*** and in this case the disturbance terms $u_i s$ have unequal or heterogeneous variances. The multiple linear regression model with *heteroscedastic disturbances* is described as heteroscedasticity**,** which may be written symbolically,

$$Var\left(u_i\right) = \sigma_i^2 \qquad i = 1, 2, \ldots, n \qquad (16.1)$$

Notice the subscript of σ^2, which reminds us that the conditional variances of u_i (or equivalently conditional variances of Y_i) are no longer constant. The nature, sources and consequences of *heteroscedasticity* are studied in the following sections of this lesson.

16.2. The Nature or Sources of Heteroscedasticity

To make the difference between homoscedasticity and heteroscedasticity clear, assume that in the two-variable model
$Y_i = \beta_1 + \beta_2 X_i + u_i$, Y represents savings and X represents income.



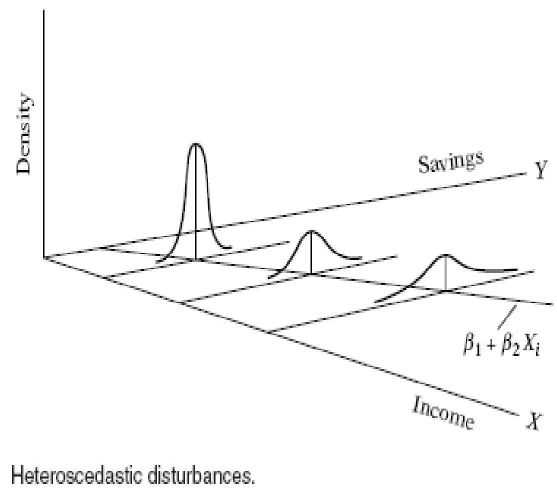


Figure 16.1 Figure 16.2

The above figures 16.1 and 16.2 show that as income increases, savings on the average also increase. But in Figure 11.1 the variance of savings remains the same at all levels of income, whereas in Figure 16.2 it increases with income. It seems that in Figure 16.2 the higher income families on the average save more than the lower-income families, but there is also more variability in their savings.

There are several reasons why the variances of u_i may be variable, some of which are as follows.

1. Following the *error-learning models*, as people learn, their errors of behavior become smaller over time. In this case, σ_i^2 is expected to decrease. As an example, consider Figure 16.3, which relates the number of typing errors made in a given time period on a test to the hours put in typing practice. As Figure 16.3 shows, as the number of hours of typing practice increases, the average number of typing errors as well as their variances decreases.

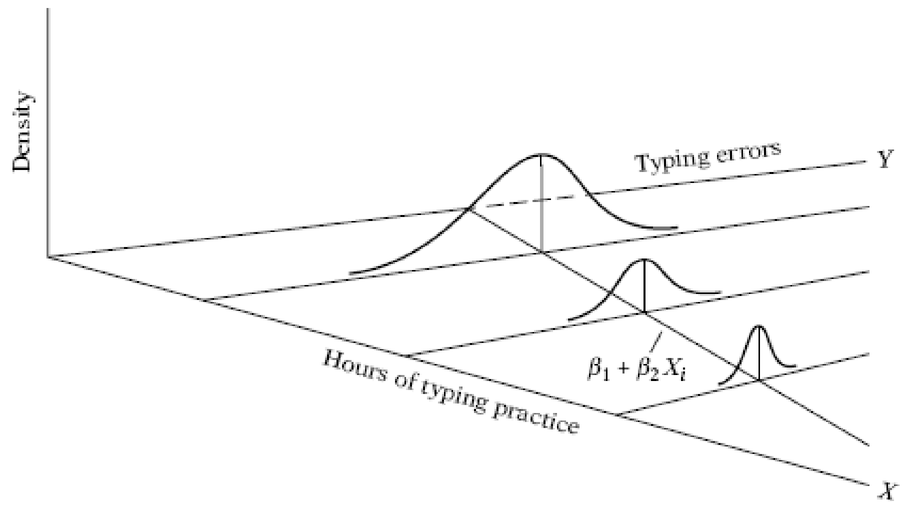


Illustration of heteroscedasticity.

Figure 16.3

2. As incomes grow, people have more *discretionary income* and hence more scope for choice about the disposition of their income. Hence, σ_i^2 is likely to increase with income. Thus in the regression of savings on income one is likely to find σ_i^2 increasing with income (as in Figure 16.2) because people have more choices about their savings behavior. Similarly, companies with larger profits are generally expected to show greater variability in their dividend policies than companies with lower profits. Also, growth oriented companies are likely to show more variability in their dividend payout ratio than established companies.

3. As data collecting techniques improve, σ_i^2 is likely to decrease. Thus, banks that have sophisticated data processing equipment are likely to commit fewer errors in the monthly or quarterly statements of their customers than banks without such facilities.

4. Heteroscedasticity can also arise as a result of the presence of *outliers***.** An outlying observation, or outlier, is an observation that is much different (either very small or very large) in relation to the observations in the sample. More precisely, an outlier is an observation from a different population to that generating the remaining sample observations. The inclusion or exclusion of such an observation, especially if the sample size is small, can substantially alter the results of regression analysis.

5. Another source of heteroscedasticity arises when there are specification errors in the regression model. Very often, heteroscedasticity may be due to the fact that some important variables are omitted from the model. Thus, in the demand function for a commodity, if we do not include the prices of commodities complementary to or competing with the commodity in question (the omitted variable bias), the residuals obtained from the regression may give the distinct impression that the error variance

may not be constant. But if the omitted variables are included in the model, that impression may disappear.

6. Another source of heteroscedasticity is skewness in the distribution of one or more regressors included in the model. Examples are economic variables such as income, wealth, and education. It is well known that the distribution of income and wealth in most societies is uneven, with the bulk of the income and wealth being owned by a few at the top.

7. Other sources of heteroscedasticity: (1) incorrect data transformation (e.g., ratio or first difference transformations) and (2) incorrect functional form (e.g., linear versus log–linear models).

Note that the problem of heteroscedasticity is likely to be more common in cross-sectional than in time series data. In cross-sectional data, one usually deals with members of a population at a given point in time, such as individual consumers or their families, firms, industries, or geographical subdivisions such as state, country, city, etc. Moreover, these members may be of different sizes, such as small, medium, or large firms or low, medium, or high income. In time series data, on the other hand, the variables tend to be of similar orders of magnitude because one generally collects the data for the same entity over a period of time. Examples are GNP, consumption expenditure, savings, or employment in India.

16.3. OLS Estimation in the Presence of Heteroscedasticity

In case of heteroscedasticity the assumed model is

$$\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}}, \tag{16.2}$$

with $E\left(\underset{\sim}{\mathbf{u}}\right) = \underset{\sim}{\mathbf{0}}$ and $Var\left(\underset{\sim}{\mathbf{u}}\right) = E\left(\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}'\right) = \mathbf{V} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \cdots & \cdots & & \cdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$ (from Eq. (16.1))

where $\mathbf{X}$ is taken as a nonstochastic matrix with full column rank,

The OLS estimator of $\boldsymbol{\beta}$ may be expressed as usual as

$$\hat{\underset{\sim}{\boldsymbol{\beta}}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{y}} = \underset{\sim}{\boldsymbol{\beta}} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{u}}$$

Thus $E\left(\hat{\underset{\sim}{\boldsymbol{\beta}}}\right) = \underset{\sim}{\boldsymbol{\beta}}$ so that OLS is still unbiased.

The variance-covariance matrix of $\hat{\underset{\sim}{\boldsymbol{\beta}}}$ is given by

$$\begin{aligned}\operatorname{var}\left(\hat{\underset{\sim}{\boldsymbol{\beta}}}\right) &= E\left(\hat{\underset{\sim}{\boldsymbol{\beta}}}-\underset{\sim}{\boldsymbol{\beta}}\right)\left(\hat{\underset{\sim}{\boldsymbol{\beta}}}-\underset{\sim}{\boldsymbol{\beta}}\right)' \\ &= \mathrm{E}\left\{\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right\} \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \end{aligned} \qquad (16.3)$$

Thus in the case of heteroscedasticity, the conventional formula

$$\operatorname{var}\left(\hat{\underset{\sim}{\boldsymbol{\beta}}}\right) = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

no longer measures the sampling variances of the OLS estimator, and any application of it is potentially misleading. More importantly, even if one could use Eq. (16.3) to estimate the sampling variances the substitution of these numbers in the conventional t formulas and confidence interval formulas is strictly invalid since the assumptions used in deriving those inference procedures no longer apply. For the same reason the optimal minimum variance property of OLS no longer holds.

Thus in the presence of heteroscedastic disturbances, even though the OLS estimator $\hat{\underset{\sim}{\boldsymbol{\beta}}}$ is unbiased estimator of $\underset{\sim}{\boldsymbol{\beta}}$, it is not the BLUE of $\underset{\sim}{\boldsymbol{\beta}}$. Hence, there is a need to the development of a more appropriate estimator.

16.4. Consequences of using OLS in Presence of Heteroscedasticity

The OLS estimators are derived under the assumption of homoscedasticity and hence in the presence of the problem of heteroscedasticity the OLS estimators are not valid due to the following reasons:

1. The OLS estimators are still unbiased and consistent (in case of large sample), but they are not BLUEs that is they are not possessing minimum variance. In other words, though the OLS estimators are unbiased and consistent, they are not efficient (variances are large) in small as well as large samples.

2. In view of (1), the standard errors of the OLS estimates become large and as a consequence the tests of significance are less powerful. Even if we use the formula (16.3) for obtaining the estimators of the variances of $\hat{\beta}_i$'s, the standard errors of $\hat{\beta}_i$'s will become large and as a consequence, the tests based on them will be misleading. Therefore, the wrong decisions may be taken regarding the inclusion of the variables in the analysis.

3. When there is the problem of heteroscedasticity and if we mistakenly apply OLS formulae (derived under the assumption of homoscedasticity) for the estimate of the common variance of the disturbances u_i, viz. $\hat{\sigma}^2 = \dfrac{\underset{\sim}{e}'\underset{\sim}{e}}{(n-k)}$ then $\hat{\sigma}^2$ is a biased estimator of σ^2 (when the disturbances are homoscedastic $\hat{\sigma}^2$ is unbiased of σ^2). The OLS

estimators of the variances of the estimators $\hat{\beta}_i$'s viz., $\text{var}\left(\hat{\beta}_i\right) = \hat{\sigma}^2 a^{ii}$ (where a^{ii} is the i^{th} diagonal element of $\mathbf{X'X}$) are biased since $\hat{\sigma}^2 = \dfrac{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}}{(n-k)}$ is biased estimator of σ^2. Hence, t and F tests based on it will be misleading. And as a consequence, the usual t and F test (based on $\hat{\sigma}^2$) are very much likely to exaggerate the statistical significance of the conventionally estimated parameters.

Thus if we erroneously disregard heteroscedasticity and use the conventional OLS estimators of the variances of the regression coefficients, the t and F tests of significance based on it will be highly misleading because in situations of heteroscedasticity the usual estimator of σ^2, viz.,

$$\hat{\sigma}^2 = \frac{\underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}}{(n-k)} \tag{16.4}$$

is no longer unbiased.

Thus in case of heteroscedasticity problem instead of using BLUEs of the regression coefficients (those can be obtained using generalized least squares (GLS) method to the regression equation) if we use OLS estimators mainly we get the above problems:

16.5 Self Assessment Questions

1. Explain the problem of heteroscedasticity. What are its sources and consequences?
2. Detail the problem of heteroscedasticity and describe a test procedure for detection of this problem.
3. Explain heteroscedasticity with suitable examples.
4. What is meant by Heteroscedasticity? What are the consequences of using OLS in it's presence?

16.6 References

1. Gujarati, D.N. (2005): *Basic Econometrics, 4th Ed., Tata McGraw-Hill.*
2. Johnston, J. (1984): *Econometric Methods, 3rd Ed., McGraw-Hill,* New York.
3. Montgomery, D.C., Peck, E.A. and Geoffrey Vining, G. (2003): *Introduction to Linear Regression Analysis, 3rd Ed., Wiley*
4. Draper, N.R., and H. Smith(1998): *Applied Regression Analysis, 3rd Ed., John Wiley & Sons,* New York.
5. G.S. Maddala (2001): *Introduction to Econometrics, 3rd Ed., John Wiley & Sons, Ltd.*
6. Johnston, J. and DiNardo J (1997): *Econometric Methods*, 4th Ed., McGraw Hill.
7. Hill, Carter, William Griffiths, and George Judge(2001): *Undergraduate Econometrics, John Willey & Sons,* New York.
8. *Koutsoyiannis, A(1973): Theory of Econometrics, Harper & Row,* New York.

Lesson 17

HETEROSCEDASTICITY: DETECTION AND REMEDIES

17.0 Objective:

This lesson is continuation of Lesson 16 and after studying this lesson, the student will understand a number of detection methods of heteroscedasticity and some remedies of heteroscedasticity.

Structure of the Lesson:

17.1. Introduction

As with multicollinearity, the important practical question is: How does one know that heteroscedasticity is present in a specific situation? Again, as in the case of multicollinearity, there are no hard-and-fast rules for detecting heteroscedasticity, only a few rules of thumb. But this situation is inevitable because σ_i^2 can be known only if we have the entire Y population corresponding to the chosen X's. But such data are an exception rather than the rule in most economic investigations. In this respect the econometrician differs from scientists in fields such as agriculture and biology, where researchers have a good deal of control over their subjects. More often than not, in economic studies there is only one sample Y value corresponding to a particular value of X. And there is no way one can know σ_i^2 from just one Y observation. Therefore, in most cases involving econometric investigations, heteroscedasticity may be a matter of intuition, educated guesswork, prior empirical experience, or sheer speculation.

With the preceding caveat in mind, let us examine some of the informal and formal methods of detecting heteroscedasticity. As the following discussion will reveal, most of these methods are based on the examination of the OLS residuals e_i since they are the ones we observe, and not the disturbances u_i. One hopes that they are good estimates of u_i, a hope that may be fulfilled if the sample size is fairly large.

As we have seen, heteroscedasticity does not destroy the unbiasedness and consistency properties of the OLS estimators, but they are no longer efficient, not even asymptotically (i.e., large sample size). This lack of efficiency makes the usual hypothesis-testing procedure of dubious value. Therefore, remedial measures may be called for. There are two approaches to remediation: when σ_i^2 is known and when σ_i^2 is not known.

17.2. Detection of Heteroscedasticity

For detection of heteroscedasticity, we have several methods which are given below.

1. Park test
2. Glejser's test
3. Spearman's rank correlation test
4. Gold field-Quandt Test
5. Breusch–Pagan–Godfrey Test
6. White's General Heteroscedasticity Test
7. A test for homogeneity of variances
8. Bartlet's test for homogeneity of variances

Let us discuss these methods one by one

17.2.1. Park test

Park suggested that σ_i^2 is some function of the explanatory variable X_i. The functional form he suggested was

$$\sigma_i^2 = \sigma^2 X_i^{\beta} e^{v_i} \Rightarrow \log\sigma_i^2 = \log\sigma^2 + \beta\log X_i + v_i \qquad (17.1)$$

where v_i is the stochastic disturbance term. Since σ_i^2 is generally not known, Park suggests using e_i^2 as a proxy and running the following regression:

$$\boxed{\begin{aligned}\log\ e_i^2 &= \log\sigma^2 + \beta\log X_i + v_i \\ &= \alpha + \beta\log X_i + v_i \qquad \text{where } \alpha = \log\sigma^2\end{aligned}} \qquad (17.2)$$

If β turns out to be statistically significant, it would suggest that heteroscedasticity is present in the data. If it turns out to be insignificant, we may accept the assumption of homoscedasticity. The Park test is thus a two stage procedure. In the first stage we run the OLS regression disregarding the heteroscedasticity question. We obtain e_i^2 from this regression, and then in the second stage we run the regression (17.2).

Although empirically appealing, the Park test has some problems. Goldfeld and Quandt have argued that the error term v_i entering into (17.2) may not satisfy the OLS assumptions and may itself be heteroscedastic. Nonetheless, as a strictly exploratory method, one may use the Park test.

17.2.2. Glejser Test:

The Glejser test is similar in spirit to the Park test. After obtaining the residuals e_i from the OLS regression, Glejser suggests regressing the absolute values of e_i on the X variable that is thought to be closely associated with σ_i^2. In his experiments, Glejser used the following functional forms:

$$\boxed{\begin{aligned} |e_i| &= \beta_1 + \beta_2 X_i + v_i \\ |e_i| &= \beta_1 + \beta_2 \sqrt{X_i} + v_i \\ |e_i| &= \beta_1 + \beta_2 \frac{1}{X_i} + v_i \\ |e_i| &= \beta_1 + \beta_2 \frac{1}{\sqrt{X_i}} + v_i \\ |e_i| &= \sqrt{\beta_1 + \beta_2 X_i} + v_i \\ |e_i| &= \sqrt{\beta_1 + \beta_2 X_i^2} + v_i \end{aligned}} \tag{17.3}$$

where v_i is the error term.

Again as an empirical or practical matter, one may use the Glejser approach. But Goldfeld and Quandt point out that the error term v_i has some problems in that its expected value is nonzero, it is serially correlated, and ironically it is heteroscedastic. An additional difficulty with the Glejser method is that models such as

$$|e_i| = \sqrt{\beta_1 + \beta_2 X_i} + v_i \text{ and } |e_i| = \sqrt{\beta_1 + \beta_2 X_i^2} + v_i \tag{17.4}$$

are nonlinear in the parameters and therefore cannot be estimated with the usual OLS procedure.

Glejser has found that for large samples the first four of the preceding models give generally satisfactory results in detecting heteroscedasticity. As a practical matter, therefore, the Glejser technique may be used for large samples and may be used in the small samples strictly as a qualitative device to learn something about heteroscedasticity.

17.2.3. Spearman's Rank Correlation Test:

We define the Spearman's rank correlation coefficient as

$$\boxed{r_s = 1 - 6\sum_{i=1}^{n} d_i^2 \Big/ \left[n(n-1)\right]} \tag{17.5}$$

where d_i = difference in the ranks assigned to two different characteristics of the i^{th} individual or phenomenon and n = number of individuals or phenomena ranked. The preceding rank correlation coefficient can be used to detect heteroscedasticity as follows:

Assume $|Y_i| = \beta_0 + \beta_1 X_i + u_i$

Step 1. Fit the regression to the data on *Y* and *X* and obtain the residuals e_i .

Step 2. Ignoring the sign of e_i , that is, taking their absolute value $|e_i|$, rank both $|e_i|$ and X_i (or $\hat{Y}_i$) according to an ascending or descending order and compute the Spearman's rank correlation coefficient r_s given in Eq. (17.5).

Step 3. Assuming that the population rank correlation coefficient ρ_s is zero and *n* > 8, the significance of the sample r_s can be tested by the *t* test as follows:

$$\boxed{t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}} \qquad (17.6)$$

If the computed t value exceeds the critical t value at the chosen level of significance with d.f. = n – 2, we may accept the hypothesis of heteroscedasticity; otherwise we may reject it. If the regression model involves more than one X variable, r_s can be computed between $|e_i|$ and each of the X variables separately and can be tested for statistical significance by the *t* test given in Eq. (17.6).

17.2.4. Goldfeld-Quandt Test:

This popular method is applicable if one assumes that the heteroscedastic variance σ_i^2 is positively related to *one* of the explanatory variables in the regression model. For simplicity, consider the usual two-variable model:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad (17.7)$$

Suppose σ_i^2 is positively related to X_i as

$$\sigma_i^2 = \sigma^2 X_i^2 \qquad (17.8)$$

where σ^2 is a constant.. Assumption (17.8) postulates that σ_i^2 is proportional to the square of the X variable. Such an assumption has been found quite useful in the study of family budgets.

If Eq. (17.8) is appropriate, it would mean σ_i^2 would be larger, the larger the values of X_i. If that turns out to be the case, heteroscedasticity is most likely to be present in the model. To test this explicitly, Goldfeld and Quandt suggest the following steps:

Step 1. Order or rank the observations according to the values of X_i, beginning with the lowest *X* value.

Step 2. Omit *c* central observations, where *c* is specified a priori, and divide the remaining $(n - c)$ observations into two groups each of $(n - c)/2$ observations.

Step 3. Fit separate OLS regressions to the first $(\text{n - c})/2$ observations and the last $(\text{n - c})/2$ observations, and obtain the respective residual sums of squares RSS_1 and RSS_2, RSS_1 representing the RSS from the regression corresponding to the smaller X_i values (the small variance group) and RSS_2 that from the larger X_i values (the large variance group). These RSS each have

$$\frac{(n-c)}{2}-k \text{ or } \left(\frac{n-c-2k}{2}\right) \text{ d.f.} \qquad (17.9)$$

where *k* is the number of parameters to be estimated, including the intercept. For the two-variable case *k* is of course 2.

Step 4. Compute the ratio

$$\boxed{F=\frac{\text{RSS}_2/\text{df}}{\text{RSS}_1/\text{df}}} \qquad (17.10)$$

If u_i are assumed to be normally distributed (which we usually do), and *if the assumption of homoscedasticity is valid*, then it can be shown that the computed F follows the F - distribution with numerator and denominator d.f. each of $(n-c-2k)/2$.

If in an application the computed F is greater than the critical F at the chosen level of significance, we can reject the hypothesis of homoscedasticity, that is, we can say that heteroscedasticity is very likely.

