



ECONOMETRICS

**A TEACHING MATERIAL FOR DISTANCE
STUDENTS MAJORING IN ECONOMICS**

Module I

Prepared By:

Bedru Babulo

Seid Hassen

*Department of Economics
Faculty of Business and Economics
Mekelle University*

*June, 2005
Mekelle*

Econometrics Module I

Introduction to the module

The principal objective of the course, “*Introduction to Econometrics*”, is to provide an elementary but comprehensive introduction to the art and science of econometrics. It enables students to see how economic theory, statistical and mathematical methods are combined in the analysis of economic data, with a purpose of giving empirical content to economic theories and verify or refute them.

Module I of the course includes the first three chapters. The first chapter introduces students with the definition and some fundamental conceptualization of econometrics. In chapter two a fairly detailed treatment of the simple classical linear regression model will be made. In this chapter students will be introduced with the basic logic, concepts, assumptions, estimation methods, and interpretations of the simple classical linear regression models and their applications in economic science. Chapter three, which deals with Multiple Regression Models, is basically the extension of the simple regression models. But in chapter three attempts will be made to expand the linear regression model by incorporating more than one explanatory variables or regressors to the model. In both chapters (chapter one and chapter two), due attention will be given to the basics of ordinary least square (OLS) method of estimation and investigating the statistical properties of the parameter estimates which are summarized by the Gauss-Markov’s BLUE (Best, Linear, Unbiased, estimator) properties.

Contents of the Module in Brief

Chapter 1. Introduction

- Definition and scope of econometrics
- Economic models vs. econometric models
- Methodology of econometrics
- Desirable properties of an econometric model
- Goals of econometrics

Chapter 2. The Classical Regression Analysis: The Simple Linear regression Models

- Stochastic and non-stochastic relationships
- The Simple Regression model
- The basic Assumptions the Classical Regression Model
- OLS Method of Estimation
- Properties of OLS Estimators
- Inferences/Predictions

Chapter 3. The Classical Regression Analysis: The Multiple Linear Regression Models

- Assumptions
- Ordinary Least Squares (OLS) estimation
- Matrix Approach to Multiple Regression Model
- Properties of the OLS estimators
- Inferences/Predictions

Chapter One

Introduction

1.1 Definition and scope of econometrics

The economic theories we learn in various economics courses suggest many relationships among economic variables. For instance, in microeconomics we learn demand and supply models in which the quantities demanded and supplied of a good depend on its price. In macroeconomics, we study ‘investment function’ to explain the amount of aggregate investment in the economy as the rate of interest changes; and ‘consumption function’ that relates aggregate consumption to the level of aggregate disposable income.

Each of such specifications involves a relationship among economic variables. As economists, we may be interested in questions such as: If one variable changes in a certain magnitude, by how much will another variable change? Also, given that we know the value of one variable; can we forecast or predict the corresponding value of another? The purpose of studying the relationships among economic variables and attempting to answer questions of the type raised here, is to help us understand the real economic world we live in.

However, economic theories that postulate the relationships between economic variables have to be checked against data obtained from the real world. If empirical data verify the relationship proposed by economic theory, we accept the theory as valid. If the theory is incompatible with the observed behavior, we either reject the theory or in the light of the empirical evidence of the data, modify the theory. To provide a better understanding of economic relationships and a better

guidance for economic policy making we also need to know the quantitative relationships between the different economic variables. We obtain these quantitative measurements taken from the real world. *The field of knowledge which helps us to carryout such an evaluation of economic theories in empirical terms is econometrics.*

Distance students! Having said the background statement in our attempt for defining ‘*ECONOMETRICS*’, we may now formally define what econometrics is.

WHAT IS ECONOMETRICS?

Literally interpreted, econometrics means “economic measurement”, but the scope of econometrics is much broader as described by leading econometricians. Various econometricians used different ways of wordings to define econometrics. But if we distill the fundamental features/concepts of all the definitions, we may obtain the following definition.

“Econometrics is the science which integrates economic theory, economic statistics, and mathematical economics to investigate the empirical support of the general schematic law established by economic theory. It is a special type of economic analysis and research in which the general economic theories, formulated in mathematical terms, is combined with empirical measurements of economic phenomena. Starting from the relationships of economic theory, we express them in mathematical terms so that they can be measured. We then use specific methods, called econometric methods in order to obtain numerical estimates of the coefficients of the economic relationships.”

Measurement is an important aspect of econometrics. However, the scope of econometrics is much broader than measurement. As D.Intriligator rightly stated

the “*metric*” part of the word econometrics signifies ‘*measurement*’, and hence econometrics is basically concerned with measuring of economic relationships.

In short, econometrics may be considered as the integration of economics, mathematics, and statistics for the purpose of providing numerical values for the parameters of economic relationships and verifying economic theories.

1.2 Econometrics vs. mathematical economics

Mathematical economics states economic theory in terms of mathematical symbols. There is no essential difference between mathematical economics and economic theory. Both state the same relationships, but while economic theory use verbal exposition, mathematical symbols. Both express economic relationships in an exact or deterministic form. Neither mathematical economics nor economic theory allows for random elements which might affect the relationship and make it stochastic. Furthermore, they do not provide numerical values for the coefficients of economic relationships.

Econometrics differs from mathematical economics in that, although econometrics presupposes, the economic relationships to be expressed in mathematical forms, it does not assume exact or deterministic relationship. Econometrics assumes random relationships among economic variables. Econometric methods are designed to take into account random disturbances which relate deviations from exact behavioral patterns suggested by economic theory and mathematical economics. Further more, econometric methods provide numerical values of the coefficients of economic relationships.

1.3 Econometrics vs. statistics

Econometrics differs from both mathematical statistics and economic statistics. An economic statistician gathers empirical data, records them, tabulates them or

charts them, and attempts to describe the pattern in their development over time and perhaps detect some relationship between various economic magnitudes. Economic statistics is mainly a descriptive aspect of economics. It does not provide explanations of the development of the various variables and it does not provide measurements the coefficients of economic relationships.

Mathematical (or inferential) statistics deals with the method of measurement which are developed on the basis of controlled experiments. But statistical methods of measurement are not appropriate for a number of economic relationships because for most economic relationships controlled or carefully planned experiments cannot be designed due to the fact that the nature of relationships among economic variables are stochastic or random. Yet the fundamental ideas of inferential statistics are applicable in econometrics, but they must be adapted to the problem economic life. Econometric methods are adjusted so that they may become appropriate for the measurement of economic relationships which are stochastic. The adjustment consists primarily in specifying the stochastic (random) elements that are supposed to operate in the real world and enter into the determination of the observed data.

1.4 Economic models vs. econometric models

i) Economic models:

Any economic theory is an observation from the real world. For one reason, the immense complexity of the real world economy makes it impossible for us to understand all interrelationships at once. Another reason is that all the interrelationships are not equally important as such for the understanding of the economic phenomenon under study. The sensible procedure is therefore, to pick up the important factors and relationships relevant to our problem and to focus our attention on these alone. Such a deliberately simplified analytical framework is called on economic model. It is an organized set of relationships that describes the

functioning of an economic entity under a set of simplifying assumptions. All economic reasoning is ultimately based on models. Economic models consist of the following three basic structural elements.

1. A set of variables
2. A list of fundamental relationships and
3. A number of strategic coefficients

ii) Econometric models:

The most important characteristic of economic relationships is that they contain a random element which is ignored by mathematical economic models which postulate exact relationships between economic variables.

Example: Economic theory postulates that the demand for a commodity depends on its price, on the prices of other related commodities, on consumers' income and on tastes. This is an exact relationship which can be written mathematically as:

$$Q = b_0 + b_1P + b_2P_0 + b_3Y + b_4t$$

The above demand equation is exact. However, many more factors may affect demand. In econometrics the influence of these 'other' factors is taken into account by the introduction into the economic relationships of random variable. In our example, the demand function studied with the tools of econometrics would be of the stochastic form:

$$Q = b_0 + b_1P + b_2P_0 + b_3Y + b_4t + u$$

where u stands for the random factors which affect the quantity demanded.

1.5 Methodology of econometrics

Econometric research is concerned with the measurement of the parameters of economic relationships and with the predication of the values of economic variables. The relationships of economic theory which can be measured with

econometric techniques are relationships in which some variables are postulated as causes of the variation of other variables. Starting with the postulated theoretical relationships among economic variables, econometric research or inquiry generally proceeds along the following lines/stages.

1. Specification the model
2. Estimation of the model
3. Evaluation of the estimates
4. Evaluation of he forecasting power of the estimated model

1. Specification of the model

In this step the econometrician has to express the relationships between economic variables in mathematical form. This step involves the determination of three important tasks:

- i) the dependent and independent (explanatory) variables which will be included in the model.
- ii) the *a priori* theoretical expectations about the size and sign of the parameters of the function.
- iii) the mathematical form of the model (number of equations, specific form of the equations, etc.)

Note: The specification of the econometric model will be based on economic theory and on any available information related to the phenomena under investigation. Thus, specification of the econometric model presupposes knowledge of economic theory and familiarity with the particular phenomenon being studied.