The ability of the Goldfeld–Quandt test to do this successfully depends on how *c* is chosen. The power will be low if c is too large, so that RSS_1 and RSS_2 have very few degrees of freedom. However, if c is too small, the power will also be low, since any contrast between RSS_1 and RSS_2 is reduced. A rough guide is to set c at approximately $n/3$. For the two-variable model the Monte Carlo experiments done by Goldfeld and Quandt suggest that *c* is about 8 if the sample size is about 30, and it is about 16 if the sample size is about 60.

Before moving on, it may be noted that in case there is more than one X variable in the model, the ranking of observations, the first step in the test, can be done according to any one of them. Thus in the model: $Y_i=\beta_1+\beta_2X_{2i}+\beta_3X_{3i}+\beta_4X_{4i}+u_i$, we can rank or order the data according to any one of these X's. If a priori we are not sure which X variable is appropriate, we can conduct the test on each of the X variables, or via a Park test, in turn, on each X.

17.2.5. Breusch–Pagan–Godfrey (BPG) Test:

The success of the Goldfeld–Quandt test depends not only on the value of *c* (the number of central observations to be omitted) but also on identifying the correct X variable with which to order the observations. This limitation of this test can be avoided if we consider the Breusch–Pagan–Godfrey (BPG) test.

To illustrate this test, consider the *k*-variable linear regression model

$$Y_i=\beta_1+\beta_2X_{2i}+\ldots+\beta_kX_{ki}+u_i\ ,\ \ i=1,2,...,n \qquad (17.11)$$

Assume that the error variance σ_i^2 is described as

$$\sigma_i^2 = f\left(\alpha_1 + \alpha_2 Z_{2i} + \cdots + \alpha_m Z_{mi}\right),\ i = 1,2,...,n \tag{17.12}$$

that is, σ_i^2 is some function of the nonstochastic variables Z's; some or all of the X's can serve as Z's. Specifically, assume that

$$\sigma_i^2 = \alpha_1 + \alpha_2 Z_{2i} + \cdots + \alpha_m Z_{mi},\ i = 1,2,...,n \tag{17.13}$$

that is, σ_i^2 is a linear function of the Z's. If $\alpha_2 = \alpha_3 = \cdots = \alpha_m = 0, \sigma_i^2 = \alpha_1$, which is a constant. Therefore, to test whether σ_i^2 is homoscedastic, one can test the hypothesis that $H_0 : \alpha_2 = \alpha_3 = \cdots = \alpha_m = 0$.

This is the basic idea behind the BPG test. The actual test procedure is as follows.

Step 1. Estimate (17.11) by OLS and obtain the residuals $e_1, e_2, \ldots, e_n$.

Step 2. Obtain the maximum likelihood (ML) estimator of σ^2 given by $\tilde{\sigma}^2 = \sum e_i^2 / n$.

[*Note:* The OLS estimator is $\sum e_i^2 / (n-k)$.]

Step 3. Construct an auxiliary variable p_i defined as

$$p_i = e_i^2 / \tilde{\sigma}^2 \tag{17.14}$$

which is simply each residual squared divided by $\tilde{\sigma}^2$.

Step 4. Regress p_i thus constructed on the Z's as

$$p_i = \alpha_1 + \alpha_2 Z_{2i} + \cdots + \alpha_m Z_{mi} + v_i \quad , \mathrm{i} = 1,2,\cdots,\mathrm{n}$$

where v_i is the residual term of this regression. Compute

$$\hat{p}_i = \hat{\alpha}_1 + \hat{\alpha}_2 Z_{2i} + \cdots + \hat{\alpha}_m Z_{mi} \quad \mathrm{i} = 1,2,\cdots,\mathrm{n} \tag{17.15}$$

Step 5. Obtain the ESS (explained sum of squares $= \sum \hat{p}_i^2 - n\overline{p}^2$ from Eq. (17.15) and define

$$\Theta = ESS/2 \tag{17.16}$$

Assuming u_i are normally distributed, one can show that if there is homoscedasticity and if the sample size *n* increases indefinitely, then

$$\boxed{\Theta \overset{asy}{\sim} \chi_{m-1}^2} \tag{17.17}$$

that is, Θ follows the chi-square distribution with (*m* − 1) degrees of freedom. (*Note: asy* means asymptotically)

Therefore, if in an application the computed Θ (obtained in Eq. (17.16)) exceeds the critical χ^2 value at the chosen level of significance with m-1 d.f., one can reject the hypothesis of homoscedasticity; otherwise one does not reject it.

17.2.6. White's General Heteroscedasticity Test:

Unlike the Goldfeld–Quandt test, which requires reordering the observations with respect to the X variable that supposedly caused heteroscedasticity, or the BPG test, which is sensitive to the normality assumption, the general test of heteroscedasticity proposed by White does not rely on the normality assumption and is easy to implement. As an illustration of the basic idea, consider the following three-variable regression model (the generalization to the *k*-variable model is straightforward):

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \tag{17.18}$$

The White test procedure as follows:

Step 1. Based on the given the data, we estimate the model (17.18) and obtain the residuals, e_i .

Step 2. We then run the following (*auxiliary*) regression:

$$e_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{3i}^2 + \alpha_6 X_{2i} X_{3i} + v_i \tag{17.19}$$

That is, the squared residuals from the original regression are regressed on the original X variables or regressors, their squared values, and the cross product(s) of the regressors. Higher powers of regressors can also be introduced. Note that there is a constant term in this equation even though the original regression may or may not contain it. Obtain the R^2 from this (auxiliary) regression.

Step 3. Under the null hypothesis that there is no heteroscedasticity, it can be shown that sample size (n) times the R^2 obtained from the auxiliary regression *asymptotically* follows the chi-square distribution with d.f. equal to the number of regressors (excluding the constant term) in the auxiliary regression. That is,

$$\boxed{nR^2 \underset{asy}{\sim} \chi^2_{\text{df}}} \tag{17.20}$$

where d.f. is as defined previously. In our example, the d.f. is 5 since there are 5 regressors in the auxiliary regression.

Step 4. If the chi-square value obtained in Eq. (17.20) exceeds the critical chi-square value at the chosen level of significance, the conclusion is that there is heteroscedasticity. If it does not exceed the critical chi-square value, there is no heteroscedasticity, which is to say that in the auxiliary regression Eq. (17.19), $\alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0$.

A comment with regard to the White test is, if a model has several regressors, then introducing all the regressors, their squared (or higher powered) terms, and their cross products can quickly consume degrees of freedom. Therefore, one must use caution in using the test.

17.2.7. A test for homogeneity of variances:

If we have plentiful cross-section data we may apply this standard test for homogeneous variances to the Y data. If we split the data on endogenous variable Y in to m classes according to the size of Y and compute

$$\boxed{\lambda = \prod_{i=1}^{m} \frac{\left(s_i/n_i\right)^{n_i/2}}{\left(\sum_{i=1}^{n} s_i \Big/ \sum_{i=1}^{n} n_i\right)^{\sum_{i=1}^{n} n_i \big/ 2}}} \tag{17.23}$$

where $n_i =$ number of observations in i[th] class

$$s_i = \sum_{j=1}^{n} \left(Y_{ij} - \overline{Y}_i\right)^2 \tag{17.24}$$

Now under $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_m^2 = \sigma^2$

$$\boxed{\begin{array}{l} \mu = -2\log_e \lambda \sim \chi^2_{m-1} \\ \text{and } \mu = -2\log_e \lambda = \left(\sum_{i=1}^{n} n_i\right)\log_e\left(\sum_{i=1}^{n} s_i \Big/ \sum_{i=1}^{n} n_i\right) - \sum_{i=1}^{m}\left(n_i \log_e\left(s_i/n_i\right)\right) \end{array}} \tag{17.25}$$

Then under the assumption of homogeneous variances, if calculated μ is greater than the table χ^2 value at given level of significance with m-1 d.f. then we may conclude that there is a problem of heteroscedasticity.

17.2.8. Bartlet's test for homogeneity of variances:

It is a slightly modified and powerful of the above test and is given below:

$$A = f\log_e\left(\sum_{i=1}^{m} s_i \Big/ \sum_{i=1}^{m} f_i\right) - \sum_{i=1}^{m}\left(f_i \log_e\left(s_i/f_i\right)\right)$$

$$B = 1 + \frac{1}{3(m-1)}\left(\sum_{i=1}^{m}\frac{1}{f_i} - \frac{1}{f}\right)$$

where $f_i = n_i - 1$, i=1,2,…,m f$=\sum_{i=1}^{m} f_i$

Under H_0 : homoscedasticity variances

$$\boxed{\frac{A}{B} \sim \chi^2_{m-1}(approximately)} \tag{17.26}$$

If A/B greater than χ^2 table value at 5% level of significance with m-1 d.f., then there is a problem of heteroscedasticity.

Illustration 17.1:

The following table gives Per capita personal consumption expenditure (Y) and per capita disposable personal income (X) (in dollars) for the United States, 1970-1984, collected from 20 families. The data is arranged with respect to the order of per capita disposable personal income (X). Using this data, *<u>examine for the presence of heteroscedasticity using Goldfeld–Quandt test</u>*.

Table 17.1

Family no.	Y (Expenditure)	X (Income)	Family no.	Y (Expenditure)	X (Income)
1	6.1	6.2	11	25.5	26.1
2	8.0	8.1	12	25.0	28.3
3	10.3	10.3	13	29.3	30.1
4	12.1	12.1	14	31.2	32.3
5	13.1	14.1	15	33.1	34.5
6	14.8	16.4	16	31.8	36.6
7	17.9	18.2	17	33.5	38.0
8	19.8	20.1	18	38.8	40.2
9	19.9	22.3	19	40.7	42.3
10	21.6	24.1	20	38.6	44.7

Source: Introduction to Econometrics, Third Edition, G.S.MADDALA, 3rd Edition, John Wiley & Sons Ltd, p. 200.

Solution:
As suggested in the Goldfeld–Quandt test, let us omit the central c=4 (9th to 12th) observations and treat the sample pertaining to first 8 families as LOWER sample and the sample pertaining to last 8 families as UPPER sample.
In the first step let us obtain the OLS regressions for lower and upper samples separately as shown below.

Lower Sample			Upper Sample		
Family no.	Y (Expenditure)	X (Income)	Family no.	Y (Expenditure)	X (Income)
1	6.1	6.2	13	29.3	30.1
2	8.0	8.1	14	31.2	32.3
3	10.3	10.3	15	33.1	34.5
4	12.1	12.1	16	31.8	36.6
5	13.1	14.1	17	33.5	38.0
6	14.8	16.4	18	38.8	40.2
7	17.9	18.2	19	40.7	42.3
8	19.8	20.1	20	38.6	44.7
From the Lower sample we have $\overline{X}=13.20$ $\overline{Y}=12.7625$ $\sum X^2=1561$ $\sum Y^2=1457$			From the Upper sample we have $\overline{X}=37.3375$ $\overline{Y}=34.6250$ $\sum X^2=11327$ $\sum Y^2=9713$		

$\sum XY = 1507$ $\hat{\beta} = 0.9539$ $\hat{\alpha} = 0.1715$ Regression equation: $\hat{Y} = 0.1715 + 0.9539\, X$	$\sum XY = 10475$ $\hat{\beta} = 0.7641$ $\hat{\alpha} = 6.0938$ Regression equation: $\hat{Y} = 6.0938 + 0.7641\, X$

In the second step we compute RSS's for lower and upper samples as shown below.

Lower Sample			Upper Sample		
Y	$\hat{Y}$	e_i	Y	$\hat{Y}$	e_i
6.1	6.1808	-0.0808	29.3	29.0945	0.2055
8.0	7.8978	0.1022	31.2	30.7756	0.4244
10.3	9.9963	0.3037	33.1	32.4567	0.6433
12.1	11.7132	0.3868	31.8	34.0614	-2.2614
13.1	13.6210	-0.5210	33.5	35.1313	-1.6312
14.8	15.8149	-1.0149	38.8	36.8124	1.9876
17.9	17.5318	0.3682	40.7	38.4171	2.2829
19.8	19.3442	0.4558	38.6	40.2510	-1.6510
	Total	0.0000		Total	0.0001
$\mathrm{RSS}_1 = \sum e_i^2 = 1.9035$			$\mathrm{RSS}_2 = \sum e_i^2 = 20.2995$		

Now, from Eq. (17.10), the F-ratio of Goldfeld–Quandt test is computed as

$$F = \frac{\mathrm{RSS}_2/\mathrm{d.f.}}{\mathrm{RSS}_1/\mathrm{d.f.}} \quad \text{where d.f.} = \left(\frac{n-c}{2}\right) - k = \frac{20-4}{2} - 2 = 8 - 2 = 6$$

$$F = \frac{20.2995/6}{1.9035/6} = 10.6643$$

The critical F-values from the F-tables are $F_{6,6}$ =4.28 at 5% l.o.s. and $F_{6,6}$ =8.47 at 1% l.o.s.

Conclusion drawn:

Since, the calculated F-ratio exceeds the critical F-values at both 5% and 1% l.o.s., we ***conclude that there is evidence of heteroscedasticity in the given data***.

Note: The student is advised to *examine for the presence of heteroscedasticity for the above given example using other tests viz., Park test, Glejser test, Spearman's rank correlation test, and BPG test and see whether same conclusion is drawn or not.*

17.3 Remedies of Heteroscedasticity

1. When σ_i^2 is Known: The Method of Weighted Least Squares
2. When σ_i^2 is Not Known

   a. Generalized Lest Square Method

b. Log Transformation

17.3.1. When σ_i^2 is Known: The Method of Weighted Least Squares

If σ_i^2 is known, the most straightforward method of correcting heteroscedasticity is by means of weighted least squares, for the estimators thus obtained are BLUE.
To illustrate the method, we use the two-variable model $Y = \beta_1 + \beta_2 X_i + u_i$. The unweighted least-squares method minimizes

$$\sum \hat{u}_i^2 = \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)^2 \qquad (17.27)$$

to obtain the estimates, whereas the weighted least-squares method minimizes the weighted residual sum of squares:

$$\sum w_i \hat{u}_i^2 = \sum w_i \left(Y_i - \hat{\beta}_1^* - \hat{\beta}_2^* X_i\right)^2 \qquad (17.28)$$

where $\hat{\beta}_1^*$ and $\hat{\beta}_2^*$ are the weighted least-squares estimators and where the weights w_i are such that

$$w_i = \frac{1}{\sigma_i^2} \qquad (17.29)$$

that is, the weights are inversely proportional to the variance of u_i or Y_i conditional upon the given X_i, it being understood that $\operatorname{var}\left(u_i / X_i\right) = \operatorname{var}\left(Y_i / X_i\right) = \sigma_i^2$.
Differentiating Eq. (17.28) with respect to $\hat{\beta}_1^*$ and $\hat{\beta}_2^*$, we obtain

$$\frac{\partial \sum w_i \hat{u}_i^2}{\partial \hat{\beta}_1^*} = 2\sum w_i \left(Y_i - \hat{\beta}_1^* - \hat{\beta}_2^* X_i\right)(-1)$$

$$\frac{\partial \sum w_i \hat{u}_i^2}{\partial \hat{\beta}_2^*} = 2\sum w_i \left(Y_i - \hat{\beta}_1^* - \hat{\beta}_2^* X_i\right)(-X_i)$$

Setting the preceding expressions equal to zero, we obtain the following two normal equations:

$$\sum w_i Y_i = \hat{\beta}_1^* \sum w_i + \hat{\beta}_2^* \sum w_i X_i \qquad (17.30)$$

$$\sum w_i X_i Y_i = \hat{\beta}_1^* \sum w_i X_i + \hat{\beta}_2^* \sum w_i X_i^2 \qquad (17.31)$$

Notice the similarity between these normal equations and the normal equations of the unweighted least squares.
Solving these equations simultaneously, we obtain

$$\hat{\beta}_1^* = \overline{Y}^* - \hat{\beta}_2^* X^* \qquad (17.32)$$

and

$$\boxed{\hat{\beta}_2^* = \frac{\left(\sum w_i X_i\right)\left(\sum w_i X_i Y_i\right) - \left(\sum w_i X_i\right)\left(\sum w_i Y_i\right)}{\left(\sum w_i\right)\left(\sum w_i X_i^2\right) - \left(\sum w_i X_i\right)^2}} \qquad (17.33)$$

Note: $\overline{Y}^* = \sum w_i Y_i / \sum w_i$ and $X^* = \sum w_i X_i / \sum w_i$. As can be readily verified, these weighted means coincide with the usual or unweighted means $\overline{Y}$ and $\overline{X}$ when $w_i = w$, a constant, for all i.

17.3.2. When σ_i^2 is Not Known: Generalized Least Squares Method:

As noted earlier, if true σ_i^2 are known, we can use the WLS method to obtain BLUE estimators. Since the true σ_i^2 are rarely known, is there a way of obtaining *consistent* (in the statistical sense) estimates of the variances and covariances of OLS estimators even if there is heteroscedasticity? The answer is yes.

When there is a problem of heteroscedasticity we should not apply OLS method for regression analysis, instead that we should apply generalized least squares (GLS) method for regression analysis. The GLS method is as given below:-
In the case of heteroscedasticity, we may rewrite the general linear model as follows:-

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} \ldots + \beta_k X_{ki} + u_i \quad i = 1, 2, \ldots, n$$

with $E(u_i) = 0, E(u_i u_j) = 0 \; if \; i \neq j \;\; \& \;\; E(u_i^2) = \sigma_i^2 = \sigma^2 \lambda_i$ (17.34)

Where σ^2 is unknown and λ_i is known, which may be taken any one of the following forms, depending upon the application.

$$\begin{array}{ll} \text{(i) } \lambda_i = X_i^2 & \text{(v) } \lambda_i = \left(a_0 + a_1 X_i\right)^2 \\ \text{(ii) } \lambda_i = X_i & \text{(vi) } \lambda_i = \left(a_0 + a_1 / X_i\right)^2 \\ \text{(iii) } \lambda_i = 1 / X_i^2 & \text{(vii) } \lambda_i = \left(a_0 + a_1 \sqrt{X_i}\right)^2 \\ \text{(iv) } \lambda_i = 1 / X_i & \text{(viii) } \lambda_i = \left(a_0 + a_1 / \sqrt{X_i}\right)^2 \end{array} \qquad (17.35)$$

Here X is an important explanatory variable which is expected to be well associated with the $\operatorname{var}(u_i)$. In the above a_0 and a_1 may be obtained by regressing the OLS residuals e_i on X_i or X_i^{-1} or $X_i^{1/2}$ or $X_i^{-1/2}$.

The general linear models given in (17.34) can be written in the matrix notation as follows:-

$$\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}} \qquad (17.36)$$

With $E(\underset{\sim}{\mathbf{u}}) = \underset{\sim}{\mathbf{0}}$ and $\operatorname{var}(\underset{\sim}{\mathbf{u}}) = E(\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}') = \sigma^2 \Omega$

When $\Omega = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$

Since Ω is positive definite matrix there exists a nonsingular matrix $\mathbf{P}$ such that

$$\Omega = \mathbf{PP'} \text{ where } \mathbf{P} = diag\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n}\right)$$

So that $\mathbf{P}^{-1}\boldsymbol{\Omega}(\mathbf{P}')^{-1} = \mathbf{I_n}$ (17.37)

Premultiplying (17.36) with $\mathbf{P}^{-1}$ we get

$$\underset{\sim}{\mathbf{y}}^* = \mathbf{X}^* \underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{u}}^* \qquad (17.38)$$

where $\underset{\sim}{\mathbf{y}}^* = \mathbf{P}^{-1}\underset{\sim}{\mathbf{y}}$, $\mathbf{X}^* = \mathbf{P}^{-1} X$, and $\underset{\sim}{\mathbf{u}}^* = \mathbf{P}^{-1}\underset{\sim}{\mathbf{u}}$ (17.39)

with i. $E(\underset{\sim}{\mathbf{u}}^*) = \underset{\sim}{\mathbf{0}}$,

ii. $E\left(\underset{\sim}{\mathbf{u}}^{*}\underset{\sim}{\mathbf{u}}^{*\prime}\right)=E\left(\mathbf{P}^{-1}\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}'\left(\mathbf{P}'\right)^{-1}\right)=\sigma^{2}\mathbf{P}^{-1}\Omega\left(\mathbf{P}'\right)^{-1}=\sigma^{2}I_{n}$ (17.40)

$\left(\because \text{from Eqs. (17.36) \& (17.37)}\right)$

Thus Eq. (17.38) is now the classical linear model for which we can apply OLS method. The OLS estimator of $\underset{\sim}{\boldsymbol{\beta}}$ in (17.38) is given by

$$
\begin{aligned}
\underset{\sim}{\mathbf{b}} &= \left(\mathbf{X}^{*\prime}\mathbf{X}^{*}\right)^{-1}\mathbf{X}^{*}\underset{\sim}{\mathbf{y}}^{*} \\
&= \left(\mathbf{X}'\left(P'\right)^{-1}\mathbf{P}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{P}'\right)^{-1}\mathbf{P}^{-1}\underset{\sim}{\mathbf{y}} \\
&= \left(\mathbf{X}'\left(\mathbf{P}\mathbf{P}'\right)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{P}\mathbf{P}'\right)^{-1}\underset{\sim}{\mathbf{y}} \\
&= \left(\mathbf{X}'\Omega^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Omega^{-1}\underset{\sim}{\mathbf{y}} \qquad (17.41)
\end{aligned}
$$

which is called the GLS estimator of $\underset{\sim}{\boldsymbol{\beta}}$ of the model (17.36), which is also BLUE of $\underset{\sim}{\boldsymbol{\beta}}$ and its minimum variance is given by

$$\boxed{\operatorname{var}\left(\underset{\sim}{b}\right)=\sigma^{2}\left(\mathbf{X}^{*\prime}\mathbf{X}^{*}\right)^{-1}=\sigma^{2}\left(\mathbf{X}'\Omega^{-1}\mathbf{X}\right)^{-1}} \qquad (17.42)$$

An unbiased estimator of σ^{2} is given by

$$\boxed{\tilde{\sigma}^{2}=\frac{\underset{\sim}{\mathbf{e}}^{*\prime}\underset{\sim}{\mathbf{e}}^{*}}{(n-k)}=\frac{\underset{\sim}{\mathbf{e}}'\Omega\underset{\sim}{\mathbf{e}}}{(n-k)}} \qquad (17.43)$$

where

$$
\begin{aligned}
\underset{\sim}{\mathbf{e}}^{*} &= \underset{\sim}{\mathbf{y}}^{*}-\mathbf{X}^{*}\underset{\sim}{\mathbf{b}} \quad \left(\text{using OLS formulae}\right) \\
&= \mathbf{P}^{-1}\left(\underset{\sim}{\mathbf{y}}-\mathbf{X}\underset{\sim}{\mathbf{b}}\right)=\mathbf{P}^{-1}\underset{\sim}{\mathbf{e}}
\end{aligned}
$$

Here $\underset{\sim}{\mathbf{e}}^{*}$ is OLS residual vector of (17.38)
and $\underset{\sim}{\mathbf{e}}$ is GLS residual vector of (17.36)

Thus GLS estimator $\left(\underset{\sim}{\mathbf{b}}\right)$ of $\underset{\sim}{\boldsymbol{\beta}}$ for model can be obtained by applying OLS method to the model (17.38).