Specification of the model is the most important and the most difficult stage of any econometric research. It is often the weakest point of most econometric applications. In this stage there exists enormous degree of likelihood of

committing errors or incorrectly specifying the model. Some of the common reasons for incorrect specification of the econometric models are:

1. the imperfections, looseness of statements in economic theories.
2. the limitation of our knowledge of the factors which are operative in any particular case.
3. the formidable obstacles presented by data requirements in the estimation of large models.

The most common errors of specification are:

- a. Omissions of some important variables from the function.
- b. The omissions of some equations (for example, in simultaneous equations model).
- c. The mistaken mathematical form of the functions.

2. Estimation of the model

This is purely a technical stage which requires knowledge of the various econometric methods, their assumptions and the economic implications for the estimates of the parameters. This stage includes the following activities.

- a. Gathering of the data on the variables included in the model.
- b. Examination of the identification conditions of the function (especially for simultaneous equations models).
- c. Examination of the aggregations problems involved in the variables of the function.
- d. Examination of the degree of correlation between the explanatory variables (i.e. examination of the problem of multicollinearity).
- e. Choice of appropriate economic techniques for estimation, i.e. to decide a specific econometric method to be applied in estimation; such as, OLS, MLM, Logit, and Probit.

3. Evaluation of the estimates

This stage consists of deciding whether the estimates of the parameters are *theoretically meaningful and statistically satisfactory*. This stage enables the econometrician to evaluate the results of calculations and determine the reliability of the results. For this purpose we use various criteria which may be classified into three groups:

- i. *Economic a priori criteria*: These criteria are determined by economic theory and refer to the size and sign of the parameters of economic relationships.
- ii. *Statistical criteria (first-order tests)*: These are determined by statistical theory and aim at the evaluation of the statistical reliability of the estimates of the parameters of the model. Correlation coefficient test, standard error test, t-test, F-test, and R^2 -test are some of the most commonly used statistical tests.
- iii. *Econometric criteria (second-order tests)*: These are set by the theory of econometrics and aim at the investigation of whether the assumptions of the econometric method employed are satisfied or not in any particular case. The econometric criteria serve as a second order test (as test of the statistical tests) i.e. they determine the reliability of the statistical criteria; they help us establish whether the estimates have the desirable properties of unbiasedness, consistency etc. Econometric criteria aim at the detection of the violation or validity of the assumptions of the various econometric techniques.

4) Evaluation of the forecasting power of the model:

Forecasting is one of the aims of econometric research. However, before using an estimated model for forecasting by some way or another the predictive power of the model. It is possible that the model may be economically

meaningful and statistically and econometrically correct for the sample period for which the model has been estimated; yet it may not be suitable for forecasting due to various factors (reasons). Therefore, this stage involves the investigation of the stability of the estimates and their sensitivity to changes in the size of the sample. Consequently, we must establish whether the estimated function performs adequately outside the sample of data. i.e. we must test an extra sample performance the model.

1.6 Desirable properties of an econometric model

An econometric model is a model whose parameters have been estimated with some appropriate econometric technique. The ‘goodness’ of an econometric model is judged customarily according to the following desirable properties.

1. *Theoretical plausibility.* The model should be compatible with the postulates of economic theory. It must describe adequately the economic phenomena to which it relates.
2. *Explanatory ability.* The model should be able to explain the observations of the actual world. It must be consistent with the observed behaviour of the economic variables whose relationship it determines.
3. *Accuracy of the estimates of the parameters.* The estimates of the coefficients should be accurate in the sense that they should approximate as best as possible the true parameters of the structural model. The estimates should if possible possess the desirable properties of unbiasedness, consistency and efficiency.

4. *Forecasting ability.* The model should produce satisfactory predictions of future values of the dependent (endogenous) variables.
5. *Simplicity.* The model should represent the economic relationships with maximum simplicity. The fewer the equations and the simpler their mathematical form, the better the model is considered, *ceteris paribus* (that is to say provided that the other desirable properties are not affected by the simplifications of the model).

1.7 Goals of Econometrics

Three main goals of Econometrics are identified:

- i) *Analysis* i.e. testing economic theory
- ii) *Policy making* i.e. Obtaining numerical estimates of the coefficients of economic relationships for policy simulations.
- iii) *Forecasting* i.e. using the numerical estimates of the coefficients in order to forecast the future values of economic magnitudes.

Review questions

- How would you define econometrics?
- How does it differ from mathematical economics and statistics?
- Describe the main steps involved in any econometrics research.
- Differentiate between economic and econometric model.
- What are the goals of econometrics?