The GLS estimator, thus, can be computed form (17.41) for a given Ω and standard errors of the estimates can be obtained using (17.42) & (17.43) so that the usual significance tests and confidence intervals can be constructed for β_i's .

Thus applying GLS method to the original data is equivalent to applying OLS method to the transformed data (where the type of transformation depends upon the nature of heteroscedasticity given in Eq. (17.35)).

17.3.3. Log Transformation:

If instead of running the regression (17.34) we can run the regression equation .

$$\log Y_i = \beta_1 + \beta_2 \log X_{2i} + \beta_3 \log X_{3i} + \ldots + \beta_k \log X_{ki} + u_i \quad (17.44)$$

Very often this model reduce the problem of heteroscedasticity. This is because log transformation compresses the scales in which the variables are measured. There by reducing a tenfold difference between two values to a two fold difference. Thus, the number 80 is 10 times of 8, $\log_e 80(=4.382)$ is only twice as range as $\log_e 8(=2.0794)$.

An additional advantage of the log-transformation is that the slope coefficients β_i's measure the elasticities of Y with respect to X_l, explanatory variable that is percentage change in Y for a percentage change in X_l for example, if Y is consumption and X_2 is income, β_2 in (17.34) will measure only the rate of change of mean consumption for a unit change in income, where as β_1 in transformed model (17.44) measures income elasticity.

It is one reason why the log models are quite popular in empirical econometrics.

Suppose we want to run the regression function

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad (17.45)$$

where X_i = Labour productivity from i[th] firm

Y_i = Labour compensation for i[th] firm .

1. Test for heteroscedasticity using Glejser approach

   (form is $|e_i| = a_0 + a_1\sqrt{X_i}$)

2. If heteroscedasticity presents, obtain the estimates of β_0 & β_1 using GLS method.

Note :

If $Y_i = \beta_0 + \beta_1 X_i + u_i$ and $\operatorname{var}(u_i) = \sigma^2 \lambda_i$ (17.46)

Then transform the above model by dividing it with $\sqrt{\lambda_i}$ which results into the model

$$Y_i^* = \beta_0 z_{0i} + \beta_1 z_{1i} + v_i \quad (17.47)$$

Where

$$Y_i^* = Y_i \big/ \sqrt{\lambda_i}$$

$$z_{0i} = 1 \big/ \sqrt{\lambda_i}$$

$$z_{1i} = X_i \big/ \sqrt{\lambda_i}$$

$$v_i = u_i \big/ \sqrt{\lambda_i}$$

Now $\operatorname{var}(v_i) = \operatorname{var}\left(u_i \big/ \sqrt{\lambda_i}\right) = \dfrac{v\operatorname{ar}(u_i)}{\lambda_i} = \sigma^2$

Now applying GLS (WLS) method to model (17.46) equivalent to applying OLS method to model (17.47), which is optimum in practical situation.

REMARKS:

1. Documenting the consequences of heteroscedasticity is easier than detecting it. There are several diagnostic tests available, but one cannot tell for sure which will work in a given situation.

2. Even if heteroscedasticity is suspected and detected, it is not easy to correct the problem. If the sample is large, one can obtain White's heteroscedasticity corrected standard errors of OLS estimators and conduct statistical inference based on these standard errors.

3. Otherwise, on the basis of OLS residuals, one can make educated guesses of the likely pattern of heteroscedasticity and transform the original data in such a way that in the transformed data there is no heteroscedasticity.

3. We emphasize that all the transformations discussed previously are ad hoc; we are essentially speculating about the nature of σ_i^2 . Which of the transformations discussed previously will work will depend on the nature of the problem and the severity of heteroscedasticity. There are some additional problems with the transformations we have considered that should be borne in mind:

   i). When we go beyond the two-variable model, we may not know a priori which of the *X* variables should be chosen for transforming the data.

   ii). Log transformation as discussed in section 17.3 is not applicable if some of the *Y* and *X* values are zero or negative.

   iii). When σ_i^2 are not directly known and are estimated from one or more of the transformations that we have discussed earlier, all our testing procedures using the *t* tests, *F* tests, etc., are *strictly speaking valid only in large samples.* Therefore, one has to be careful in interpreting the results based on the various transformations in small or finite samples.

17.4 Self Assessment Questions

1. Explain various detection methods of heteroscedasticity.
2. Explain Park test for detecting the heteroscedasticity.
3. Explain Glejser's test for detecting the heteroscedasticity.
4. Describe Spearman rank correlation test for detecting the heteroscedasticity.
5. Describe Gold-field test for detecting the heteroscedasticity.
6. Explain Breusch–Pagan–Godfrey (BPG) test for detecting the heteroscedasticity.
7. Explain White's test for detecting the heteroscedasticity.
8. Explain Bartlett's test for testing the homogeneity of variances.
9. Detail the problem of heteroscedasticity and describe a test procedure for detection of this problem.
10. Explain the method of weighted least squares.
11. Distinguish between weighted least squares method and ordinary least squares method.
12. Derive the weighted least squares estimators of the parameters of a linear model with heteroscedastic disturbances.

13. Distinguish between weighted least squares method and ordinary least squares method.
14. Derive the weighted least squares estimators in a linear model with heteroscedastic disturbances.
15. Derive the generalized least squares estimators in a linear model with heteroscedastic disturbances.
16. Distinguish between weighted least squares method and generalized least squares method.

17.5 References

1. Gujarati, D.N. (2005): *Basic Econometrics, 4th Ed., Tata McGraw-Hill.*

2. Johnston, J. (1984): *Econometric Methods, 3rd Ed., McGraw-Hill,* New York.

3. Montgomery, D.C., Peck, E.A. and Geoffrey Vining, G. (2003): *Introduction to Linear Regression Analysis, 3rd Ed., Wiley*

4. Draper, N.R., and H. Smith(1998): *Applied Regression Analysis, 3rd Ed., John Wiley & Sons,* New York.

5. G.S. Maddala (2001): *Introduction to Econometrics, 3rd Ed., John Wiley & Sons, Ltd.*

6. Johnston, J. and DiNardo J (1997): *Econometric Methods*, 4th Ed., McGraw Hill.

7. Hill, Carter, William Griffiths, and George Judge(2001): *Undergraduate Econometrics, John Willey & Sons,* New York.

8. *Koutsoyiannis, A(1973): Theory of Econometrics, Harper & Row,* New York.

Lesson 18

AUTO CORRELATION - NATURE, SOURCES AND CONSEQUENCES

18.0 Objective:

In Lesson 16 we considered the consequences of relaxing the assumption that the disturbance (error) terms have common variance (homoscedasticity). We now come to the next assumption that the disturbance terms in the regression model are independent. This lesson relaxes this assumption, which means the disturbances are correlated. The objective of this lesson is to discuss the nature, sources and consequences of correlated disturbances.

Structure of the Lesson:

18.1 Introduction

The student may note that there are generally three types of data that are available for empirical analysis: (1) cross section, (2) time series, and (3) combination of cross section and time series, also known as pooled data. In developing the classical linear regression model (CLRM) we made several assumptions. However, we noted that *not* all these assumptions would hold in every type of data. As a matter of fact, we saw in the previous lessons that the assumption of homoscedasticity, or equal error variance, may not be always tenable in cross-sectional data. In other words, cross-sectional data are often plagued by the problem of heteroscedasticity. However, in cross-section studies, data are often collected on the basis of a random sample of cross-sectional units, such as households (in a consumption function analysis) or firms (in an investment study analysis) so that there is no prior reason to believe that the error term pertaining to one household or a firm is correlated with the error term of another household or firm. If by chance such a correlation is observed in cross-sectional units, it is called spatial autocorrelation, that is, correlation in space rather than over time. However, it is important to remember that, in cross-sectional analysis, the ordering of the data must have some logic, or economic interest, to make sense of any determination of whether (spatial) autocorrelation is present or not.

The situation, however, is likely to be very different if we are dealing with time series data, for the observations in such data follow a natural ordering over time so that successive observations are likely to exhibit inter-correlations, especially if the time interval between successive observations is short, such as a day, a week, or a month rather than a year. If you observe stock price indexes, such as the BSE SENSEX or NSE NIFTY over successive days, it is not unusual to find that these indexes move up or down for several days in succession. Obviously, in situations like this, the assumption of no autocorrelation (no serial correlation) in the error terms that underlies the CLRM will be violated.

In this lesson we take a critical look at this assumption with a view to answering the following questions:

1. What is the nature and sources of autocorrelation?
2. What are the theoretical and practical consequences of autocorrelation?

The student will find this lesson is in many ways similar to Lesson 16 on heteroscedasticity in that under both heteroscedasticity and autocorrelation, the usual OLS estimators, although linear, unbiased, and asymptotically (i.e., in large samples) normally distributed, are no longer minimum variance among all linear unbiased estimators. In short, they are not efficient relative to other linear and unbiased estimators. Put differently, they may not be BLUEs. As a result, the usual, t, F, and χ^2 may not be valid.

18.2 The Nature and Sources of Autocorrelation

The term autocorrelation may be defined as "correlation between members of series of observations ordered in time [as in time series data] or space [as in cross-sectional data]''. In the regression context, the CLRM assumes that such autocorrelation does not exist in the disturbances u_i. Symbolically,

$$E\left(u_i u_j\right) = 0 \quad \text{for all } i \neq j$$

Put simply, the CLRM assumes that the disturbance term relating to any observation is not influenced by the disturbance term relating to any other observation. For example, if we are dealing with quarterly time series data involving the regression of output on labor and capital inputs and if, say, there is a labour strike affecting output in one quarter, there is no reason to believe that this disruption will be carried over to the next quarter. That is, if output is lower this quarter, there is no reason to expect it to be lower next quarter. Similarly, if we are dealing with cross-sectional data involving the regression of family consumption expenditure on family income, the effect of an increase of one family's income on its consumption expenditure is not expected to affect the consumption expenditure of another family. However, if there is such dependence, we have autocorrelation. Symbolically,

$$E\left(u_i u_j\right) \neq 0 \quad \text{for all } i \neq j$$

In this situation, the disruption caused by a strike this quarter may very well affect output next quarter, or the increases in the consumption expenditure of one family may very well prompt another family to increase its consumption expenditure.

Before we find out why autocorrelation exists, it is essential to clear up some terminological questions. Although it is now a common practice to treat the terms autocorrelation and serial correlation synonymously, some authors prefer to distinguish the two terms. For example, Tintner defines autocorrelation as "lag correlation of a given series with itself, lagged by a number of time units,'' whereas he reserves the term serial correlation to "lag correlation between two different series.'' Thus, correlation between two time series such as $u_1, u_2, \ldots, u_{10}$ and $u_2, u_3, \ldots, u_{11}$, where the former is the latter series lagged by one time period, is *autocorrelation,* whereas correlation between time series such as $u_1, u_2, \ldots, u_{10}$ and $v_2, v_3, \ldots, v_{11}$, where *u* and *v* are two different time series, is called *serial correlation*.

The following are the reasons or sources of autocorrelation:

1. **Inertia.** A salient feature of most economic time series is inertia, or sluggishness. As is well known, time series such as GNP, price indexes, production, employment, and unemployment exhibit (business) cycles. Starting at the bottom of the recession, when economic recovery starts, most of these series start moving upward. In this upswing, the value of a series at one point in time is greater than its previous value. Thus there is a "momentum'' built into them, and it continues until something happens (e.g., increase in interest rate or taxes or both) to slow them down. Therefore, in regressions involving time series data, successive observations are likely to be interdependent.

2. **Specification Bias: Excluded Variables Case.** In empirical analysis the researcher often starts with a plausible regression model that may not be the most "perfect'' one. After the regression analysis, the researcher does the postmortem to find out whether the results accord with a priori expectations. If not, surgery is begun. For example, the researcher may plot the residuals e_i obtained from the fitted regression and may observe the patterns of the plots. These residuals (which are proxies for u_i) may suggest that some variables that were originally members but were not included in the model for a variety of reasons should be included. This is the case of excluded variable specification bias. Often the inclusion of such variables removes the correlation pattern observed among the residuals. For example, suppose we have the following demand model:

   $$\mathrm{Y_t} = \beta_1 + \beta_2 \mathrm{X_{2t}} + \beta_3 \mathrm{X_{3t}} + \beta_4 \mathrm{X_{4t}} + \mathrm{u_t} \qquad (18.1)$$

   where Y = quantity of beef demanded, X_2 = price of beef, X_3 = consumer income, X_4 = price of pork, and t = time. However, for some reason we run the following regression:

   $$\mathrm{Y_t} = \beta_1 + \beta_2 \mathrm{X_{2t}} + \beta_3 \mathrm{X_{3t}} + v_\mathrm{t} \qquad (18.2)$$

   Now if model (18.1) is the "correct'' model or "true" relation, running model (18.2) is tantamount to letting $v_\mathrm{t} = \beta_4 \mathrm{X_{4t}} + u_t$. And to the extent the price of pork affects the consumption of beef, the error or disturbance term v will reflect a systematic pattern, thus creating (false) autocorrelation. A simple test of this would be to run both models (18.1) and (18.2) and see whether autocorrelation, if any, observed in model (18.2) disappears when model (18.1) is run. The actual mechanics of detecting autocorrelation will be discussed in next lesson.

3. **Specification Bias: Incorrect Functional Form.** Suppose the true or correct model in a cost-output study is as follows:

   $$\text{Marginal cost}_i = \beta_1 + \beta_2 \text{output}_i + \beta_3 \text{output}_i^2 + \mathrm{u}_i \qquad (18.3)$$

   but we fit the following model:

   $$\text{Marginal cost}_i = \alpha_1 + \alpha_2 \text{output}_i + v_i \qquad (18.4)$$

The disturbance term v_i is, in fact, equal to β_3 $\text{output}^2 + u_i$, and hence will catch the systematic effect of the output^2 term on marginal cost. In this case, v_i will reflect autocorrelation because of the use of an incorrect functional form.

4. **Cobweb Phenomenon.** The supply of many agricultural commodities reflects the so-called Cobweb phenomenon, where supply reacts to price with a lag of one time period because supply decisions take time to implement (the gestation period). Thus, at the beginning of this year's planting of crops, farmers are influenced by the price prevailing last year, so that their supply function is

$$\text{supply}_t = \beta_1 + \beta_2 P_{t-1} + u_t \qquad (18.5)$$

Suppose at the end of period t, price P_t turns out to be lower than P_{t-1}. Therefore, in period $t+1$ farmers may very well decide to produce less than they did in period t. Obviously, in this situation the disturbances u_t are not expected to be random because if the farmers overproduce in year t, they are likely to reduce their production in $t+1$, and so on, leading to a Cobweb pattern.

5. **Lags.** In a time series regression of consumption expenditure on income, it is not uncommon to find that the consumption expenditure in the current period depends, among other things, on the consumption expenditure of the previous period. That is,

$$\text{Consumption}_t = \beta_1 + \beta_2 \text{income}_t + \beta_3 \text{consumption}_{t-1} + u_t \qquad (18.6)$$

A regression such as (18.6) is known as **autoregression** because one of the explanatory variables is the lagged value of the dependent variable. The rationale for a model such as (18.6) is simple. Consumers do not change their consumption habits readily for psychological, technological, or institutional reasons. Now if we neglect the lagged term in (18.6), the resulting error term will reflect a systematic pattern due to the influence of lagged consumption on current consumption.

6. **Manipulation of Data.** In empirical analysis, the raw data are often "manipulated''. For example, in time series regressions involving quarterly data, such data are usually derived from the monthly data by simply adding three monthly observations and dividing the sum by 3. This averaging introduces smoothness into the data by dampening the fluctuations in the monthly data. Therefore, the graph plotting the quarterly data looks much smoother than the monthly data, and this smoothness may itself lend to a systematic pattern in the disturbances, thereby introducing autocorrelation. Another source of manipulation is interpolation or extrapolation of data. For example, the Census of Population is conducted every 10 years in our country, the last being in 2010 and the one before that in 2000. Now if there is a need to obtain data for some year within the intercensus period2000–2010, the common practice is to interpolate on the basis of some adhoc assumptions. All such data "massaging'' techniques might impose upon the data a systematic pattern that might not exist in the original data.

7. **Data Transformation.** As an example of this, consider the following model

$$Y_t = \beta_1 + \beta_2 X_t + u_t \qquad (18.7)$$

where, say, Y = consumption expenditure and X = income. Since Eq. (18.7) holds true at every time period, it holds true also in the previous time period 't-1'. So, we can write Eq. (18.7) as

$$Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + u_{t-1} \tag{18.8}$$

Y_{t-1}, X_{t-1}, and u_{t-1} are known as the lagged values of Y, X, and u ,respectively, here lagged by one period. Now if we subtract Eq. (18.8) from Eq. (18.7), we obtain

$$\Delta Y_t = \beta_2 \Delta X_t + \Delta u_t \tag{18.9}$$

where Δ, known as the first difference operator**,** tells us to take successive differences of the variables in question. Thus, $\Delta Y_t = Y_t - Y_{t-1}$, $\Delta X_t = X_t - X_{t-1}$, and $\Delta u_t = u_t - u_{t-1}$. For empirical purposes, we write(18.9) as

$$\Delta Y_t = \beta_2 \Delta X_t + v_t \tag{18.10}$$

where $v_t = \Delta u_t = (u_t - u_{t-1})$.

Equation (18.8) is known as the level form and Eq. (18.9) is known as the (first) difference form. Both forms are often used in empirical analysis. For example, if in Eq. (18.8) Y and X represent the logarithms of consumption expenditure and income, then in Eq. (18.9) ΔY and ΔX will represent changes in the logs of consumption expenditure and income. But as we know, a change in the log of a variable is a relative change, or a percentage change, if the former is multiplied by 100. So, instead of studying relationships between variables in the level form, we may be interested in their relationships in the growth form.

Now if the error term in Eq. (18.7) satisfies the standard OLS assumptions, particularly the assumption of no autocorrelation, it can be shown that the error term v_t in Eq. (18.10) is autocorrelated. It may be noted here that models like Eq. (18.10) are known as dynamic regression models**,** that is, models involving lagged regressand.

The point of the preceding example is that sometimes autocorrelation may be induced as a result of transforming the original model.

18.3 OLS Estimation in the presence of Autocorrelation

What happens to the OLS estimators and their variances if we introduce autocorrelation in the disturbances by assuming that $E(u_t u_{t+s}) \neq 0 (\mathrm{s} \neq 0)$ but retain all the other assumptions of the classical model? Note again that we are now using the subscript t on the disturbances to emphasize that we are dealing with time series data.