Chapter Two

THE CLASSICAL REGRESSION ANALYSIS

[*The Simple Linear Regression Model*]

Economic theories are mainly concerned with the relationships among various economic variables. These relationships, when phrased in mathematical terms, can predict the effect of one variable on another. The functional relationships of these variables define the dependence of one variable upon the other variable (s) in the specific form. The specific functional forms may be linear, quadratic, logarithmic, exponential, hyperbolic, or any other form.

In this chapter we shall consider a simple linear regression model, i.e. a relationship between two variables related in a linear form. We shall first discuss two important forms of relation: *stochastic* and *non-stochastic*, among which we shall be using the former in econometric analysis.

2.1. Stochastic and Non-stochastic Relationships

A relationship between X and Y , characterized as $Y = f(X)$ is said to be deterministic or non-stochastic if for each value of the independent variable (X) there is one and only one corresponding value of dependent variable (Y). On the other hand, a relationship between X and Y is said to be stochastic if for a particular value of X there is a whole probabilistic distribution of values of Y . In such a case, for any given value of X , the dependent variable Y assumes some specific value only with some probability. Let's illustrate the distinction between stochastic and non stochastic relationships with the help of a supply function.

Assuming that the supply for a certain commodity depends on its price (other determinants taken to be constant) and the function being linear, the relationship can be put as:

$$Q = f(P) = \alpha + \beta P \text{-----} (2.1)$$

The above relationship between P and Q is such that for a particular value of P, there is only one corresponding value of Q. This is, therefore, a deterministic (non-stochastic) relationship since for each price there is always only one corresponding quantity supplied. This implies that all the variation in Y is due solely to changes in X, and that there are no other factors affecting the dependent variable.

If this were true all the points of price-quantity pairs, if plotted on a two-dimensional plane, would fall on a straight line. However, if we gather observations on the quantity actually supplied in the market at various prices and we plot them on a diagram we see that they do not fall on a straight line.

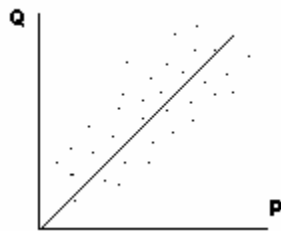


Fig 2.a The scatter diagram

The derivation of the observation from the line may be attributed to several factors.

- a. Omission of variables from the function
- b. Random behavior of human beings
- c. Imperfect specification of the mathematical form of the model
- d. Error of aggregation
- e. Error of measurement

In order to take into account the above sources of errors we introduce in econometric functions a random variable which is usually denoted by the letter ‘u’ or ‘ ε ’ and is called error term or random disturbance or stochastic term of the function, so called because u is supposed to ‘disturb’ the exact linear relationship which is assumed to exist between X and Y. By introducing this random variable in the function the model is rendered stochastic of the form:

$$Y_i = \alpha + \beta X + u_i \dots\dots\dots(2.2)$$

Thus a stochastic model is a model in which the dependent variable is not only determined by the explanatory variable(s) included in the model but also by others which are not included in the model.

2.2. Simple Linear Regression model.

The above stochastic relationship (2.2) with one explanatory variable is called simple linear regression model.

The true relationship which connects the variables involved is split into two parts: a part represented by a line and a part represented by the random term ‘u’.

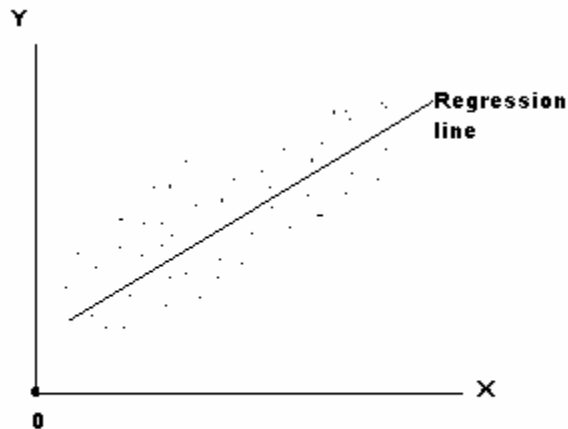


Fig 2.b the scatter diagram and the regression line

The scatter of observations represents the true relationship between Y and X. The line represents the exact part of the relationship and the deviation of the observation from the line represents the random component of the relationship.

- Were it not for the errors in the model, we would observe all the points on the line $Y'_1, Y'_2, \dots, Y'_n$ corresponding to $X_1, X_2, \dots, X_n$. However because of the random disturbance, we observe $Y_1, Y_2, \dots, Y_n$ corresponding to $X_1, X_2, \dots, X_n$. These points diverge from the regression line by $u_1, u_2, \dots, u_n$.

$$\underbrace{Y_i}_{\text{the dependent variable}} = \underbrace{\alpha + \beta x_i}_{\text{the regression line}} + \underbrace{u_i}_{\text{random variable}}$$

- The first component in the bracket is the part of Y explained by the changes in X and the second is the part of Y not explained by X, that is to say the change in Y is due to the random influence of u_i .

2.2.1 Assumptions of the Classical Linear Stochastic Regression Model.

The classicals made important assumption in their analysis of regression. The most important of these assumptions are discussed below.

1. *The model is linear in parameters.*

The classicals assumed that the model should be linear in the parameters regardless of whether the explanatory and the dependent variables are linear or not. This is because if the parameters are non-linear it is difficult to estimate them since their value is not known but you are given with the data of the dependent and independent variable.

Example 1. $Y = \alpha + \beta x + u$ is linear in both parameters and the variables, so it

Satisfies the assumption

2. $\ln Y = \alpha + \beta \ln x + u$ is linear only in the parameters. Since the classicals worry on the parameters, the model satisfies the assumption.

Dear distance students! Check yourself whether the following models satisfy the above assumption and give your answer to your tutor.

a. $\ln Y^2 = \alpha + \beta \ln X^2 + U_i$

b. $Y_i = \sqrt{\alpha + \beta X_i} + U_i$

2. U_i is a random real variable

This means that the value which u may assume in any one period depends on chance; it may be positive, negative or zero. Every value has a certain probability of being assumed by u in any particular instance.

2. The mean value of the random variable(U) in any particular period is zero

This means that for each value of x , the random variable(u) may assume various values, some greater than zero and some smaller than zero, but if we considered all the possible and negative values of u , for any given value of X , they would have on average value equal to zero. In other words the positive and negative values of u cancel each other.

$$\text{Mathematically, } E(U_i) = 0 \dots\dots\dots(2.3)$$

3. The variance of the random variable(U) is constant in each period (The assumption of homoscedasticity)

For all values of X , the u 's will show the same dispersion around their mean. In Fig.2.c this assumption is denoted by the fact that the values that u can assume lie within the same limits, irrespective of the value of X . For X_1 , u can assume any value within the range AB ; for X_2 , u can assume any value within the range CD which is equal to AB and so on.

Graphically;

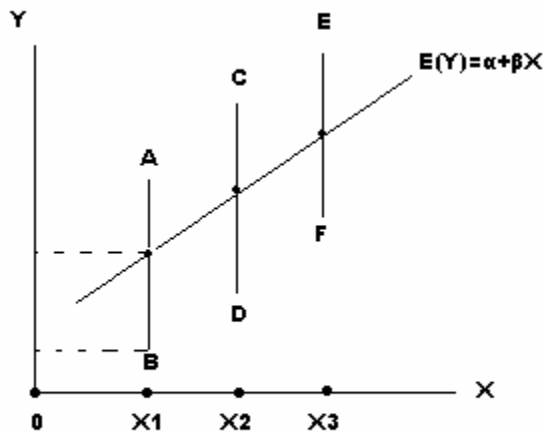


Fig. 2.c Homoscedastic Variance

Mathematically;

$Var(U_i) = E[U_i - E(U_i)]^2 = E(U_i)^2 = \sigma^2$ (Since $E(U_i) = 0$). This constant variance is called **homoscedasticity** assumption and the constant variance itself is called homoscedastic variance.

4. The random variable (U) has a normal distribution

This means the values of u (for each x) have a bell shaped symmetrical distribution about their zero mean and constant variance σ^2 , i.e.

$$U_i \sim N(0, \sigma^2) \dots\dots\dots 2.4$$

5. The random terms of different observations (U_i, U_j) are independent.

(The assumption of no autocorrelation)

This means the value which the random term assumed in one period does not depend on the value which it assumed in any other period.

Algebraically,

$$\begin{aligned} \text{Cov}(u_i, u_j) &= E[(u_i - E(u_i))(u_j - E(u_j))] \\ &= E(u_i u_j) = 0 \dots\dots\dots (2.5) \end{aligned}$$

6. The X_i are a set of fixed values in the hypothetical process of repeated sampling which underlies the linear regression model.

- This means that, in taking large number of samples on Y and X, the X_i values are the same in all samples, but the u_i values do differ from sample to sample, and so of course do the values of y_i .