We revert once again to the two-variable regression model to explain the basic ideas involved, namely, $\mathrm{Y}_t = \beta_1 + \beta_2 \mathrm{X_t} + u_t$. To make any headway, we must assume the mechanism that generates u_t, for $E(u_t u_{t+s}) \neq 0 (\mathrm{s} \neq 0)$ is too general an assumption to be of any practical use. As a starting point, or first approximation, one can assume that the disturbances are generated by the following mechanism

$$u_t = \rho u_{t-1} + \varepsilon_t, \quad -1<\rho<1 \tag{18.11}$$

where ρ is known as the **coefficient of autocovariance** and where ε_t is the stochastic disturbance term such that it satisfied the standard OLS assumptions, namely,

$$
\begin{aligned}
E(\varepsilon_t) &= 0 \\
\operatorname{var}(\varepsilon_t) &= \sigma_\varepsilon^2 \qquad (18.12) \\
\operatorname{cov}(\varepsilon_t \varepsilon_{t+s}) &= 0 \qquad s \neq 0
\end{aligned}
$$

The scheme in Eq. (18.11) is known as Markov first-order autoregressive scheme**,** or simply **a first-order autoregressive scheme,** usually denoted as **AR(1).** The name *autoregressive* is appropriate because Eq. (18.11) can be interpreted as the regression of u_t on itself lagged one period. It is first order because u_t and its immediate past value are involved; that is, the maximum lag is 1. If the model were $u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \varepsilon_t$, it would be an AR(2), o second-order, autoregressive scheme, and so on. The coefficient of autocovariance ρ in Eq. (18.11), can also be interpreted as the first-order coefficient of autocorrelation, or more accurately, **the coefficient of autocorrelation at lag 1.**

Given the AR(1) scheme, it can be shown that

$$\operatorname{var}(u_t) = E(u_t^2) = \frac{\sigma_\varepsilon^2}{\rho^2} \qquad (18.13)$$

$$\operatorname{cov}(u_t u_{t+s}) = E(u_t u_{t-s}) = \rho^s \frac{\sigma_\varepsilon^2}{1-\rho^2} \qquad (18.14)$$

$$\operatorname{cor}(u_t u_{t+s}) = \rho^s, \quad s=1,2,\ldots \qquad (18.15)$$

Note that because of the symmetry property of covariances and correlations, $\operatorname{cov}(u_t u_{t+s}) = \operatorname{cov}(u_t u_{t-s})$ and $\operatorname{cor}(u_t u_{t+s}) = \operatorname{cor}(u_t u_{t-s})$.

Since ρ is a constant between −1 and +1, Eq. (18.13) shows that under the AR(1) scheme, the variance of u_t is *still homoscedastic,* but u_t is correlated not only with its immediate past value but its values several periods in the past. It is *critical* to note that $|\rho| < 1$. If, for example $\rho = 1$, the variances and covariances listed above are not defined. If $|\rho| < 1$, then it is clear from Eq. (18.14) that the value of the covariance will decline as we go into the distant past.

One reason we use the AR(1) process is not only because of its simplicity compared to higher-order AR schemes, but also because in many applications it has proved to be quite useful. Additionally, a considerable amount of theoretical and empirical work has been done on the AR(1) scheme.

Now return to our two-variable regression model: $Y_t = \beta_1 + \beta_2 X_t + u_t$. We know that the OLS estimator of the slope coefficient is

$$\hat{\beta}_2 = \frac{\sum x_t y_t}{\sum x_t^2}, \quad \text{where } x_t = X_t - \bar{X} \textit{ and } y_t = Y_t - \bar{Y} \qquad (18.16)$$

and its variance is given by

$$\operatorname{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} \qquad (18.17)$$

Now under the AR(1) scheme, it can be shown that the variance of this estimator is:

$$\operatorname{var}\left(\hat{\beta}_2\right)_{AR1} = \frac{\sigma^2}{\sum x_t^2}\left[1 + 2\rho\frac{\sum x_t x_{t-1}}{\sum x_t^2} + 2\rho^2\frac{\sum x_t x_{t-2}}{\sum x_t^2} + \cdots + 2\rho^{n-1}\frac{\sum x_1 x_n}{\sum x_t^2}\right] \quad (18.18)$$

where $\operatorname{var}\left(\hat{\beta}_2\right)_{AR1}$ means the variance of $\hat{\beta}_2$ under first-order autoregressive scheme.

A comparison of Eq. (18.18) with Eq. (18.17) shows the former is equal to the latter times a term that depends on *ρ* as well as the sample autocorrelations between the values taken by the regressor X at various lags. And in general we cannot foretell whether $\operatorname{var}\left(\hat{\beta}_2\right)$ is less than or greater than $\operatorname{var}\left(\hat{\beta}_2\right)_{AR1}$. Of course, if $\rho = 0$, the two formulas will coincide. Also, if the correlations among the successive values of the regressor are very small, the usual OLS variance of the slope estimator will not be seriously biased. But, as a general principle, the two variances will not be the same.

To give some idea about the difference between the variances given in Eqs. (18.17) and (18.18), assume that the regressor X also follows the first-order autoregressive scheme with a coefficient of autocorrelation of *r*. Then it can be shown that Eq. (18.18) reduces to:

$$\operatorname{var}\left(\hat{\beta}_2\right)_{AR(1)} = \frac{\sigma^2}{\sum x_t^2}\left[\frac{1+r\rho}{1-r\rho}\right] = \operatorname{var}\left(\hat{\beta}_2\right)_{OLS}\left[\frac{1+r\rho}{1-r\rho}\right] \quad (18.19)$$

If, for example, *r* = 0.6 and *ρ* = 0.8, using Eq. (18.19) we can check that $\operatorname{var}\left(\hat{\beta}_2\right)_{AR1} = 2.8461\operatorname{var}\left(\hat{\beta}_2\right)_{OLS}$ .To put it another way, $\operatorname{var}\left(\hat{\beta}_2\right)_{OLS} = \frac{1}{2.8461}\operatorname{var}\left(\hat{\beta}_2\right)_{AR1}$ $= 0.3513\operatorname{var}\left(\hat{\beta}_2\right)_{AR1}$. That is, the usual OLS formula [i.e. Eq. (18.17)] will underestimate the variance of $\left(\hat{\beta}_2\right)_{AR1}$ by about 65 percent. The point of this exercise is to warn you that a blind application of the usual OLS formulae to compute the variances and standard errors of the OLS estimators could give seriously misleading results.

Suppose we continue to use the OLS estimator $\hat{\beta}_2$ and adjust the usual variance formula by taking into account the AR(1) scheme. That is, we use $\hat{\beta}_2$ given by Eq. (18.16) but use the variance formula given by Eq. (18.18). What now are the properties of $\hat{\beta}_2$? It is easy to prove that $\hat{\beta}_2$ is still linear and unbiased. As a matter of fact, the assumption of no serial correlation, like the assumption of no heteroscedasticity, is not required to prove that $\hat{\beta}_2$ is unbiased. Is $\hat{\beta}_2$ still BLUE? Unfortunately, it is in the class of linear unbiased estimators, but it does not have minimum variance. In short, $\hat{\beta}_2$, although linear unbiased, is not efficient. The student will notice that this finding is quite similar to the finding that $\hat{\beta}_2$ is less efficient in the presence of heteroscedasticity. There we saw that the weighted least-square estimator $\hat{\beta}_2^*$ studied in Section 17.3 of Lesson 17, a special case of the generalized least-squares (GLS) estimator, was efficient. In the case of autocorrelation can we find an estimator that is BLUE? The answer is yes, as can be seen from the discussion in the next lesson.

18.4 Consequences of using OLS in the presence of autocorrelation

As in the case of heteroscedasticity, in the presence of autocorrelation the OLS estimators are still linear unbiased as well as consistent and asymptotically normally distributed, but they are no longer efficient (i.e., minimum variance). What then happens to our usual hypothesis testing procedures if we continue to use the OLS estimators? Again, as in the case of heteroscedasticity, we distinguish two cases. We continue to work with the two-variable model, although the following discussion can be extended to multiple regressions without much trouble.

18.4.1 OLS estimation allowing for autocorrelation:

As noted, $\hat{\beta}_2$ is not BLUE, and even if we use $\operatorname{var}\left(\hat{\beta}_2\right)_{AR1}$, the confidence intervals derived from there are likely to be wider than those based on the GLS procedure. This result is likely to be the case even if the sample size increases indefinitely. That is, $\hat{\beta}_2$ *is not asymptotically efficient*. The implication of this finding for hypothesis testing is clear: We are likely to declare a coefficient statistically insignificant (i.e., not different from zero) even though in fact (i.e., based on the correct GLS procedure) it may be.

The message is: To establish confidence intervals and to test hypotheses, one should use GLS and not OLS even though the estimators derived from the latter are unbiased and consistent.

18.4.2 OLS estimation disregarding autocorrelation:

The situation is potentially very serious if we not only use $\hat{\beta}_2$ but also continue to use $\operatorname{var}\left(\hat{\beta}_2\right)=\sigma^2 \Big/ \sum x_t^2$, which completely disregards the problem of autocorrelation, that is, we mistakenly believe that the usual assumptions of the classical model hold true. Errors will arise for the following reasons:

1. The residual variance $\hat{\sigma}^2=\sum e_i^2 \Big/ (n-2)$ is likely to underestimate the true σ^2.
2. As a result, we are likely to under estimate RSS and hence to overestimate R^2.
3. Even if σ^2 is not underestimated, $\operatorname{var}\left(\hat{\beta}_2\right)$ may underestimate its variance under (first-order) autocorrelation $\operatorname{var}\left(\hat{\beta}_2\right)_{AR1}$ [given in Eq. (18.18)].
4. Therefore, the usual t and F tests of significance are no longer valid, and if applied, are likely to give seriously misleading conclusions about the statistical significance of the estimated regression coefficients.

To establish some of these propositions, let us revert to the two-variable model. We know that under the classical assumption

$$\hat{\sigma}^2=\sum e_i^2 \Big/ (n-2) \qquad (18.20)$$

provides an unbiased estimator of σ^2, that is, $E\left(\hat{\sigma}^2\right)=\sigma^2$. But, if there is autocorrelation, given by AR(1), it can be shown that

$$E\left(\hat{\sigma}^2\right)=\frac{\sigma^2\left[n-\left[2/(1-\rho)\right]-2\rho r\right]}{n-2} \qquad (18.21)$$

where $r = \sum_{t=1}^{n-1} x_t x_{t-1} \Big/ \sum_{t=1}^{n} x_t^2$, which can be interpreted as the (sample)correlation coefficient between successive values of the X's. If ρ and *r* are both positive (not an unlikely assumption for most economic time series), it is apparent from Eq. (18.21) that $E\left(\hat{\sigma}^2\right) < \sigma^2$; that is, the usual residual variance formula, on average, will underestimate the true σ^2. In other words, $\hat{\sigma}^2$ will be biased downward. Needless to say, this bias in σ^2 will be transmitted to $\operatorname{var}\left(\hat{\beta}_2\right)$ because in practice we estimate the latter by the formula $\hat{\sigma}^2 \Big/ \sum x_i^2$.

But even if $\hat{\sigma}^2$ is not underestimated, $\operatorname{var}\left(\hat{\beta}_2\right)$ is a *biased* estimator of $\operatorname{var}\left(\hat{\beta}_2\right)_{AR1}$, which can be readily seen by comparing Eq. (18.17) with Eq. (18.18),for the two formulas are not the same. As a matter of fact, if ρ is positive (which is true of most economic time series) and the X's are positively correlated (also true of most economic time series), then it is clear that

$$\operatorname{var}\left(\hat{\beta}_2\right) < \operatorname{var}\left(\hat{\beta}_2\right)_{AR1} \qquad (18.22)$$

that is, the usual OLS variance of $\hat{\beta}_2$ underestimates its variance under AR(1) [see Eq. (18.19)]. Therefore, if we use $\operatorname{var}\left(\hat{\beta}_2\right)$, we shall inflate the precision or accuracy (i.e., underestimate the standard error) of the estimator $\hat{\beta}_2$. As a result, in computing the t ratio as $t = \hat{\beta}_2 \Big/ SE\left(\hat{\beta}_2\right)$ (under H_0: β_2 = 0), we shall be overestimating the t value and hence the statistical significance of the estimated β_2. The situation is likely to get worse if additionally σ^2 is underestimated, as noted previously.

18.5 SUMMARY AND CONCLUSIONS

1. If the assumption of the classical linear regression model—that the errors or disturbances entering into the model are random or uncorrelated—is violated, the problem of serial or auto correlation arises.

2. Autocorrelation can arise for several reasons, such as inertia or sluggishness of economic time series, specification bias resulting from excluding important variables from the model or using incorrect functional form, the cobweb phenomenon, data massaging, and data transformation.

3. Although in the presence of autocorrelation the OLS estimators remain unbiased, consistent, and asymptotically normally distributed, they are no longer efficient. As a consequence, the usual *t*, *F* and χ^2 tests cannot be legitimately applied. Hence, remedial measures may be called for and are discussed in the next lesson.

18.6 Self Assessment Questions

1. Define the concept of Auto correlation with reference to a two-variable linear model with first order-Auto regression scheme.

2. Distinguish between simple correlation and serial correlation.
3. Define the concept of autocorrelation with reference to a two-variable linear model with first order-autoregressive scheme.
4. Explain Auto-correlation with suitable examples.
5. Explain various sources of auto correlation.
6. Define Auto Correlation. Explain the problem of auto Correlation in two variable linear model.
7. Explain the consequences of autocorrelation if we apply OLS estimation method.
8. Explain the consequences if we apply OLS estimation method disregarding autocorrelation.

18.7 References

1. Gujarati, D.N. (2005): *Basic Econometrics, 4th Ed., Tata McGraw-Hill.*
2. Johnston, J. (1984): *Econometric Methods, 3rd Ed., McGraw-Hill,* New York.
3. Montgomery, D.C., Peck, E.A. and Geoffrey Vining, G. (2003): *Introduction to Linear Regression Analysis, 3rd Ed., Wiley*
4. Draper, N.R., and H. Smith(1998): *Applied Regression Analysis, 3rd Ed., John Wiley & Sons,* New York.
5. G.S. Maddala (2001): *Introduction to Econometrics, 3rd Ed., John Wiley & Sons, Ltd.*
6. Johnston, J. and DiNardo J (1997): *Econometric Methods*, 4th Ed., McGraw Hill.
7. Hill, Carter, William Griffiths, and George Judge(2001): *Undergraduate Econometrics, John Willey & Sons,* New York.
8. *Koutsoyiannis, A(1973): Theory of Econometrics, Harper & Row,* New York.

Lesson 19

AUTOCORRELATION – DETECTION AND REMEDIES

19.0 Objective:

This lesson is continuation of Lesson 18 and after studying this lesson, the student will be familiarized with some popular detection methods such as Durbin-Watson d test as well as some remedies of autocorrelation.

Structure of the Lesson:

19.1 Introduction

Recall that the assumption of no autocorrelation of the classical linear regression model (CLRM) relates to the population disturbances u_t, which are not directly observable. Therefore, how does one know that there is autocorrelation in any given situation? How does one remedy the problem of autocorrelation? Instead of the unobservable disturbances, we have their proxies, the residuals $e_t's$, which can be obtained by the usual OLS procedure. Although the $e_t's$ are not the same thing as $u_t's$, very often a visual examination of the $e_t's$ gives us some clues about the likely presence of autocorrelation in the u's. Actually, a visual examination of $e_t's$ or ($e_t^2's$) can provide useful information not only about autocorrelation but also about heteroscedasticity. In Section 19.2, we study some popular methods of the detection of autocorrelation while in Section 19.3, we explain some estimation methods in the presence of autocorrelation. In Section 19.4, we present the brief summary and conclusions.

19.2 Detection of Autocorrelation

1. Graphical Method

The importance of producing and analyzing plots of [residuals] as a standard part of statistical analysis cannot be overemphasized. Besides occasionally providing an easy to understand summary of a complex problem, they allow the simultaneous examination of the data as an aggregate while clearly displaying the behavior of individual cases.

There are various ways of examining the residuals. We can simply plot them against time, the **time sequence plot.** Alternatively, we can plot the standardized residuals against time. The standardized residuals are simply the residuals (e_t's) divided by the standard error of the regression $(\hat{\sigma})$, that is, they are $(e_t/\hat{\sigma})$. Notice that the residuals e_t's and $\hat{\sigma}$ are measured in the units in which the regress and Y is measured. The values of the standardized residuals will therefore be pure numbers (devoid of units of measurement) and can be compared with the standardized residuals of other regressions. Moreover, the standardized residuals, like e_t's, have zero mean and *approximately* unit variance. In large samples $(e_t/\hat{\sigma})$ is approximately normally distributed with zero mean and unit variance.

The graphical method we have just discussed, although powerful and suggestive, is subjective or qualitative in nature. But there are several quantitative tests that one can use to supplement the purely qualitative approach. We now consider some of these tests.

2. Durbin–Watson *d* Test

The most celebrated test for detecting autocorrelation is that developed by statisticians Durbin and Watson. It is popularly known as the **Durbin–Watson *d* statistic,** which is computed from the vector of OLS residuals $\underset{\sim}{\mathbf{e}} = \underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\hat{\boldsymbol{\beta}}}$. It is denoted in the literature variously as d or DW and is defined as

$$\boxed{d = \frac{\sum_{t=2}^{n}\left(e_t - e_{t-1}\right)^2}{\sum_{t=1}^{n} e_t^2}} \qquad (19.1)$$

which is simply the ratio of the sum of squared differences in successive residuals to the RSS. Note that in the numerator of the d statistic the number of observations is $n-1$ because one observation is lost in taking successive differences. A great advantage of the d statistic is that it is based on the estimated residuals, which are routinely computed in regression analysis. Because of this advantage, it is now a common practice to report the Durbin–Watson *d* along with summary measures, such as R^2, adjusted R^2, t, and F. Although it is now routinely used, it is important to note the assumptions underlying the d statistic.

1. The regression model includes the intercept term. If it is not present, as in the case of the regression through the origin, it is essential to rerun the regression including the intercept term to obtain the RSS.
2. The explanatory variables, the X's are nonstochastic, or fixed in repeated sampling.

3. The disturbances u_t's are generated by the first-order autoregressive scheme, $u_t = \rho u_{t-1} + \varepsilon_t$. Therefore, it cannot be used to detect higher-order autoregressive schemes.
4. The error term u_t is assumed to be normally distributed.
5. The regression model does not include the lagged value(s) of the dependent variable as one of the explanatory variables. Thus, the test is inapplicable in **autoregressive models**.

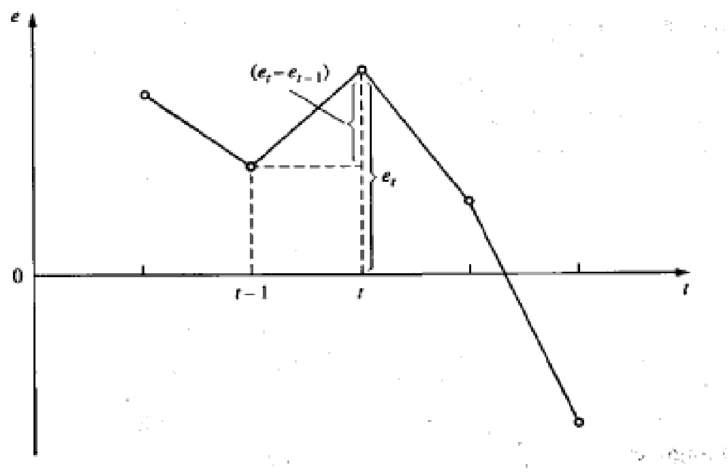


Figure 19.1 (a): Positive auto correlation

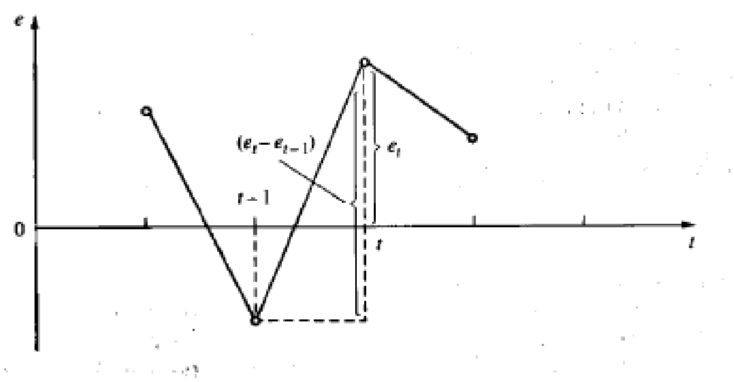


Figure 19.1 (b): Negative auto correlation

The above figures indicate why d might be expected to measure the extent of first-order autocorrelation. The mean of the residuals is zero, so the residuals will be scattered around the horizontal axis. If the e's are positively auto correlated, successive values will tend to be close to each other, runs above and below the horizontal axis will occur, and the first differences will tend to be numerically smaller than the residuals themselves. Alternatively, if the e's have a first-order negative autocorrelation, there is a tendency for successive observations to be on opposite sides of the horizontal axis so that first differences tend to be numerically larger than the residuals. Thus d will tend to be "small" for positively auto (serial) correlated e's and "large" for negatively auto (serial) correlated e's. If the e's are random, we have an in-between situation with no tendency for runs above and below the axis or for alternate swings across it and d will take on an intermediate value.

The Durbin-Watson statistic is closely related to the sample first order autocorrelation coefficient of the e's. Expanding Eq. (19.1), we have

$$d = \frac{\sum_{t=2}^{n} e_t^2 + \sum_{t=2}^{n} e_{t-1}^2 - 2\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2} \tag{19.2}$$

Since $\sum_{t=2}^{n} e_t^2$ and $\sum_{t=2}^{n} e_{t-1}^2$ differ in only one observation, they are approximately equal to $\sum_{t=1}^{n} e_t^2$. Therefore, Eq. (19.2) may be rewritten as

$$d \simeq 2(1-\hat{\rho}) \tag{19.3}$$

where $\hat{\rho} = \sum_{t=2}^{n} e_t e_{t-1} \Big/ \sum_{t=1}^{n} e_t^2$ is the coefficient in the OLS regression of e_t on e_{t-1}. Ignoring end-point discrepancies, $\hat{\rho}$ is seen to be the simple correlation coefficient between e_t and e_{t-1} and hence, $-1 \le \hat{\rho} \le 1$. Thus, Eq. (19.3) implies that

$$0 \le d \le 4 \tag{19.4}$$

that is the range of d is from 0 to 4 and as well as we have the following:

d < 2 for positive autocorrelation of the e's
d > 2 for negative autocorrelation of the e's
d = 2 for zero autocorrelation of the e's

It is also apparent from Eq. (19.3) that

- if $\hat{\rho} = 0,\ d \approx 2$; that is, if there is no serial correlation (of the first-order), d is expected to be about 2. *Therefore, as a rule of thumb, if d is found to be closer to 2 in an application, one may assume that there is no first-order autocorrelation, either positive or negative.*
- If $\hat{\rho} = +1$, indicating perfect positive correlation in the residuals, $d \approx 0$. Therefore, the closer d is to 0, the greater the evidence of positive autocorrelation.
- If $\hat{\rho} = -1$, that is, there is perfect negative correlation among successive residuals, $d \approx 4$. Hence, the closer d is to 4, the greater the evidence of negative autocorrelation.

The exact sampling or probability distribution of the *d* statistic given in Eq. (19.1) is difficult to derive because, as Durbin and Watson have shown, it depends in a complicated way on the $\mathbf{X}$ matrix of the given sample. This difficulty should be understandable because *d* is computed from $e_t's$, which are, of course, dependent on the given $\mathbf{X}$ matrix. Therefore, unlike the t, F, or χ^2 tests, there is no unique critical value that will lead to the rejection or the acceptance of the null hypothesis that there is no first-order autocorrelation in the disturbances $u_t's$. However, Durbin and Watson were successful in deriving a lower bound d_L and an upper bound d_U such that if the computed d from Eq. (19.1) lies outside these critical values, a decision can be made regarding the presence of positive or negative auto (serial) correlation.

Moreover, these limits depend only on the number of observations *n* and the number of explanatory variables (k-1) and do not depend on the values taken by these explanatory variables. These limits, for *n* going from 6 to 200 and up to 20 explanatory variables, have been tabulated by Durbin and Watson.

The mechanics of the Durbin–Watson test are as follows, assuming that the assumptions underlying the test are fulfilled:

1. Run the OLS regression for the given data and obtain the residuals.
2. Compute d from Eq. (19.1). (Most computer programs now do this routinely.)
3. For the given sample size and given number of explanatory variables, find out the critical d_L and d_U values.
4. Now follow the decision rules given below.

The testing procedure is as follows:
Set the null hypothesis **H_0: zero auto correlation**
If d<=2: Set the alternative hypothesis **H_1: positive first-order auto correlation**

Decision Rules:
1. If $d < d_L$, reject the null hypothesis H_0 in favor of the alternative hypothesis of H_1.
2. If $d > d_U$, do not reject the null hypothesis.
3. If $d_L < d < d_U$, the test is inconclusive.

(19.5)

If d>2: Set the alternative hypothesis **H_1: negative first-order auto correlation**
Decision Rules: Replace d with 4- d and follow the above decision rules.

Remark: Even when the conditions for the validity of the Durbin-Watson test are satisfied, the inconclusive range is an awkward problem, especially as it becomes fairly large at low degrees of freedom. A conservative practical procedure is to use d_U as if it were a conventional critical value and simply reject the null hypothesis if $d < d_U$. The consequences of accepting H_0 when autocorrelation is present are almost certainly more serious than the consequences of incorrectly presuming its presence.

Note: The student is advised to refer any text book on Econometrics for Durbin and Watson d-tables of d_L and d_U constructed at 1% and 5% levels of significance.

Illustration 19.1

The following table gives data on indexes of real compensation per hour (Y) and output per hour(*X*) in the business sector of the U.S. economy for the period 1980–1997, the base of the indexes being 1992 = 100. For this data examine for the presence of auto correlation using Durbin-Watson d-test.

Table 19.1: INDEXES OF REAL COMPENSATION AND PRODUCTIVITY, UNITED STATES, 1980–1997

YEAR	Y_t	X_t	YEAR	Y_t	X_t
1980	89.7	79.8	1989	95.8	93.3

1981	89.8	81.4	1990	96.4	94.5
1982	91.1	81.2	1991	97.4	95.9
1983	91.2	84.0	1992	100.0	100.0
1984	91.5	86.4	1993	99.9	100.1
1985	92.8	88.1	1994	99.7	101.4
1986	95.9	90.7	1995	99.1	102.2
1987	96.3	91.3	1996	99.6	105.2
1988	97.3	92.4	1997	101.1	107.5

Source: Economic Report of the President, 2000, Table B-47, p.362

Solution:

From the above data we have

n = 18 $\sum X^2$ = 157172.00

$\sum X$ =1675.40 $\sum Y^2$ = 165488.30

$\sum Y$ =1724.60 $\sum XY$ = 161060.40

$$\hat{\beta}_2 = \left(\sum XY - n\bar{X}\bar{Y}\right) / \left(\sum X^2 - n\bar{X}^2\right) = 0.4379$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 55.0485$$

Estimated regression equation: $\hat{Y}_t = 55.0485 + 0.4379\mathrm{X}_t$

Computation of Durbin-Watson d-statistic:

t	Y_t	X_t	$\hat{Y}_t = \hat{\alpha} + \hat{\beta}\mathrm{X}_t$	$e_t = Y_t - \hat{Y}_t$	e_{t-1}	$e_t - e_{t-1}$	$(e_t - e_{t-1})^2$	e_t^2
1	89.7	79.8	89.9962	-0.2962	--	--	---	0.0878
2	89.8	81.4	90.6969	-0.8969	-0.2962	-0.6007	0.3608	0.8045
3	91.1	81.2	90.6093	0.4907	-0.8969	1.3876	1.9254	0.2407
4	91.2	84.0	91.8356	-0.6356	0.4907	-1.1262	1.2684	0.4040
5	91.5	86.4	92.8866	-1.3866	-0.6356	-0.7511	0.5641	1.9228
6	92.8	88.1	93.6311	-0.8311	-1.3866	0.5555	0.3086	0.6908
7	95.9	90.7	94.7698	1.1302	-0.8311	1.9614	3.8469	1.2774
8	96.3	91.3	95.0325	1.2675	1.1302	0.1372	0.0188	1.6064
9	97.3	92.4	95.5143	1.7857	1.2675	0.5183	0.2686	3.1888
10	95.8	93.3	95.9084	-0.1084	1.7857	-1.8941	3.5878	0.0118
11	96.4	94.5	96.4340	-0.0340	-0.1084	0.0745	0.0055	0.0012
12	97.4	95.9	97.0471	0.3529	-0.0340	0.3869	0.1497	0.1246
13	100.0	100.0	98.8426	1.1574	0.3529	0.8044	0.6471	1.3395
14	99.9	100.1	98.8864	1.0136	1.1574	-0.1438	0.0207	1.0273
15	99.7	101.4	99.4558	0.2442	1.0136	-0.7693	0.5919	0.0597
16	99.1	102.2	99.8061	-0.7061	0.2442	-0.9504	0.9032	0.4986
17	99.6	105.2	101.1199	-1.5199	-0.7061	-0.8138	0.6623	2.3102
18	101.1	107.5	102.1272	-1.0272	-1.5199	0.4927	0.2428	1.0551
SUMS	1724.6	1675.4	1724.6000	0.0000			15.3726	16.6509

From the above table we have

$$\sum_{t=2}^{n}\left(e_t - e_{t-1}\right)^2 = 15.3726 \text{ and } \sum_{t=1}^{n} e_t^2 = 16.6509$$

Now substituting these values in Eq. (19.1), we get Durbin-Watson statistic value

d=0.9232

Let us set the null hypothesis

H_0: zero autocorrelated disturbances

Since, D-W statistic value d<2, let us set the alternative hypothesis

H_1: positive first-order autocorrelated disturbances

Conclusion drawn:

From Durbin-Watson d-tables, the critical d-values at 5% l.o.s. are

d_L=1.158 and d_U=1.391

Since, the calculated d-value (0.9232) is less than the critical d_L value, by applying the decision rules given in (19.5), we reject H_0. Thus, there is a problem of autocorrelation in the given data. Hence, we conclude that the estimated regression model using the above wages-productivity data yields the first order positive auto correlated residuals. Therefore, the estimated model

$$\hat{Y}_t = 55.0485 + 0.4379\mathrm{X}_\mathrm{t} \qquad (19.6)$$

is not the correct estimated model and we have to estimate it using a different estimation method which we will discuss in the next section.

3. The Wallis Test for Fourth-order Autocorrelation

Wallis has pointed out that many applied studies employ quarterly data, and in such cases one might expect to find fourth-order autocorrelation in the disturbance term. The appropriate specification is then

$$u_t = \rho_4 u_{t-4} + \varepsilon_t \qquad (19.7)$$

To test the null hypothesis, $H_0 : \rho_4 = 0$, Wallis proposes a modified Durbin-Watson statistic,

$$\boxed{d_4 = \frac{\sum_{t=5}^{n}\left(e_t - e_{t-4}\right)^2}{\sum_{t=1}^{n} e_t^2}} \qquad (19.8)$$

where the e's are the usual OLS residuals. Wallis derives upper and lower bounds for d_4 under the assumption of a nonstochastic $\mathbf{X}$ matrix.

4. Durbin's *h*-test for a Regression Model with Lagged Dependent Variables

We know that the Durbin-Watson test procedure was derived under the assumption of a non-stochastic $\mathbf{X}$ matrix, which is violated by the presence of lagged values of the dependent variable among the regressors. Durbin has derived a large-sample (asymptotic) test for the more general case. It is still a test against first-order autocorrelation, and one must specify the complete set of regressors. Consider the relation,

$$y_t = \beta_1 y_{t-1} + \cdots + \beta_r y_{t-r} + \beta_{r+1} x_{1t} + \cdots + \beta_{r+s} X_{st} + u_t \tag{19.9}$$

with $u_t = \rho u_{t-1} + \varepsilon_t \qquad |\rho| < 1$ and $\boldsymbol{\varepsilon} \sim N\left(0, \sigma_{\boldsymbol{\varepsilon}}^2 \mathbf{I}\right)$

Durbin's basic result is that under the null hypothesis $H_0 : \rho = 0,$ the statistic

$$\boxed{h = \hat{\rho}\sqrt{\frac{n}{1 - n\,\mathrm{var}\left(\hat{\beta}_1\right)}} \overset{asy}{\sim} N\left(0,1\right)} \tag{19.10}$$

where n = sample size

$\mathrm{var}\left(\hat{\beta}_1\right)$ = estimated sampling variance of the coefficient of y_{t-1} in the OLS regression of

Eq. (19.9) $\hat{\rho} = \sum_{t=2}^{n} e_t e_{t-1} \Big/ \sum_{t=2}^{n} e_{t-1}^2,$ the estimate of ρ from the regression of e_t on e_{t-1}, the e's in turn being the residuals from the OLS regression of Eq. (19.9).

The test procedure is as follows.

1. Fit the OLS regression denoted by Eq. (19.9) and note $\mathrm{var}\left(\hat{\beta}_1\right)$.
2. From the residuals compute $\hat{\rho}$ or, alternatively, if the Durbin-Watson statistic has been computed, we may use the approximation $\hat{\rho} \simeq 1 - d/2$.
3. Substitute $\mathrm{var}\left(\hat{\beta}_1\right)$ and $\hat{\rho}$ in Eq. (19.10) to obtain h, and if h >1.645, reject the null hypothesis at 5 percent level of significance in favor of the hypothesis of a positive first-order autocorrelation.
4. A similar one-sided test for negative autocorrelation can be carried out for negative h.

5. The Breusch–Godfrey (BG) Test for Higher Order Autocorrelation

To avoid some of the pitfalls of the Durbin–Watson *d* test of autocorrelation, statisticians Breusch and Godfrey have developed a test of autocorrelation that is general in the sense that it allows for (1) nonstochastic regressors, such as the lagged values of the regressand; (2) higher-order autoregressive schemes, such as AR(1), AR(2), etc.; and (3) simple or higher-order moving averages of white noise error terms.

We use the two-variable regression model to illustrate the **BG test**, which is also known as the **LM test,** although many regressors can be added to the model. Also, lagged values of the regressand can be added to the model. Let

$$Y_t = \beta_1 + \beta_2 X_t + u_t \tag{19.11}$$

Assume that the error term u_t follows the p^{th} order autoregressive, AR(p), scheme as follows:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \ldots + \rho_p u_{t-p} + \varepsilon_t \tag{19.12}$$

where ε_t is a white noise error term as discussed previously. As you will recognize, this is simply the extension of AR(1) scheme.

The null hypothesis H_0 to be tested is that

$$H_0 : \rho_1 = \rho_2 = \ldots = \rho_p = 0 \tag{19.13}$$

That is, there is no serial correlation of any order. The BG test involves the following steps:

1. Estimate (19.11) by OLS and obtain the residuals e_t's.
2. Regress e_t on the original X_t (if there is more than one X variable in the original model, include them also) and $e_{t-1}, e_{t-2}, \ldots, e_{t-p}$, where the latter are the lagged values of the estimated residuals in step 1. Thus, if *p* = 4, we will introduce four lagged values of the residuals as additional regressors in the model. In short, run the following regression

$$e_t = \alpha_1 + \alpha_2 X_t + \rho_1 e_{t-1} + \rho_2 e_{t-2} + \ldots + \rho_p e_{t-p} + \varepsilon_t \tag{19.14}$$

and obtain R^2 from this (auxiliary) regression. Since there are only *n* values of *e* available, this regression might be carried out using only the last (*n* − *p*) observations.

3. If the sample size *n* is large (technically, infinite), Breusch and Godfrey have shown that

$$\boxed{(n - p)R^2 \sim \chi^2_p} \tag{19.15}$$

That is, asymptotically, $(n-p)\,R^2$ value obtained from the auxiliary regression (19.14) follows the chi-square distribution with *p* d.f.. If in an application, $(n-p)R^2$ exceeds the critical chi-square value at the chosen level of significance, we reject the null hypothesis H_0 [Eq. (19.13)], in which case at least one of $\rho_1, \rho_2, \ldots, \rho_p$ is statistically significantly different from zero.

The following *practical points* about the BG test may be noted:

1. The regressors included in the regression model may contain lagged values of the regressand Y, that is, Y_{t-1}, Y_{t-2}, etc., may appear as explanatory variables. Contrast this model with the Durbin–Watson test restriction that there are no lagged values of the regressand among the regressors.

2. As noted earlier, the BG test is applicable even if the disturbances follow a *p*th order moving average (MA)process, that is, the u_t are generated as follows:

$$u_t = \varepsilon_t + \lambda_1 \varepsilon_{t-1} + \lambda_2 \varepsilon_{t-2} + \ldots + \lambda_p \varepsilon_{t-p} \tag{19.16}$$

where ε_t is a white noise error term, that is, the error term that satisfies all the classical assumptions.

3. If in Eq. (19.12) p = 1, meaning first-order auto regression, then the BG test is known as Durbin's *M*-test.

4. A drawback of the BG test is that the value of *p*, the length of the lag, cannot be specified a priori.

19.3 Estimation of Relationships with Autocorrelated Disturbances

Knowing the consequences of autocorrelation, especially the lack of efficiency of OLS estimators, we may need to remedy the problem. If one or more of the diagnostic tests described in the previous section suggest autocorrelated disturbances, then we have to apply one of the following estimation methods.

1. The method of Generalized Least Squares(GLS):

Consider the two-variable regression model

$$Y_t = \beta_1 + \beta_2 X_t + u_t \tag{19.17}$$

and assume that the error term u_t follows the AR(*1*) scheme, namely,

$$u_t = \rho u_{t-1} + \varepsilon_t \ , \qquad -1<\rho<1 \tag{19.18}$$

If we replace *t=t-1* in Eq. (19.17), we get

$$Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + u_{t-1} \tag{19.19}$$

Multiplying Eq. (19.19) with ρ on both sides, we obtain

$$\rho Y_{t-1} = \rho\beta_1 + \rho\beta_2 X_{t-1} + \rho u_{t-1} \tag{19.20}$$

Subtracting Eq. (19.20) from Eq. (19.19) and using Eq. (19.18), we get

$$Y_t - \rho Y_{t-1} = \beta_1(1-\rho) + \beta_2(X_t - \rho X_{t-1}) + \varepsilon_t \tag{19.21}$$

We can express Eq. (19.21) as

$$\boxed{\begin{aligned} & Y_t^* = \beta_1^* + \beta_2 X_t^* + \varepsilon_t \\ & \text{where} \\ & Y_t^* = Y_t - \rho Y_{t-1} \\ & X_t^* = X_t - \rho X_{t-1} \\ & \beta_1^* = \beta_1(1-\rho) \end{aligned}} \tag{19.22}$$

since the error term ε_t in Eq. (19.18) satisfies the usual OLS assumptions, we can apply OLS to the transformed variables Y_t^* and X_t^*; and obtain estimators with all the optimum properties of BLUE. Here, it may be noted applying OLS to Eq. (19.22) is equivalent to the application of generalized least squares (GLS).

Regression equation (19.22) is known as the generalized difference equation. In this equation we lose one observation because the first observation has no antecedent. Although conceptually straightforward to apply, the method of generalized difference in Eq. (19.22) is difficult to implement because ρ is rarely known in practice. Therefore, we need to find ways of estimating ρ and we have given below some of the methods.

ρ based on Durbin–Watson *d* statistic:

We have an easy method of estimating ρ from the relationship between *d* and ρ established previously in Eq. (19.3), from which we can estimate ρ as follows.

$$\hat{\rho} \approx 1 - d/2 \tag{19.23}$$

Thus, in reasonably large samples one can obtain ρ from (19.23) and use it to transform the data as shown in the generalized difference equation (19.22). Keep in mind that the relationship between ρ and *d* given in (19.23) may not hold true in small samples.

ρ estimated from the residuals:

If the AR(1) scheme

$$u_t = \rho u_{t-1} + \varepsilon_t \ , \qquad -1 < \rho < 1$$

is valid, a simple way to estimate ρ is to regress the residuals e_t on e_{t-1}, for the e_t's are consistent estimators of the true u_t, as noted previously. That is, we run the following regression

$$e_t = \rho e_{t-1} + v_t \tag{19.24}$$

where e_t's are the residuals obtained from the original regression and where v_t's are the error term of this regression. Note that there is no need to introduce the intercept term in Eq. (19.24), for we know the OLS residuals sum is zero.

Iterative methods of estimating *ρ*:

All the methods of estimating ρ discussed previously provide us with only a single estimate of ρ. But there are the so-called iterative methods that estimate ρ iteratively, that is, by successive approximation, starting with some initial value of ρ. Among these methods the following may be mentioned: the **Cochrane–Orcutt iterative procedure,** the **Cochrane–Orcutt two-step method,** the **Durbin two–step method.** Of these, the most popular is the Cochran–Orcutt iterative method. Remember that the ultimate objective of these methods is to provide an estimate of ρ that may be used to obtain GLS estimates of the parameters. One advantage of the Cochrane–Orcutt iterative method is that it can be used to estimate not only an AR(1) scheme, but also higher-order auto regressive schemes, such as $e_t = \rho_1 e_{t-1} + \rho_2 e_{t-2} + v_t$, which is AR(2). Having obtained ρ_1 and ρ_2, one can easily extend the generalized difference equation (19.22).

2. The Cochrane–Orcutt (C–O) iterative procedure.

Consider the two-variable regression model

$$Y_t = \beta_1 + \beta_2 X_t + u_t \tag{19.25}$$

and assume that the error term u_t follows the AR(*1*) scheme, namely,

$$u_t = \rho u_{t-1} + \varepsilon_t \ , \qquad -1 < \rho < 1 \tag{19.26}$$

where the error terms ε_t's are well-behaved.

Now, the above model can be rearranged in two equivalent forms as

$$Y_t - \rho Y_{t-1} = \beta_1(1-\rho) + \beta_2(X_t - \rho X_{t-1}) + \varepsilon_t \qquad (19.27)$$

Or

$$Y_t - \beta_1 - \beta_2 X_t = \rho\left(Y_{t-1} - \beta_1 - \beta_2 X_{t-1}\right) + \varepsilon_t \qquad (19.28)$$

If ρ is known in Eq. (19.27), the unknown parameters β_1 and β_2 could be estimated by applying OLS straightforwardly. Similarly, if β_1 and β_2 are known, ρ could be estimated by applying OLS to the regression equation (19.28).

It is important to note that the above two equivalent forms of equations (19.27) and (19.28) are the basis for Cochrane-Orcutt iterative method for estimation of model (19.25) with AR(1) scheme. Starting with any value for ρ, the quasi first differences in the generalized difference equation (19.27) could be computed, and OLS applied to it would then yield estimates of β_1 *and* β_2. These estimates in turn can be used to compute the $Y_t - \beta_1 - \beta_2 X_t$ series. Regressing this series on itself lagged on period in Eq. (19.28) yields a revised estimate of ρ, which can then be fed back into Eq. (19.27), and the iteration process continues until a satisfactory degree of convergence is reached.

Various steps of Cochrane-Orcutt iterative method of estimation of model (19.25) are as follows.

Step 1: Estimate the two variable model (19.25) by applying standard OLS technique and obtain the estimates of β_1 *and* β_2 denoted by $\hat{\beta}_1$ *and* $\hat{\beta}_2$.