7. The random variable (U) is independent of the explanatory variables.

This means there is no correlation between the random variable and the explanatory variable. If two variables are unrelated their covariance is zero.

$$\text{Hence } \text{Cov}(X_i, U_i) = 0 \dots\dots\dots (2.6)$$

Proof:-

$$\text{cov}(XU) = E[(X_i - E(X_i))(U_i - E(U_i))]$$

$$\begin{aligned}
 &= E[(X_i - E(X_i))(U_i)] \text{ given } E(U_i) = 0 \\
 &= E(X_i U_i) - E(X_i)E(U_i) \\
 &= E(X_i U_i) \\
 &= X_i E(U_i), \text{ given that the } x_i \text{ are fixed} \\
 &= 0
 \end{aligned}$$

8. *The explanatory variables are measured without error*

- U absorbs the influence of omitted variables and possibly errors of measurement in the y's. i.e., we will assume that the regressors are error free, while y values may or may not include errors of measurement.

Dear students! We can now use the above assumptions to derive the following basic concepts.

A. *The dependent variable Y_i is normally distributed.*

$$\text{i.e } \therefore Y_i \sim N[(\alpha + \beta x_i), \sigma^2] \dots \dots \dots (2.7)$$

Proof:

$$\text{Mean: } E(Y) = E(\alpha + \beta x_i + u_i)$$

$$= \alpha + \beta x_i \text{ Since } E(u_i) = 0$$

$$\text{Variance: } Var(Y_i) = E(Y_i - E(Y_i))^2$$

$$= E(\alpha + \beta x_i + u_i - (\alpha + \beta x_i))^2$$

$$= E(u_i)^2$$

$$= \sigma^2 \text{ (since } E(u_i)^2 = \sigma^2)$$

$$\therefore \text{var}(Y_i) = \sigma^2 \dots \dots \dots (2.8)$$

The shape of the distribution of Y_i is determined by the shape of the distribution of u_i which is normal by assumption 4. Since α and β , being constant, they don't

affect the distribution of y_i . Furthermore, the values of the explanatory variable, x_i , are a set of fixed values by assumption 5 and therefore don't affect the shape of the distribution of y_i .

$$\therefore Y_i \sim N(\alpha + \beta x_i, \sigma^2)$$

B. successive values of the dependent variable are independent, i.e

$$\text{Cov}(Y_i, Y_j) = 0$$

Proof:

$$\begin{aligned}\text{Cov}(Y_i, Y_j) &= E\{[Y_i - E(Y_i)][Y_j - E(Y_j)]\} \\ &= E\{[\alpha + \beta X_i + U_i - E(\alpha + \beta X_i + U_i)][\alpha + \beta X_j + U_j - E(\alpha + \beta X_j + U_j)]\} \\ &\quad (\text{Since } Y_i = \alpha + \beta X_i + U_i \text{ and } Y_j = \alpha + \beta X_j + U_j) \\ &= E[(\alpha + \beta X_i + U_i - \alpha - \beta X_i)(\alpha + \beta X_j + U_j - \alpha - \beta X_j)], \text{Since } E(u_i) = 0 \\ &= E(U_i U_j) = 0 \quad (\text{from equation (2.5)})\end{aligned}$$

Therefore, $\text{Cov}(Y_i, Y_j) = 0$.

2.2.2 Methods of estimation

Specifying the model and stating its underlying assumptions are the first stage of any econometric application. The next step is the estimation of the numerical values of the parameters of economic relationships. The parameters of the simple linear regression model can be estimated by various methods. Three of the most commonly used methods are:

1. Ordinary least square method (OLS)
2. Maximum likelihood method (MLM)
3. Method of moments (MM)

But, here we will deal with the OLS and the MLM methods of estimation.

2.2.2.1 The ordinary least square (OLS) method

The model $Y_i = \alpha + \beta X_i + U_i$ is called the true relationship between Y and X because Y and X represent their respective population value, and α and β are called the true parameters since they are estimated from the population value of Y and X. But it is difficult to obtain the population value of Y and X because of technical or economic reasons. So we are forced to take the sample value of Y and X. The parameters estimated from the sample value of Y and X are called the estimators of the true parameters α and β and are symbolized as $\hat{\alpha}$ and $\hat{\beta}$.

The model $Y_i = \hat{\alpha} + \hat{\beta}X_i + e_i$, is called estimated relationship between Y and X since $\hat{\alpha}$ and $\hat{\beta}$ are estimated from the sample of Y and X and e_i represents the sample counterpart of the population random disturbance U_i .