Step 2: Now, compute the residuals given by

$$e_t = Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_t\ , \quad t = 1, 2, \ldots, n.$$

and run the regression

$$e_t = \hat{\rho} e_{t-1} + \hat{\varepsilon}_t$$

which is the estimated equation of (19.28). Here $\hat{\rho}$ is OLS estimate of ρ given by

$$\boxed{\hat{\rho} = \sum_{t=2}^{n} e_t e_{t-1} \Big/ \sum_{t=2}^{n} e_{t-1}^2} \qquad (19.29)$$

Step 3: Using the above estimated $\hat{\rho}$, compute the quasi first differences given by

$$Y_t^* = Y_t - \hat{\rho} Y_{t-1} \text{ and } X_t^* = X_t - \hat{\rho} X_{t-1}$$

Now, the generalized difference equation (19.27) with $\rho = \hat{\rho}$ can be written as

$$Y_t^* = \beta_1^* + \beta_2 X_t^* + \varepsilon_t\ , \quad t = 2, 3, \ldots, n \qquad (19.30)$$

where $\beta_1^* = \beta_1\left(1 - \hat{\rho}\right)$. (19.31)

Now by applying OLS to Eq. (19.30) we can get the estimates of β_1 [from the relation (19.31)] and β_2. Now, denote these second round estimates of β_1 *and* β_2 by $\hat{\hat{\beta}}_1$ *and* $\hat{\hat{\beta}}_2$.

Step 4: Now, repeat the above steps (2) and (3) by replacing $\hat{\beta}_1$ *and* $\hat{\beta}_2$ with the above 2nd round estimates $\hat{\hat{\beta}}_1$ *and* $\hat{\hat{\beta}}_2$ respectively to compute 3rd round estimates $\hat{\hat{\hat{\beta}}}_1$ *and* $\hat{\hat{\hat{\beta}}}_2$.
Repeat the above steps (2)-(4) until the successive estimates of the parameter ρ differ by less than some prescribed amount.

Illustration 19.2: Consider the data, given in Illustration 19.1, on indexes of real compensation (wages) and productivity in the business sector of the U.S. economy for the period 1980–1997. We have already seen that in Illustration 19.1, the OLS method is not a suitable estimation method, since autocorrelation is presented in the data. Hence, let us re-estimate the model using Cochrane-Orcutt iterative method.

Solution:

STEP 1: Estimating the model $Y_t = \beta_1 + \beta_2 X_t + u_t$ using OLS method:

From the data of illustration 1, we have

n = 18 $\sum X^2$ = 157172.00

$\sum X$ =1675.40 $\sum Y^2$ = 165488.30

$\sum Y$ =1724.60 $\sum XY$ = 161060.40

$\hat{\beta}_2 = \left(\sum XY - n\overline{X}\overline{Y}\right) / \left(\sum X^2 - n\overline{Y}^2\right) = 0.4379$

$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X} = 55.0485$

Estimated regression equation: $\hat{Y}_t = 55.0485 + 0.4379\mathrm{X}_t$

FIRST ITERATION :

STEP 2: COMPUTATION OF OLS RESIDUALS AND HENCE ESTIMATING ρ

t	Y_t	X_t	$\hat{Y}_t = \hat{\beta}_1 + \hat{\beta}_2 \mathrm{X}_t$	$e_t = Y_t - \hat{Y}_t$	e_{t-1}	$e_t * e_{t-1}$	e_{t-1}^2
1	89.7	79.8	89.9962	-0.2962	--	--	--
2	89.8	81.4	90.6969	-0.8969	-0.2962	0.2657	0.0878
3	91.1	81.2	90.6093	0.4907	-0.8969	-0.4401	0.8045
4	91.2	84.0	91.8356	-0.6356	0.4907	-0.3119	0.2407
5	91.5	86.4	92.8866	-1.3866	-0.6356	0.8813	0.4040
6	92.8	88.1	93.6311	-0.8311	-1.3866	1.1525	1.9228
7	95.9	90.7	94.7698	1.1302	-0.8311	-0.9394	0.6908
8	96.3	91.3	95.0325	1.2675	1.1302	1.4325	1.2774
9	97.3	92.4	95.5143	1.7857	1.2675	2.2633	1.6064
10	95.8	93.3	95.9084	-0.1084	1.7857	-0.1936	3.1888
11	96.4	94.5	96.4340	-0.0340	-0.1084	0.0037	0.0118
12	97.4	95.9	97.0471	0.3529	-0.0340	-0.0120	0.0012
13	100.0	100.0	98.8426	1.1574	0.3529	0.4085	0.1246
14	99.9	100.1	98.8864	1.0136	1.1574	1.1731	1.3395
15	99.7	101.4	99.4558	0.2442	1.0136	0.2476	1.0273

16	99.1	102.2	99.8061	-0.7061	0.2442	-0.1725	0.0597
17	99.6	105.2	101.1199	-1.5199	-0.7061	1.0732	0.4986
18	101.1	107.5	102.1272	-1.0272	-1.5199	1.5613	2.3102
SUMS	1724.6	1675.4	1724.6000	0.0000		8.3932	15.5958

From the above table we have

$$\sum_{t=2}^{n} e_t e_{t-1} = 8.3932 \text{ and } \sum_{t=2}^{n} e_{t-1}^2 = 15.5958$$

Now substituting these values in Eq. (19.29) we get

$$\hat{\rho} = 0.5382$$

STEP 3: REESTIMATING β_1 and β_2 by estimating $Y_t^* = \beta_1^* + \beta_2 X_t^* + \varepsilon_t$ using OLS method

t	Y_t	X_t	Y_{t-1}	X_{t-1}	$Y_t^* = Y_t - \hat{\rho} Y_{t-1}$	$X_t^* = X_t - \hat{\rho} X_{t-1}$
1	89.7	79.8	--	--	--	--
2	89.8	81.4	89.7	79.8	41.5261	38.4541
3	91.1	81.2	89.8	81.4	42.7723	37.3930
4	91.2	84.0	91.1	81.2	42.1727	40.3006
5	91.5	86.4	91.2	84.0	42.4189	41.1937
6	92.8	88.1	91.5	86.4	43.5574	41.6021
7	95.9	90.7	92.8	88.1	45.9578	43.2873
8	96.3	91.3	95.9	90.7	44.6895	42.4880
9	97.3	92.4	96.3	91.3	45.4742	43.2651
10	95.8	93.3	97.3	92.4	43.4360	43.5731
11	96.4	94.5	95.8	93.3	44.8433	44.2888
12	97.4	95.9	96.4	94.5	45.5204	45.0430
13	100.0	100.0	97.4	95.9	47.5822	48.3895
14	99.9	100.1	100.0	100.0	46.0830	46.2830
15	99.7	101.4	99.9	100.1	45.9368	47.5292
16	99.1	102.2	99.7	101.4	45.4444	47.6296
17	99.6	105.2	99.1	102.2	46.2673	50.1991
18	101.1	107.5	99.6	105.2	47.4983	50.8845
				Totals	761.1800	751.8000

n* = 17 $\qquad \sum Y^{*2} = 34135.850$

$\sum Y^* = 761.18 \qquad \sum X^{*2} = 33489.070$

$\sum X^* = 751.80 \qquad \sum X^* Y^* = 33762.686$

$\hat{\beta}_2 = 0.4157$

$\hat{\beta}_1^* = 26.3919 \Rightarrow \quad \hat{\beta}_1 = \hat{\beta}_1^* / (1 - \hat{\rho}) = 57.1464$

SECOND ITERATION:

STEP 2: RECOMPUTING THE RESIDUALS USING LATEST $\hat{\beta}_1$ $and\hat{\beta}_2$ AND HENCE REESTIMATING ρ

t	Y_t	X_t	$\hat{Y}_t = \hat{\beta}_1 + \hat{\beta}_2 X_t$	$e_t = Y_t - \hat{Y}_t$	e_{t-1}	$e_t * e_{t-1}$	e_{t-1}^2
1	89.7	79.8	90.3186	-0.6186	--	--	--
2	89.8	81.4	90.9837	-1.1837	-0.6186	0.7322	0.3827
3	91.1	81.2	90.9006	0.1994	-1.1837	-0.2361	1.4011
4	91.2	84.0	92.0645	-0.8645	0.1994	-0.1724	0.0398
5	91.5	86.4	93.0622	-1.5622	-0.8645	1.3505	0.7474
6	92.8	88.1	93.7688	-0.9688	-1.5622	1.5135	2.4403
7	95.9	90.7	94.8496	1.0504	-0.9688	-1.0176	0.9386
8	96.3	91.3	95.0990	1.2010	1.0504	1.2614	1.1033
9	97.3	92.4	95.5563	1.7437	1.2010	2.0941	1.4423
10	95.8	93.3	95.9304	-0.1304	1.7437	-0.2274	3.0405
11	96.4	94.5	96.4293	-0.0293	-0.1304	0.0038	0.0170
12	97.4	95.9	97.0112	0.3888	-0.0293	-0.0114	0.0009
13	100.0	100.0	98.7156	1.2844	0.3888	0.4994	0.1511
14	99.9	100.1	98.7571	1.1429	1.2844	1.4679	1.6498
15	99.7	101.4	99.2975	0.4025	1.1429	0.4600	1.3061
16	99.1	102.2	99.6301	-0.5301	0.4025	-0.2133	0.1620
17	99.6	105.2	100.8772	-1.2772	-0.5301	0.6770	0.2810
18	101.1	107.5	101.8333	-0.7333	-1.2772	0.9365	1.6311
					Totals	9.1180	16.7350

From the above table we have

$$\sum_{t=2}^{n} e_t e_{t-1} = 9.1180 \text{ and } \sum_{t=2}^{n} e_{t-1}^2 = 16.7350$$

Now substituting these values in Eq. (19.29) we get

$\hat{\rho} = 0.5448$

STEP 3: REESTIMATING β_1 $and\beta_2$ by estimating $Y_t^* = \beta_1^* + \beta_2 X_t^* + \varepsilon_t$ using OLS method

t	Y_t	X_t	Y_{t-1}	X_{t-1}	$Y_t^* = Y_t - \hat{\rho} Y_{t-1}$	$X_t^* = X_t - \hat{\rho} X_{t-1}^*$
1	89.7	79.8	--	--	--	--
2	89.8	81.4	89.7	79.8	40.9272	37.9212
3	91.1	81.2	89.8	81.4	42.17272	36.8494
4	91.2	84.0	91.1	81.2	41.56442	39.7584
5	91.5	86.4	91.2	84.0	41.80993	40.6328
6	92.8	88.1	91.5	86.4	42.94648	41.0252
7	95.9	90.7	92.8	88.1	45.33817	42.699

8	96.3	91.3	95.9	90.7	44.04915	41.8824
9	97.3	92.4	96.3	91.3	44.83121	42.6554
10	95.8	93.3	97.3	92.4	42.78636	42.9561
11	96.4	94.5	95.8	93.3	44.20363	43.6658
12	97.4	95.9	96.4	94.5	44.87672	44.4119
13	100.0	100.0	97.4	95.9	46.93188	47.7491
14	99.9	100.1	100.0	100.0	45.41527	45.6153
15	99.7	101.4	99.9	100.1	45.26976	46.8608
16	99.1	102.2	99.7	101.4	44.77873	46.9525
17	99.6	105.2	99.1	102.2	45.60564	49.5166
18	101.1	107.5	99.6	105.2	46.83321	50.1821
				Totals	750.4100	741.4100

n* = 17 $\quad \sum Y^{*2} =$ 33177.60

$\sum Y^{*}$ = 750.41 $\quad \sum X^{*2} =$ 32570.00

$\sum X^{*}$ = 741.41 $\quad \sum X^{*}Y^{*} =$ 32825.12

$\hat{\beta}_2$ = 0.4153

$\hat{\beta}_1^{*}$ = 26.0283 $\Rightarrow \hat{\beta}_1 = \hat{\beta}_1^{*}/(1-\hat{\rho}) = 57.18588$

THIRD ITERATION:

STEP 2: RECOMPUTING THE RESIDUALS USING LATEST $\hat{\beta}_1\ and\,\hat{\beta}_2$ AND HENCE REESTIMATING ρ

t	Y_t	X_t	$\hat{Y}_t = \hat{\alpha}+\hat{\beta}\mathrm{X}_t$	$e_t = Y_t - \hat{Y}_t$	e_{t-1}	$e_t * e_{t-1}$	e_{t-1}^2
1	89.7	79.8	90.3250	-0.6250	--	--	--
2	89.8	81.4	90.9895	-1.1895	-0.6250	0.7434	0.3906
3	91.1	81.2	90.9064	0.1936	-1.1895	-0.2303	1.4148
4	91.2	84.0	92.0692	-0.8692	0.1936	-0.1683	0.0375
5	91.5	86.4	93.0658	-1.5658	-0.8692	1.3610	0.7555
6	92.8	88.1	93.7718	-0.9718	-1.5658	1.5217	2.4519
7	95.9	90.7	94.8515	1.0485	-0.9718	-1.0189	0.9444
8	96.3	91.3	95.1007	1.1993	1.0485	1.2574	1.0993
9	97.3	92.4	95.5575	1.7425	1.1993	2.0898	1.4383
10	95.8	93.3	95.9313	-0.1313	1.7425	-0.2287	3.0363
11	96.4	94.5	96.4296	-0.0296	-0.1313	0.0039	0.0172
12	97.4	95.9	97.0110	0.3890	-0.0296	-0.0115	0.0009
13	100.0	100.0	98.7136	1.2864	0.3890	0.5004	0.1513
14	99.9	100.1	98.7551	1.1449	1.2864	1.4727	1.6548
15	99.7	101.4	99.2950	0.4050	1.1449	0.4637	1.3107
16	99.1	102.2	99.6272	-0.5272	0.4050	-0.2135	0.1640

17	99.6	105.2	100.8731	-1.2731	-0.5272	0.6712	0.2780
18	101.1	107.5	101.8282	-0.7282	-1.2731	0.9270	1.6207
					Totals	9.1410	16.7661

From the above table we have

$\sum_{t=2}^{n} e_t e_{t-1}$ =9.1410 and $\sum_{t=2}^{n} e_{t-1}^2$ =16.7661

Now substituting these values in Eq. (19.29) we get

$\hat{\rho} = 0.5452$

STEP 3: REESTIMATING β_1 and β_2 by estimating $Y_t^* = \beta_1^* + \beta_2 X_t^* + \varepsilon_t$ using OLS method

t	Y_t	X_t	Y_{t-1}	X_{t-1}	$Y_t^* = Y_t - \hat{\rho} Y_{t-1}$	$X_t^* = X_t - \hat{\rho} X_{t-1}^*$
1	89.7	79.8	--	--	--	--
2	89.8	81.4	89.7	79.8	45.4207	32.4956
3	91.1	81.2	89.8	81.4	46.8298	32.2410
4	91.2	84.0	91.1	81.2	45.4032	34.3323
5	91.5	86.4	91.2	84.0	44.3947	36.6778
6	92.8	88.1	91.5	86.4	44.7679	38.2142
7	95.9	90.7	92.8	88.1	46.4504	40.1054
8	96.3	91.3	95.9	90.7	46.5232	39.0153
9	97.3	92.4	96.3	91.3	46.9235	39.8972
10	95.8	93.3	97.3	92.4	44.9328	40.2520
11	96.4	94.5	95.8	93.3	44.8786	42.2698
12	97.4	95.9	96.4	94.5	45.1153	43.3427
13	100.0	100.0	97.4	95.9	45.4800	46.8975
14	99.9	100.1	100.0	100.0	45.3255	45.5800
15	99.7	101.4	99.9	100.1	44.4167	46.9345
16	99.1	102.2	99.7	101.4	43.3806	47.8436
17	99.6	105.2	99.1	102.2	42.2450	51.1707
18	101.1	107.5	99.6	105.2	42.4910	53.1981
				Totals	749.7700	740.7800

n* = 17 $\sum Y^{*2} =$ 33120.2

$\sum Y^*$ = 749.77 $\sum X^{*2} =$ 32514.77

$\sum X^*$ = 740.78 $\sum X^* Y^* =$ 32768.91

$\hat{\beta}_2$ = 0.4153

$\hat{\beta}_1^*$ = 26.009 $\Rightarrow$ $\hat{\beta}_1 = \hat{\beta}_1^* / (1 - \hat{\rho}) = 57.1878$

From the above computations, we may notice that the corresponding values of $\hat{\rho},\ \hat{\beta}_1\ and\ \hat{\beta}_2$ *at SECOND and THIRD iterations are approximately equal*. Hence, we take the *values computed at THIRD iteration as the* estimates of $\hat{\rho},\ \hat{\beta}_1\ and\ \hat{\beta}_2$ and are given by

$$\hat{\rho} = 0.5452$$

$$\hat{\beta}_1 = 57.1878$$

$$\hat{\beta}_2 = 0.4153$$

Thus the estimated regression model for the given wages-productivity data, using Cochrane-Orcutt iterative method, is obtained as

$$\hat{Y}_t = 57.1878 + 0.4153\ X_t$$

Compare this estimated regression model with the regression model (19.6), which is estimated using OLS method, ignoring the presence of auto correlation. We may note that this model is different from the model (19.6).

3. The Cochrane–Orcutt two-step method.

This is a shortened version of the Cochrane-Orcutt iterative procedure. In step 1, we estimate ρ from the first iteration, and in step 2 we use that estimate of ρ to run the generalized difference equation. Sometimes in practice, this two-step method gives results quite similar to those obtained from the more elaborate Cochrane-Orcutt iterative procedure. Let us explain this method for k-variable regression model.

Consider the general linear regression model with k-1 explanatory variables $X_{2t}, X_{3t}, ..., X_{kt}$

$$Y_t = \beta_1 + \sum_{i=2}^{k} \beta_i X_{it} + u_t, \qquad t = 1, 2, ..., n \tag{19.32}$$

and assume that the error term u_t follows the AR(*1*) scheme, namely,

$$u_t = \rho u_{t-1} + \varepsilon_t\ , \qquad -1 < \rho < 1 \tag{19.33}$$

where the error terms ε_t's are well-behaved.

The above model can be rearranged in two equivalent forms as

$$Y_t - \rho Y_{t-1} = \beta_1(1-\rho) + \sum_{i=2}^{k} \beta_i (X_{it} - \rho X_{i(t-1)}) + \varepsilon_t \tag{19.34}$$

Or
$$Y_t - \beta_1 - \sum_{i=2}^{k} \beta_i X_{it} = \rho\left(Y_{t-1} - \beta_1 - \sum_{i=2}^{k} \beta_i X_{i(t-1)} \right) + \varepsilon_t \tag{19.35}$$

Now, the two steps of Cochrane-Orcutt two-step method of estimation of model (19.32) are as follows.

Step 1: Estimate the general linear model (19.32) by applying standard OLS technique and obtain the estimates of $\beta_1, \beta_2, ..., \beta_k$ denoted by $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$ respectively. Now, compute the residuals,

$$e_t = Y_t - \hat{\beta}_1 - \sum_{i=2}^{k} \hat{\beta}_i X_{it}\ , \quad t = 1, 2, \ldots, n. \tag{19.36}$$

and run the regression

$$e_t = \hat{\rho} e_{t-1} + \hat{\varepsilon}_t$$

which is the estimated equation of (19.35). Here $\hat{\rho}$ is the OLS estimate of ρ and by definition

$$\hat{\rho} = \sum_{t=2}^{n} e_t e_{t-1} \Big/ \sum_{t=2}^{n} e_{t-1}^2 \qquad (19.37)$$

Step 2: Using the above estimated $\hat{\rho}$, compute the quasi first differences given by

$$Y_t^* = Y_t - \hat{\rho} Y_{t-1} \ \text{ and } X_{it}^* = X_{it} - \hat{\rho} X_{i(t-1)}, \ \ for \ i = 2,3,...,k$$

Now, the generalized difference equation (19.34) with $\rho = \hat{\rho}$ can be written as

$$\boxed{Y_t^* = \beta_1^* + \sum_{i=2}^{k} \beta_i X_{it}^* + \varepsilon_t, \quad t = 2,3,\ldots,n} \qquad (19.38)$$

where $\beta_1^* = \beta_1 (1 - \hat{\rho})$

Now by applying OLS to the above equation we get the estimates of $\beta_1 \ (from \ the \ estimate \ of \ \beta_1^*), \beta_2, \beta_3, ..., \beta_k$. Now, if we denote these revised estimates of $\beta_1, \beta_2, ..., \beta_k$ by $b_1, b_2, ..., b_k$, then the estimated regression model, using Cochrane-Orcutt two-step method, is given by

$$\hat{Y}_t = b_1 + \sum_{i=2}^{k} b_i X_{it}, \qquad t = 1,2,...,n \qquad (19.39)$$

Exercise to the students: Apply the above Cochrane–Orcutt two-step method to the wages–productivity example discussed in illustration 2, and compare your results with those obtained from the Cochrane–Orcutt iterative method.

4. Durbin's two-step method.

Let us explain this method for k-variable regression model.

Consider the general linear regression model with k-1 explanatory variables $X_{2t}, X_{3t}, ..., X_{kt}$

$$Y_t = \beta_1 + \sum_{i=2}^{k} \beta_i X_{it} + u_t, \qquad t = 1,2,...,n \qquad (19.40)$$

and assume that the error term u_t follows the AR(*1*) scheme, namely,

$$u_t = \rho u_{t-1} + \varepsilon_t, \qquad -1 < \rho < 1 \qquad (19.41)$$

where the error terms ε_t s are well-behaved.

The above model can be rearranged in the generalized difference equation form as

$$Y_t - \rho Y_{t-1} = \beta_1 (1-\rho) + \sum_{i=2}^{k} \beta_i (X_{it} - \rho X_{i(t-1)}) + \varepsilon_t \qquad (19.42)$$

which may be rewritten as

$$Y_t = \beta_1 (1-\rho) + \rho Y_{t-1} + \sum_{i=2}^{k} \beta_i X_{it} - \sum_{i=2}^{k} \beta_i^* X_{i(t-1)} + \varepsilon_t \qquad (19.43)$$

Durbin suggests the following two-step procedure for estimation of the model (19.40).

Step 1: Treat (19.43) as a multiple regression model, regressing Y_t on the 2k-1 variables Y_{t-1}, $X_{2t}, X_{3t}, ..., X_{kt}$, $X_{2(t-1)}, X_{3(t-1)}, ..., X_{k(t-1)}$ and treat the estimated value of the regression coefficient of Y_{t-1} (= $\hat{\rho}$) as an estimate of ρ.

Step 2: Using the above estimated $\hat{\rho}$, compute the quasi first differences given by

$$Y_t^* = Y_t - \hat{\rho} Y_{t-1} \text{ and } X_{it}^* = X_{it} - \hat{\rho} X_{i(t-1)}, \; for \; i = 2,3,...,k$$

and the generalized difference equation (19.42) with $\rho = \hat{\rho}$ can be written as

$$\boxed{Y_t^* = \beta_1 \left(1 - \hat{\rho}\right) + \sum_{i=2}^{k} \beta_i X_{it}^* + \varepsilon_t \, , \quad t = 2,3,\ldots,n} \tag{19.44}$$

Now, by applying OLS to the above equation we get the estimates of $\beta_1, \beta_2, ..., \beta_k$, which are denoted by $b_1, b_2, ..., b_k$. Thus, the estimated regression model, using Durbin's two-step method, is

$$\hat{Y}_t = b_1 + \sum_{i=2}^{k} b_i X_{it}, \quad t = 1,2,...,n \tag{19.45}$$

Exercise to the students: Apply the Durbin's two-step method to the wages–productivity example discussed in illustration 2, and compare your results with those obtained from the Cochrane–Orcutt iterative procedure and the Cochrane–Orcutt two-step method.

Remark: As explained in Cochrane–Orcutt two step method and Durbin's two-step method we can also explain Cochrane–Orcutt iteration method for the case of k-variable general linear model.

19.4 Summary and Conclusions

1. The remedy depends on the nature of the interdependence among the disturbances. But since the disturbances are unobservable, the common practice is to assume that they are generated by some mechanism.

2. The mechanism that is commonly assumed is the Markov first-order autoregressive scheme, which assumes that the disturbance in the current time period is linearly related to the disturbance term in the previous time period, the coefficient of autocorrelation ρ providing the extent of the interdependence. This mechanism is known as the AR(1) scheme.

3. If the AR(1) scheme is valid and the coefficient of autocorrelation is known, the serial correlation problem can be easily tackled by transforming the data following the generalized difference procedure. The AR(1) scheme can be easily generalized to an AR(*p*).

4. Even if we use an AR(1) scheme, the coefficient of autocorrelation is not known a priori. We considered several methods of estimating ρ, such as the Durbin–Watson *d*,

Cochrane–Orcutt iterative procedure, Cochrane–Orcutt two-step method, and the Durbin two-step method.

5. In large samples, these methods generally yield similar estimates of ρ, although in small samples they perform differently. In practice, the Cochrane–Orcutt iterative method has become quite popular.

6. Using any of the methods just discussed, we can use the generalized difference method to estimate the parameters of the transformed model by OLS, which essentially amounts to GLS. But since we estimate ρ (= $\hat{\rho}$), we call the method of estimation as feasible or estimated GLS, or FGLS or EGLS for short.

7. Of course, before remediation comes detection of autocorrelation. Among the methods of detection, Durbin–Watson *d* test and Breusch–Godfrey (BG) test are the popular and routinely used is the Durbin–Watson *d* test. It is better to use the BG test, for it is much more general in that it allows for both AR and MA error structures as well as the presence of lagged regressand as an explanatory variable. But keep in mind that it is a large sample test.

19.5 Self Assessment Questions

1. Explain how the plot of residuals against the regressand variable will be useful for detecting serial correlation.
2. Explain in detail how various residual plots are useful in checking the standard assumptions on the error terms in a linear regression model.
3. Explain the role of residuals plots in regression analysis.
4. Explain various estimation procedures of a model, briefly, in the presence of autocorrelation.
5. Explain the procedure for Durbin-Watson test to detect the Autocorrelation.
6. Explain Durbin-Watson test for detection of serial correlation in a regression model and discuss the limitations of the test.
7. Describe the Durbin-Watson test of serial correlation and discuss merits and demerits of the test.
8. Describe the Wallis Test for examining fourth-order autocorrelation.
9. Describe the Breusch–Godfrey (BG) test of (higher order) serial (auto) correlation.
10. Describe Durbin's *h*-test for a regression model with lagged dependent variables.
11. Explain the Cochrane-Orcutt iterative method for estimation of the parameters in a simple linear model in the presence of autocorrelated disturbances.
12. Explain the two-step Cochrane-Orcutt method for estimating the parameters of a linear model in the presence of autocorrelated disturbances.
13. Describe the two-step Durbin's method for estimating the parameters of a linear model in the presence of autocorrelated disturbances.
14. Explain any one of estimation procedures in the case of auto Correlation. Give Durbin-Watson test for Auto Correlation.

19.6 References

1. Gujarati, D.N. (2005): *Basic Econometrics, 4th Ed., Tata McGraw-Hill.*

2. Johnston, J. (1984): *Econometric Methods, 3rd Ed., McGraw-Hill,* New York.

3. Montgomery, D.C., Peck, E.A. and Geoffrey Vining, G. (2003): *Introduction to Linear Regression Analysis, 3rd Ed., Wiley*

4. Draper, N.R., and H. Smith(1998): *Applied Regression Analysis, 3rd Ed., John Wiley & Sons,* New York.

5. G.S. Maddala (2001): *Introduction to Econometrics, 3rd Ed., John Wiley & Sons, Ltd.*

6. Johnston, J. and DiNardo J (1997): *Econometric Methods*, 4th Ed., McGraw Hill.

7. Hill, Carter, William Griffiths, and George Judge(2001): *Undergraduate Econometrics, John Willey & Sons,* New York.

8. Koutsoyiannis, A(1973): Theory of Econometrics, Harper & Row, New York.

Lesson 20

QUALITATIVE RESPONSE REGRESSION MODELS

20.0 Objective:

In this lesson we consider several models in which the dependent variable (regressand) itself is qualitative in nature. After studying this lesson, the student will be familiar with some popular binary regression models namely Logit and Probit models, which are increasingly used in areas of social sciences and medical research.

Structure of the Lesson:

20.1 Introduction

In all the regression models that we have considered so far, we have implicitly assumed that the regressand/dependent/response variable Y is quantitative, whereas the explanatory variables are either quantitative, qualitative or mixture of those two variables. But there are several practical problems or illustrations, where a dependent variable or response variable will be a dummy variable which take two or more values.

Qualitative response regression models are widely being used in areas of social sciences and medical research. However these models pose interesting estimation and interpretation challenges.

20.2 The Nature of Qualitative Response Models

Consider an example of Indian Parliamentary elections assume that there are two political parties Congress and BJP. The dependent variable here is vote choice between the two political parties. Suppose we let

Y=1, if vote is for a Congress candidate.
=0, if vote is for a BJP candidate.

Here in this example the explanatory (cause) variables used in the vote choice are unemployment, inflation rates, present ruling party, caste of the candidate etc. Here in this example it may be noted that the regressed is a qualitative variable.

One can think of several other examples (listed below) where the dummy dependent variable is qualitative in nature.

1. A family either owns a house or it does not.
2. A family either owns a car or it does not.
3. Both husband and wife are government employees or only one spouse.
4. A certain drug is effective in curing an illness or it does not.
5. A firm decides to declare a stock dividend or not.

We don't have to restrict our response variable yes or no dichotomous or binary categories only. Suppose in the parliamentary elections there are many parties are contesting for instance in Andhra Pradesh in Guntur, the candidates pertaining to Congress, TDP, BJP, CPI, and CPM. Then the dependent variable takes 1, 2, 3, 4, and 5. This is the case where the dependent variable is polychotomous (multiple category) response variable. In particular, if the response variable take three categories then it is called as trichotomous variable.

In polychotomous regression models, the regressand may be either ordinal (i.e. an ordered categorical variable such as education example: school, college, university) or the regressand is nominal where there is no inherent ordering such as religion (Hindu, Muslim, Sikh, Christian)

There are some other qualitative response regression models where the response variable may be

1. The number of visits to one's physician per year.
2. The number of patents received by a firm in a given year.
3. The number of articles published by a university professor in a year.
4. The number of telephone calls received in a span of 5 minutes.
5. The number of cars passing through a toll gate in a span of 5 minutes.

These are the examples of Poisson probability regression models. There are three approaches to develop a probability model for the above binary response variable or polychotomous response variable and they are

1. The linear probability model (LPM).
2. The LOGIT model (logistic regression model)
3. The PROBIT model.

It is important to note that the fundamental difference between a traditional regression model and qualitative response regression model is that in the traditional regression model the regressand Y is quantitative, where as in the qualitative response regression model, the regressand Y is qualitative. In a model where Y is quantitative, our objective is to estimate the expected or mean value given the values of the regressors as given below.

$$Y_i = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i \tag{20.1}$$

$$E\left(Y_i / X_{2i}, X_{3i}, ..., X_{ki}\right) = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji}$$

But, in the models where Y is qualitative, our objective is to find the probability of something happening such as voting for a Congress candidate or owning a house or a certain drug is effective etc., Hence qualitative response regression models are often known as probability models.

We have to seek the answers for the following questions:

1. How do we estimate qualitative response regression models? Can we estimate simply with the usual OLS procedure?
2. Are there any special inference problems.
3. How to measure the goodness of fit?
4. How do we estimate and interpret polychotomous regression models ? Also how do we handle models in which the regressand is ordinal or nominal?

20.3 The Linear Probability Model (LPM)

For the sake of simplicity, we consider the regression model with only one explanatory variable X as given below

$$Y_i = \beta_1 + \beta_2 X_i + u_i, \quad \text{with } E(u_i) = 0, \quad i = 1, 2, \dots, n$$

where Y_i = 1, if i^{th} family have own house.

= 0, if i^{th} family does not have own house.

X_i = Income of i^{th} family.

(20.2)

The above model is a typical linear regression model because the regressand Y_i is a binary variable/dichotomous variable and it is called a linear probability model[LPM].

From Eq. (20.2) we may write

$$E\left(Y_i / X_i\right) = \beta_1 + \beta_2 X_i \tag{20.3}$$

Let us denote

$$P_i = P(\text{the event of the i}^{\text{th}} \text{ family have } \textit{own} \text{ house}) \qquad = P(Y_i = 1)$$

$$1 - P_i = P(\text{the event of the i}^{\text{th}} \text{ family } \textit{does not} \text{ have } \textit{own} \text{ house}) = P(Y_i = 0)$$

Thus the variable Y_i has the following probability distribution

Y_i	$P(Y_i)$
0	$1 - P_i$
1	P_i
Total	1

that is Y_i follows the Bernoulli probability distribution. Now, by the definition of mathematical expectation

$$E(Y_i) = \sum Y_i P(Y_i) = 0(1 - P_i) + 1\, P_i = P_i \tag{20.4}$$

From Eqs. (20.3) and (20.4) we may write

$$E(Y_i/X_i) = \beta_1 + \beta_2 X_i = E(Y_i) = P_i$$

that is the conditional expectation of the model (20.2) can in fact be interpreted as the conditional probability of Y_i. Since the probability P_i must lie between 0 and 1, we have the restriction

$$0 \le E(Y_i/X_i) = \beta_1 + \beta_2 X_i \le 1 \tag{20.5}$$

Thus, $\beta_1 + \beta_2 X_i$ is interpreted as the probability that the event will occur at given the X_i. The calculated value of Y_i from the regression model (20.2) that is $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ will then give the estimated probability that the event will occur given the particular value of X_i. The above LPM posses the following problems.

1. Non-Normality of the disturbances:

From equation (20.2) we may write the disturbance term u_i as

$$u_i = Y_i - \beta_1 - \beta_2 X_i \qquad i = 1, 2, \ldots, n$$

Thus the probability distribution of u_i is

	u_i	$P(u_i)$
when Y_i=1	$1-\beta_1 - \beta_2 X_i$	$P_i = \beta_1 + \beta_2 X_i$
when Y_i=0	$-\beta_1 - \beta_2 X_i$	$1-P = 1-\beta_1 - \beta_2 X_i$

Obviously u_i cannot be assumed to be normally distributed as it follows the Bernoulli distribution and hence, there is a problem with the application of the usual tests of significance.

However, in large sample any distribution approaches normal distribution and hence, we can assume u_i as normal in large samples.

2. Heteroscedasticity of the disturbances:

Even if $E(u_i) = 0$ *and* $\text{cov}(u_i, u_j) = 0 \;\forall\; i \neq j$ (that is no auto correlation), the LPM disturbances are not homoscedastic as explained below.

Since the probability distribution of u_i is Bernoulli, we have

$$\begin{aligned} V(u_i) &= P(u_i)\big(1 - P(u_i)\big) \\ &= P_i(1 - P_i) \\ &= E(Y_i/X_i)\big(1 - E(Y_i/X_i)\big) \\ &= (\beta_1 + \beta_2 X_i)(1 - \beta_1 - \beta_2 X_i) \qquad \text{(using Eq.(20.3))} \\ &= W_i\,(say) \end{aligned} \tag{20.6}$$

Since $V(u_i)$ is depending on X_i which varies from one individual to other, $V(u_i)$ is not common/constant for all u_i's i.e., $V(u_i)$ is not homoscedastic. Thus u_i's are heteroscedastic disturbances.

Dividing Eq. (20.2) by $\sqrt{W_i}$ we get

$$Y_i^* = \beta_1 Z_i + \beta_2 X_i^* + u_i^*, \quad i = 1, 2, \ldots, n \tag{20.7}$$

where $Y_i^* = \dfrac{Y_i}{\sqrt{W_i}},\; Z_i = \dfrac{1}{\sqrt{W_i}},\; X_i^* = \dfrac{X_i}{\sqrt{W_i}},\; u_i^* = \dfrac{u_i}{\sqrt{W_i}}$ (20.8)

$$V(u_i^*) = V\left(\frac{u_i}{\sqrt{W_i}}\right) = \frac{V(u_i)}{W_i} = \frac{W_i}{W_i} = 1$$

Now, we can verify from the equation Eq. (20.8), $V(u_i^*) = 1$ and hence u_i^*'s are homoscedastic. We can estimate the model (20.7) by applying ordinary least square (OLS) method. But for the application of OLS method the practical problem is Z_i and Y_i^*, and X_i^* are in terms of W_i, where W_i is in terms of the unknown parameter β_1 *and* β_2.

To overcome this problem we have to estimate W_i using the following two step procedure.

Step 1: Run the OLS regression for equation Eq. (20.2). Despite the heteroscedasticity problem and obtain $\hat{Y}_i$=estimate of true $E(Y_i/X_i)$. Then obtain

$\hat{W}_i = \hat{Y}_i\left(1 - \hat{Y}_i\right)$ where $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_1 X_i$

Step 2: Use the estimated W_i to obtain Y_i^*, Z_i and X_i^* from Eq. (20.8) and apply OLS method to the transformed model (20.7) and it may be noted that the transformed model does not have the intercept.

3. Non fulfillment of $0 \le E(Y_i/X_i) \le 1$:

Since $E(Y_i/X_i)$ in the LPM measures the conditional probability of the event Y_i occurring given X_i. It must necessarily lie between 0 and 1. But there is no guarantee that $\hat{Y}_i$, the estimate of $E(Y_i/X_i)$ will necessarily fulfill this restriction and this is the real and practical problem with the OLS estimation of the LPM.

There are two ways of finding out whether the estimated $\hat{Y}_i$ lie between 0 and 1. One is to estimate the LPM by the usual OLS method and find out whether the estimated $\hat{Y}_i$ lie between 0 and 1. If $\hat{Y}_i < 0$, for some i, make $\hat{Y}_i = 0$ and if $\hat{Y}_i > 1$, for some i, make $\hat{Y}_i = 1$.

The second procedure is to devise an estimating technique that will guarantee that the estimated probability $\hat{Y}_i$ will lie between 0 and 1. The Logit and Probit model discussed later will guarantee that the estimated probabilities will indeed lie between the logical limits 0 and 1.

Note: the traditional R^2 cannot be used to measure the goodness of fit in case of probability models.

20.4 The Logit Model

An explanation of the LOGIT model (logistic regression) begins with the explanation of the logistic function

$$F(Z) = \frac{1}{1+e^{-z}}, \qquad -\infty < z < \infty \tag{20.9}$$

The logistic function is useful because it can take as an input any value from $-\infty$ to $+\infty$, where as the output is confined to the values between 0 and 1. Thus Eq. (20.9) is a probability function called logistic probability model. The variable Z is usually defined as

$$Z = \beta_1 + \beta_2 X_2 + \cdots + \beta_k X_k \; = \beta_1 + \sum_{j=2}^{k} \beta_j X_j \tag{20.10}$$

where $X_2, X_3, \cdots, X_k$ are *k-1* explanatory variables or cause variables or regressors and β_1 is the intercept and $\beta_2, \beta_3, \cdots, \beta_k$ are regression coefficients.

It is easy to verify that as Z varies to $-\infty$ to $+\infty$, $F(Z)$ ranges from 0 to 1. Thus $F(Z)$ represents the probability of a particular outcome, given that a set of factor variables. A positive β_j means that the variable X_j increases the probability of outcome, while a negative β_j means the variable X_j decreases the probability of outcome.

Now let us consider a situation where a dependent variable or response variable is a dichotomous that is which takes 0 or 1, which depends on several factors namely $X_2, X_3, ..., X_k$. Now our objective is to find

$$E\left(Y_i / X_{2i}, X_{3i}, ..., X_{ki}\right)$$

Since Y_i takes only two values,

Y_i=1, if the event occurs

=0, other wise

Y_i is Bernoulli variable and by definition we have

$$E(Y_i) = 1P(Y_i = 1) + 0P(Y_i = 0) = P(Y_i = 1) = P_i \quad (\text{say})$$

$$\therefore E\left(Y_i / X_{2i}, X_{3i}, \ldots, X_{ki}\right) = P_i \ , \quad \text{where } P_i = E(Y_i) \tag{20.11}$$

Now, if we adopt the logistic model (20.9) for P_i given in Eq. (20.11), then we get

$$P_i = \frac{1}{1+e^{-Z_i}}, \quad -\infty < Z_i < \infty, \quad i = 1, 2, \ldots, n \tag{20.12}$$

$$\Rightarrow 1 - P_i = 1 - \frac{1}{1+e^{-Z_i}} = \frac{e^{-Z_i}}{1+e^{-Z_i}}$$

$$\Rightarrow \frac{P_i}{1-P_i} = \left(\frac{1}{1+e^{-Z_i}}\right)\left(\frac{1+e^{-Z_i}}{e^{-Z_i}}\right) = e^{Z_i}$$

$$\boxed{\Rightarrow \log\left(\frac{P_i}{1-P_i}\right) = Z_i = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} \qquad (\text{from Eq. (20.10)})} \tag{20.13}$$

Now the model (20.13) is called as **LOGIT model**. The quantity $\log\left(\frac{P_i}{1-P_i}\right) = L_i$ (say) is called as the "LOGIT".

Note: The model given in Eq. (20.12) with $Z_i = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji}$ is also called as LOGIT model.

Remark: The quantity $\frac{P_i}{1-P_i}$ is simply the odds ratio in favor of happening the event $Y_i = 1$. Thus if $P_i = 0.8$, it means that the odds are $\frac{P_i}{1-P_i} = \frac{0.8}{0.2} = 4$ to 1 in favor of happening of $Y_i = 1$.

Features of Logit Model:

1. As P_i goes from 0 to 1 the LOGIT L_i goes from $-\infty$ to $+\infty$. That is although the probabilities lies between 0 and 1, the LOGITs are not so bounded.

2. Although L is linear in $X_2, X_3, \ldots, X_k$, the $P_i s$ are not so. This property is in contrast with the **LPM,** where the probability increases linearly with values of $X_2, X_3, ..., X_k$.

3. If the Logit, $L_i = \log\left(\frac{P_i}{1-P_i}\right)$ in the LOGIT model is positive it means that when the values of the regressors increase, the odds that the regressand $Y_i = 1$ (means some event of interest happens) increases.

4. If L_i is negative, the odds that the regressand $Y_i = 1$ decreases as the values of regressors increases. To put it differently the LOGIT becomes negative and increasingly large in magnitude (-∞, 0) as the odds ratio $\left(\frac{P_i}{1-P_i}\right)$ decreases from 1 to 0 and becomes positive and increasingly large (0, ∞) as the odds ratio increases from 1 to ∞.

M.L. Estimation of LOGIT model:

The LOGIT model is given by

$$\begin{aligned} P_i &= P(Y_i = 1 / X_{2i}, X_{3i}, \ldots, X_{ki}) \\ &= 1 - P(Y_i = 0 / X_{2i}, X_{3i}, \ldots, X_{ki}) \\ &= \frac{1}{1 + e^{-\left(\beta_1 + \sum_{j=2}^{k} \beta_j X_{ji}\right)}}, \qquad \text{for i=1,2,...,n} \end{aligned} \tag{20.14}$$

where Y is binary dependent variable (take the value 0 or 1) and $X_2, X_3, ..., X_k$ are k-1 explanatory variables.