Estimation of α and β by least square method (OLS) or classical least square (CLS) involves finding values for the estimates $\hat{\alpha}$ and $\hat{\beta}$ which will minimize the sum of square of the squared residuals ($\sum e_i^2$).

From the estimated relationship $Y_i = \hat{\alpha} + \hat{\beta}X_i + e_i$, we obtain:

$$e_i = Y_i - (\hat{\alpha} + \hat{\beta}X_i) \dots\dots\dots(2.6)$$

$$\sum e_i^2 = \sum (Y_i - \hat{\alpha} - \hat{\beta}X_i)^2 \dots\dots\dots(2.7)$$

To find the values of $\hat{\alpha}$ and $\hat{\beta}$ that minimize this sum, we have to partially differentiate $\sum e_i^2$ with respect to $\hat{\alpha}$ and $\hat{\beta}$ and set the partial derivatives equal to zero.

$$1. \frac{\partial \sum e_i^2}{\partial \hat{\alpha}} = -2 \sum (Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0 \dots\dots\dots(2.8)$$

Rearranging this expression we will get: $\boxed{\sum Y_i = n\hat{\alpha} + \hat{\beta}\sum X_i} \dots\dots\dots(2.9)$

If you divide (2.9) by 'n' and rearrange, we get

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \dots\dots\dots(2.10)$$

$$2. \frac{\partial \sum e_i^2}{\partial \hat{\beta}} = -2 \sum X_i (Y_i - \hat{\alpha} - \hat{\beta} X_i) = 0 \dots \dots \dots (2.11)$$

Note: at this point that the term in the parenthesis in equation 2.8 and 2.11 is the residual, $e = Y_i - \hat{\alpha} - \hat{\beta} X_i$. Hence it is possible to rewrite (2.8) and (2.11) as $-2 \sum e_i = 0$ and $-2 \sum X_i e_i = 0$. It follows that;

$$\sum e_i = 0 \text{ and } \sum X_i e_i = 0 \dots \dots \dots (2.12)$$

If we rearrange equation (2.11) we obtain;

$$\boxed{\sum Y_i X_i = \hat{\alpha} \sum X_i + \hat{\beta} \sum X_i^2} \dots \dots \dots (2.13)$$

Equation (2.9) and (2.13) are called the **Normal Equations**. Substituting the values of $\hat{\alpha}$ from (2.10) to (2.13), we get:

$$\begin{aligned} \sum Y_i X_i &= \sum X_i (\bar{Y} - \hat{\beta} \bar{X}) + \hat{\beta} \sum X_i^2 \\ &= \bar{Y} \sum X_i - \hat{\beta} \bar{X} \sum X_i + \hat{\beta} \sum X_i^2 \\ \sum Y_i X_i - \bar{Y} \sum X_i &= \hat{\beta} (\sum X_i^2 - \bar{X} \sum X_i) \\ \sum XY - n \bar{X} \bar{Y} &= \hat{\beta} (\sum X_i^2 - n \bar{X}^2) \\ \boxed{\hat{\beta} = \frac{\sum XY - n \bar{X} \bar{Y}}{\sum X_i^2 - n \bar{X}^2}} &\dots \dots \dots (2.14) \end{aligned}$$

Equation (2.14) can be rewritten in somewhat different way as follows;

$$\begin{aligned} \sum (X - \bar{X})(Y - \bar{Y}) &= \sum (XY - X\bar{Y} - \bar{X}Y + \bar{X}\bar{Y}) \\ &= \sum XY - \bar{Y} \sum X - \bar{X} \sum Y + n \bar{X} \bar{Y} \\ &= \sum XY - n \bar{Y} \bar{X} - n \bar{X} \bar{Y} + n \bar{X} \bar{Y} \\ \sum (X - \bar{X})(Y - \bar{Y}) &= \sum XY - n \bar{X} \bar{Y} \dots \dots \dots (2.15) \end{aligned}$$

$$\sum (X - \bar{X})^2 = \sum X^2 - n \bar{X}^2 \dots \dots \dots (2.16)$$

Substituting (2.15) and (2.16) in (2.14), we get

$$\hat{\beta} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(X - \bar{X})^2}$$

Now, denoting $(X_i - \bar{X})$ as x_i , and $(Y_i - \bar{Y})$ as y_i we get;

$$\boxed{\hat{\beta} = \frac{\Sigma x_i y_i}{\Sigma x_i^2}} \dots\dots\dots (2.17)$$

The expression in (2.17) to estimate the parameter coefficient is termed is the formula in *deviation form*.