We do not actually observe P_i but only observe the outcome

Y_i=1, if the event occurs

Y_i=0, if the event does not occur

Suppose we have a random sample of *n* observations $Y_1, Y_2, \ldots, Y_n$, the likelihood function of $Y_1, Y_2, \ldots, Y_n$ is given as

$$L(Y_1, Y_2, \ldots, Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} P_i^{Y_i} (1-P_i)^{1-Y_i} \tag{20.15}$$

Taking logarithms on both sides we get

$$\begin{aligned} Log\ L &= \sum_{i=1}^{n} [Y_i \log P_i + (1-Y_i) \log(1-P_i)] \\ &= \sum_{i=1}^{n} [Y_i \log P_i - Y_i \log(1-P_i) + \log(1-P_i)] \\ &= \sum_{i=1}^{n} Y_i \log\left[\frac{P_i}{1-P_i}\right] + \sum_{i=1}^{n} \log(1-P_i) \end{aligned} \tag{20.16}$$

Now from Eq. (20.14), we may write

$$1-P_i = 1-\frac{1}{1+e^{-\left(\beta_1+\sum_{j=2}^{k}\beta_j X_{ji}\right)}} = \frac{e^{-\left(\beta_1+\sum_{j=2}^{k}\beta_j X_{ji}\right)}}{1+e^{-\left(\beta_1+\sum_{j=2}^{k}\beta_j X_{ji}\right)}} \qquad (20.17)$$

From Eqs. (20.14) and (20.17), we get

$$\frac{P_i}{1-P_i} = e^{\beta_1+\sum_{j=2}^{k}\beta_j X_{ji}} \Rightarrow \log\left(\frac{P_i}{1-P_i}\right) = \beta_1 + \sum_{j=2}^{k}\beta_j X_{ji} \qquad (20.18)$$

Eq. (20.17) may be rewritten as

$$1-P_i = \frac{1}{1+e^{\left(\beta_1+\sum_{j=2}^{k}\beta_j X_{ji}\right)}} \qquad (20.19)$$

Now, substituting Eq. (20.18) and Eq. (20.19) in Eq. (20.16) we get

$$\log L = \sum_{i=1}^{n} Y_i\left(\beta_1 + \sum_{j=2}^{k}\beta_j X_{ji}\right) - \sum_{i=1}^{n}\log\left(1+e^{\left(\beta_1+\sum_{j=2}^{k}\beta_j X_{ji}\right)}\right) \qquad (20.20)$$

Differentiating the above $Log\ \mathrm{L}$ with respect to β_j , j=1,2,…,k, and setting them equal to zero, we get

We have $\dfrac{\partial \log L}{\partial \beta_j} = 0 \qquad j = 1,2,\ldots,k$ (20.21)

Since the above set of 'k' equations are in non-linear form and are not in explicit form, one has to solve the 'k' equations simultaneously using some iterative technique such as Newton-Raphson method or Gauss Newton methods to obtain the estimates of $\beta_1, \beta_2, \ldots, \beta_k$.Once, we get these estimates they can be substituted in the logistic model Eq. (20.14) to obtain the estimated logistic model.

20.5 The Probit Model

Let us assume that we have a regression model

$$Y_i^* = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i = \beta_1 + \sum_{j=2}^{k}\beta_j X_{ji} + u_i \qquad (20.22)$$

where Y_i^* is not observed it is commonly called a "latent" variable what we observe is a dummy variable,

$$\begin{aligned} Y_i &= 1, \text{ if } Y_i^* > 0 \\ &= 0, \text{ otherwise} \end{aligned} \qquad (20.23)$$

For instance if the observed dummy variable Y_i is where the given person is employed or not, then Y_i^* would be defined as "propensity" or "ability" to find employment. Similarly, if the observed dummy variable Y_i is whether the person has bought a car or not, then Y_i^* could be defined as “desired” or “ability” to buy a car. Note that in both the examples we have given there is “desire” or “ability” involved. Thus the explanatory variables in Eq. (20.22) would contain variables that explain both these variables.

Now from Eq. (20.23) that multiplying Y_i^* by any positive constant does not change Y_i. Hence, if we observe Y_i we can estimate θ in Eq. (20.22) up to a positive multiple. Hence, it is customary to assume $V(u_i)=1$, this fixes the scale of Y_i^*.

If we denote

$$Z_i = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} \qquad (20.24)$$

then we get from Eqs. (20.22) and (20.23)

$$\begin{aligned} P_i = P(Y_i = 1) &= P(Y_i^* > 0) \\ &= P(Z_i + u_i > 0) \\ &= P(u_i > -Z_i) \\ &= 1 - P(u_i < -Z_i) \\ &= 1 - F(-Z_i) \end{aligned} \qquad (20.25)$$

where F is the c.d.f. of error term u_i. If the distribution of u_i is symmetric then

$$1 - F(-z_i) = F(z_i)$$

and in this case Eq. (20.25) becomes

$$P_i = F(z_j) = P(u_i < Z_i) \qquad (20.26)$$

Now functional form *F* in Eq. (20.26) will depend on the assumption made about the error term u_i. If the cumulative distribution of u_i is a normal distribution with mean 0 and variance σ^2, then the above model become a PROBIT model and in this case

$$\begin{aligned} P_i = F(Z_i) &= P(u_i < Z_i) \\ &= P(u_i/\sigma < Z_i/\sigma) && (\because \sigma > 0) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_i/\sigma} e^{-\frac{1}{2}t^2}\, dt && \left(\because \frac{u_i}{\sigma} \sim N(0,1)\right) \\ &= \Phi(Z_i/\sigma), && \text{where } \Phi \text{ is c.d.f. of standard normal variate} \end{aligned}$$

and $F^{-1}(P_i) = Z_i = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji}$ [from Eq. (20.24)] (20.27)

Since F^{-1} is the inverse of normal c.d.f., the model (20.27) is called as **PROBIT model**.

Note: If the cumulative distribution of u_i is a logistic distribution then the Eq. (20.26) yields LOGIT model and in this case

$$P_i = F(Z_i) = \frac{1}{1+e^{-Z_i}} \Rightarrow \frac{P_i}{1-P_i} = e^{Z_i}$$

$$\Rightarrow \log\left(\frac{P_i}{1-P_i}\right) = Z_i = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} \quad \text{[from Eq. (20.24)]} \qquad (20.28)$$

Now the above equation is called LOGIT Model. The quantity $\log\left(\frac{P_i}{1-P_i}\right) = L_i$ (say) is called as the "LOGIT".

Remark: ***The LOGIT model can be derived alternatively as shown in the above note.***

M.L. Estimation of PROBIT model:

The PROBIT model is given by

$$\begin{aligned} P_i &= P(Y_i = 1/X_{2i}, X_{3i}, ..., X_{ki}) \\ &= 1 - P(Y_i = 0/X_{2i}, X_{3i}, ..., X_{ki}) \\ &= F(\beta_1 + \sum_{j=2}^{k} \beta_j X_{ji}), \qquad \text{for } i = 1, 2, ..., n \end{aligned} \qquad (20.29)$$

where Y is binary dependent variable (take the values 0 or 1) and $X_2, X_3, ..., X_k$ are $k-1$ explanatory variables. Here, $F(\ .\ )$ is the c.d.f. of the normal variate $N(0, \sigma^2)$, given by

$$F(Z_i) = \Phi(Z_i/\sigma) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_i/\sigma} e^{-\frac{1}{2}t^2} dt, \quad \text{where } \Phi \text{ is c.d.f. of standard normal variate}$$

We do not actually observe P_i but only observe the outcome

Y_i=1, if the event occurs

Y_i=0, if the event does not occur

Since each Y_i is a Bernoulli random variable, we can write

$$P(Y_i = 1) = P_i \qquad (20.30)$$

$$P(Y_i = 0) = 1 - P_i \qquad (20.31)$$

Suppose we have a random sample of 'n' observations. Letting $f(Y_i)$ denote the probability that Y_i=1 (or) 0. Now, the likelihood functions of $Y_1, Y_2, ..., Y_n$ is given as

$$L(Y_1, Y_2, ..., Y_n) = \prod_{i=1}^{n} f(Y_i) = \prod_{i=1}^{n} P_i^{Y_i} (1-P_i)^{1-Y_i} \qquad (20.32)$$

Taking logarithms on both sides of Eq. (20.32) we get

$$\text{Log } L = \sum_{i=1}^{n} [Y_i \log P_i + (1-Y_i)\log(1-P_i)]$$

Substituting P_i from Eq. (20.29), we get

$$\log L = \sum_{i=1}^{n} \left\{ Y_i \log F\left(\beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} \right) + (1 - Y_i) \log\left[1 - F\left(\beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} \right) \right] \right\} \quad (20.33)$$

Differentiating Eq. (20.33) with respect to β_j , j=1,2,…,k, and setting them equal to zero, we get

$$\frac{\partial \log L}{\partial \beta_j} = 0 \quad j = 1, 2, \ldots, k \quad (20.34)$$

Since the above set of 'k' equations are in non-linear form and are not in explicit form and hence one has to solve the above 'k' equations simultaneously using some iterative technique such as Newton-Raphson method or Gauss Newton method to obtain the estimates of $\beta_1, \beta_2, ..., \beta_k$. Once, we get these estimates they can be substituted in the PROBIT model Eq. (20.29) to obtain the estimated PROBIT model.

20.6 The Tobit Model

Let us assume that we have a regression model

$$Y_i^* = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i \quad (20.35)$$

where Y_i^* is not observed and it is commonly called a "latent" variable. In the Logit and Probit models what we observe is a dummy variable, defined by

$$\begin{aligned} Y_i &= 1, \text{ if } Y_i^* > 0 \\ &= 0, \text{ otherwise} \end{aligned} \quad (20.36)$$

Suppose, however, that Y_i^* is observed if Y_i^* >0 and is not observed if $Y_i^* \leq 0$. Then the observed Y_i will be defined as

$$Y_i = \begin{cases} \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases} \quad (20.37)$$

where $u_i \overset{iid}{\sim} N(0, \sigma^2)$

This is known as the T*obit model* (Tobin's probit) and was first analyzed in the econometrics literature by Tobin. It is also known as a *censored normal regression model* because some observations on Y_i^* (those for which $Y_i^* \leq 0$) are censored (we are not allowed to see them). Our objective is to estimate the parameters β's and *σ.*

Some Examples:

There have been a very large number of applications of the Tobit model. We present two examples below.

The first example that Tobin considered was that of automobile expenditures. Suppose that we have data on a sample of households. We wish to estimate, say, the income elasticity of demand for automobiles. Let Y_i^* denote expenditures on automobiles and X denote income, and we postulate the regression equation

$$Y_i^* = \alpha + \beta X_i + u_i \qquad u_i \overset{iid}{\sim} N(0, \sigma^2)$$

However, in the sample we would have a large number of observations for which the expenditures on automobiles are zero. Tobin argued that we should use the censored regression model. We can specify the model as

$$Y_i = \begin{cases} \alpha + \beta X_i + u_i & \text{for those with positive automobile expenditures} \\ 0 & \text{for those with no expenditures} \end{cases} \qquad (20.38)$$

The structure of this model appears to be the same as that in Eq. (20.37).
The second example that Tobin considered was the hours worked (*H*) or wages (*W*). If we have observations on a number of individuals, some of whom are employed and others not, we can specify the model for hours worked as

$$H_i = \begin{cases} \alpha + \beta X_i + u_i & \text{for those working} \\ 0 & \text{for those not working} \end{cases} \qquad (20.39)$$

Similarly, for wages we can specify the model

$$W_i = \begin{cases} \alpha + \beta Z_i + v_i & \text{for those working} \\ 0 & \text{for those not working} \end{cases} \qquad (20.40)$$

The structure of these models again appears to be the same as in Eq. (20.37).

Method of Estimation:
Let us consider the estimation of β's and σ. We cannot use OLS with the positive observations Y_i because when we write the model

$$Y_i = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i$$

the error term u_i does not have a zero mean. Since observation with $Y_i \le 0$ are omitted, it implies that only observations for which $u_i > -(\beta_1 + \sum_{j=2}^{k} \beta_j X_{ji})$ are included in the sample. Thus the distribution of u_i is a *truncated normal distribution* and its mean is not zero. In fact, it depends on $\beta_2, \beta_3, ..., \beta_k, \sigma$ and $X_{2i}, X_{3i}, ..., X_{ki}$ and is thus different for each observation. A method of estimation commonly suggested is the maximum likelihood method, which is as follows.

Note that we have two sets of observations:

1. The positive values of Y_i, for which we can write down the normal density function as usual.

   We note that $(Y_i - (\beta_1 + \sum_{j=2}^{k} \beta_j X_{ji}))/\sigma$ has a standard normal distribution.

2. The zero observations of Y_i, for which all we know is that $Y_i^* \le 0$ or $\beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i \le 0$.

   Since u_i / σ has a standard normal distribution, we will write this as $u_i / \sigma \le -(\beta_1 + \sum_{j=2}^{k} \beta_j X_{ji}) / \sigma$. The probability of this can be written as $\Phi(-(\beta_1 + \sum_{j=2}^{k} \beta_j X_{ji}) / \sigma)$, where $\Phi(.)$ is the c.d.f. of the standard normal.

Let us denote the density function of the standard normal by $\phi(.)$ and then

$$\phi(t) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2} \quad \text{and} \quad \Phi(z) = \int_{-\infty}^{z} \phi(t)dt$$

Using this notation we can write the likelihood function for the Tobit model as

$$L = \prod_{y_i > 0} \frac{1}{\sigma} \phi \left(\frac{y_i - (\beta_1 + \sum_{j=2}^{k} \beta_j X_{ji})}{\sigma} \right) \prod_{y_i \le 0} \Phi \left(- \frac{\beta_1 + \sum_{j=2}^{k} \beta_j X_{ji}}{\sigma} \right)$$

Maximizing this likelihood function with respect to $\beta_2, \beta_3, ..., \beta_k$ *and* σ, we get the ML estimates of these parameters.

20.7 Measuring goodness of fit

There is a problem with the use of conventional R^2 - type measures when the explained variable Y takes only two values. The predicted values $\hat{Y}$ are probabilities and the actual values Y are either 0 or 1. For the LPM and Logit models, we have $\sum Y = \sum \hat{Y}$, as with the linear regression model, if a constant term is also estimated. For the Probit model there is no such exact relationship although it is approximately valid.

There are several R^2 - type measures that have been suggested for models with qualitative dependent variables. The following are some of them. In the case of the linear regression model, they are all equivalent. However, they are not equivalent in the case of models with qualitative dependent variables.

1. R^2 =Squared correlation between Y and $\hat{Y}$ = $[cor(Y, \hat{Y})]^2$.
2. *Measures based on residual sum of squares (RSS)*. For the linear regression model we have

$$R^2 = 1 - \left[\frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y} \right)^2} \right]$$

We can use this same measure if we can use $\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$ as the measure of *RSS*. Effron argued that we can use it. Note that in the case of binary dependent variable,

$$\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2=\sum Y_i^2-n\overline{Y}^2=n_1-n\left(\frac{n_1}{n}\right)^2=\frac{n_1n_2}{n}$$

where $n_1 = number\ of\ 1's\ \ and \quad n_2 = number\ of\ 0's$

Hence **Effron's measure of** R^2 is

$$R^2=1-\frac{n}{n_1n_2}\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=1-\frac{nRSS}{n_1n_2}$$

Amemiya argues that it makes more sense to define that *RSS* as

$$\sum_{i=1}^{n}\frac{\left(Y_i-\hat{Y}_i\right)^2}{\hat{Y}_i\left(1-\hat{Y}_i\right)}$$

That is, to weight the squared error $\left(Y_i-\hat{Y}_i\right)^2$ by a weight that is inversely proportional to its variance.

3. *Measures based on likelihood ratios.* For the standard linear regression model.

$$Y=\beta_1+\sum_{i=2}^{k}\beta_iX_i+u, \qquad u\overset{iid}{\sim}N(0,\sigma^2)$$

Let L_{UR} be the maximum of likelihood function when maximized with respect to all the parameters and L_R be the maximum when maximized with restriction $\beta_i=0$ for $i=1,2,\ldots,k$. Then, a measure of R^2, defined by McFadden, is

$$\text{McFadden's } R^2=1-\frac{\log L_{\text{UR}}}{\log L_{\text{R}}}$$

However, this measure does not correspond to any R^2 measure in the linear regression model.

4. Finally, we can also think of R^2 in terms of the *proportion* of correct predictions. Since the dependent variable is a 0 or 1 variable, after we compute the $\hat{Y}_i$ we classify the *i*^th^ observation as belonging to group 1 if $\hat{Y}_i>=0.5$ and group 2 if $\hat{Y}_i<0.5$. We can then count the number of correct predictions. We can define a predicted value $\hat{Y}_i^*$, which is also a zero one variable such that

$$\hat{Y}_i^*=\begin{cases}1 & \text{if } \hat{Y}_i>=0.5\\ 0 & \text{if } \hat{Y}<0.5\end{cases}$$

Now define

$$\text{count } R^2=\frac{\text{number of correct predictions}}{\text{total number of observations}}$$

20.8 Summary and Conclusions

1. Qualitative response (dummy dependent variable) regression models refer to models in which the response, dependent, or regressand, variable is not quantitative or an interval scale.

2. The simplest possible qualitative response regression model is the binary model in which the regressand is of the yes/no or presence/absence type. Regarding the dummy dependent variable, there are three different models that one can use: the linear probability model (LPM), the Logit model, and the Probit model.

3. The simplest possible binary regression model is the LPM, in which the binary response variable is regressed on the relevant explanatory variables by using the standard OLS methodology. Simplicity may not be a virtue here, for the LPM suffers from several estimation problems. The LPM has the drawback that the predicted values can be outside the permissible interval (0, 1). Even if some of the estimation problems can be overcome, the fundamental weakness of the LPM is that it assumes that the probability of something happening increases linearly with the level of the regressor. This very restrictive assumption can be avoided if we use the Logit and Probit models.

4. In the analysis of models with dummy dependent variables, we assume the existence of a latent (unobserved) continuous variable which is specified as the usual regression model. However, the latent variable can be observed only as dichotomous variable. The difference, between the Logit and Probit models, is in the assumptions made about the error term. If the error term has a logistic distribution, we have the Logit model. If it has a normal distribution, we have the Probit model. From the practical point of view, there is not much to choose between the two. The results are usually very similar.

5. In the Logit model the dependent variable is the log of the odds ratio, which is a linear function of the regressors. The probability function that underlies the Logit model is the logistic distribution.

6. If we choose the normal distribution as the appropriate probability distribution, then we can use the Probit model. This model is mathematically a bit difficult as it involves integrals. But for all practical purposes, both Logit and Probit models give similar results. In practice, the choice therefore depends on the ease of computation, which is not a serious problem with sophisticated statistical packages that are now readily available.

7. A model that is closely related to the Probit model is the Tobit model, also known as a censored regression model. In this model, the response variable is observed only if certain condition(s) are met. Thus, the question of how much one spends on a car is meaningful only if one decides to buy a car to begin with.

8. For comparing the LP, Logit, and Probit models, one can look at the number of cases correctly predicted. However, this is not enough. It is better to look at some measures

of R^2's. We discuss several measures of namely of i) Squared correlation between Y and $\hat{Y}$, ii) **Effron's** R^2 iii) McFadden's R^2 and iv) count R^2.

Note: In this lesson no illustrations are given since the demonstration of these models require software packages.

20.9 Self Assessment Questions

1. Discuss the need of qualitative response models. Explain the linear probability model (LPM)
2. Discuss how logistic regression model is different from the traditional regression model.
3. Explain the LOGIT Model and an estimation procedure of the model.
4. What are the applications of logistic regression model?
5. Explain ML estimation procedure of logistic model and give two of its applications.
6. Describe the PROBIT and TOBIT models.
7. Discuss the need of qualitative response models. Explain the ML estimation of the parameters of the LOGIT model.
8. Explain in detail the PROBIT model and also explain its use in analysis of biological data.
9. Distinguish between Probit and Logit models and give their applications.
10. Explain ML estimation method of Probit model and give two applications of the model.
11. Discuss various measures of goodness of fit for LOGIT/PROBIT models.

20.10 References

1. Gujarati, D.N. (2005): *Basic Econometrics, 4th Ed., Tata McGraw-Hill.*
2. Johnston, J. (1984): *Econometric Methods, 3rd Ed., McGraw-Hill,* New York.
3. Montgomery, D.C., Peck, E.A. and Geoffrey Vining, G. (2003): *Introduction to Linear Regression Analysis, 3rd Ed., Wiley*
4. Draper, N.R., and H. Smith(1998): *Applied Regression Analysis, 3rd Ed., John Wiley & Sons,* New York.
5. G.S. Maddala (2001): *Introduction to Econometrics, 3rd Ed., John Wiley & Sons, Ltd.*
6. Johnston, J. and DiNardo J (1997): *Econometric Methods*, 4th Ed., McGraw Hill.
7. Hill, Carter, William Griffiths, and George Judge(2001): *Undergraduate Econometrics, John Willey & Sons,* New York.
8. *Koutsoyiannis, A(1973): Theory of Econometrics, Harper & Row,* New York.