2.2.2.2 Estimation of a function with zero intercept

Suppose it is desired to fit the line $Y_i = \alpha + \beta X_i + U_i$, subject to the restriction $\alpha = 0$. To estimate $\hat{\beta}$, the problem is put in a form of restricted minimization problem and then Lagrange method is applied.

$$\text{We minimize: } \Sigma e_i^2 = \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

$$\text{Subject to: } \hat{\alpha} = 0$$

The composite function then becomes

$$Z = \sum (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2 - \lambda \hat{\alpha}, \quad \text{where } \lambda \text{ is a Lagrange multiplier.}$$

We minimize the function with respect to $\hat{\alpha}, \hat{\beta}$, and λ

$$\frac{\partial Z}{\partial \hat{\alpha}} = -2 \Sigma (Y_i - \hat{\alpha} - \hat{\beta} X_i) - \lambda = 0 \dots\dots\dots (i)$$

$$\frac{\partial Z}{\partial \hat{\beta}} = -2 \Sigma (Y_i - \hat{\alpha} - \hat{\beta} X_i) (X_i) = 0 \dots\dots\dots (ii)$$

$$\frac{\partial Z}{\partial \lambda} = -2 \hat{\alpha} = 0 \dots\dots\dots (iii)$$

Substituting (iii) in (ii) and rearranging we obtain:

$$\Sigma X_i (Y_i - \hat{\beta} X_i) = 0$$

$$\Sigma Y_i X_i - \hat{\beta} \Sigma X_i^2 = 0$$

$$\hat{\beta} = \frac{\sum X_i Y_i}{\sum X_i^2} \dots \dots \dots (2.18)$$

This formula involves the *actual values* (observations) of the variables and not their deviation forms, as in the case of unrestricted value of $\hat{\beta}$.

2.2.2.3. Statistical Properties of Least Square Estimators

There are various econometric methods with which we may obtain the estimates of the parameters of economic relationships. We would like to an estimate to be as close as the value of the true population parameters i.e. to vary within only a small range around the true parameter. *How are we to choose among the different econometric methods, the one that gives ‘good’ estimates?* We need some criteria for judging the ‘goodness’ of an estimate.

‘Closeness’ of the estimate to the population parameter is measured by the mean and variance or standard deviation of the *sampling distribution* of the estimates of the different econometric methods. We assume the usual process of repeated sampling i.e. we assume that we get a very large number of samples each of size ‘n’; we compute the estimates $\hat{\beta}$ ’s from each sample, and for each econometric method and we form their distribution. We next compare the *mean (expected value)* and the *variances* of these distributions and we choose among the alternative estimates the one whose distribution is concentrated as close as possible around the population parameter.

PROPERTIES OF OLS ESTIMATORS

The ideal or optimum properties that the OLS estimates possess may be summarized by well known theorem known as the **Gauss-Markov Theorem**.

Statement of the theorem: “Given the assumptions of the classical linear regression model, the OLS estimators, in the class of linear and unbiased estimators, have the minimum variance, i.e. the OLS estimators are BLUE.

According to the this theorem, under the basic assumptions of the classical linear regression model, the least squares estimators are linear, unbiased and have minimum variance (i.e. are best of all linear unbiased estimators). Some times the theorem referred as the BLUE theorem i.e. Best, Linear, Unbiased Estimator. An estimator is called BLUE if:

- a. **Linear:** a linear function of the a random variable, such as, the dependent variable Y.
- b. **Unbiased:** its average or expected value is equal to the true population parameter.
- c. **Minimum variance:** It has a minimum variance in the class of linear and unbiased estimators. An unbiased estimator with the least variance is known as an efficient estimator.

According to the Gauss-Markov theorem, the OLS estimators possess all the BLUE properties. The detailed proof of these properties are presented below
Dear colleague lets proof these properties one by one.

a. Linearity: (for $\hat{\beta}$)

Proposition: $\hat{\alpha}$ & $\hat{\beta}$ are linear in Y.

Proof: From (2.17) of the OLS estimator of $\hat{\beta}$ is given by:

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i (Y - \bar{Y})}{\sum x_i^2} = \frac{\sum x_i Y - \bar{Y} \sum x_i}{\sum x_i^2},$$

$$(\text{but } \sum x_i = \sum (X - \bar{X}) = \sum X - n\bar{X} = n\bar{X} - n\bar{X} = 0)$$

$$\Rightarrow \hat{\beta} = \frac{\sum x_i Y}{\sum x_i^2}; \quad \text{Now, let } \frac{x_i}{\sum x_i^2} = K_i \quad (i = 1, 2, \dots, n)$$

$$\therefore \hat{\beta} = \sum K_i Y \text{------(2.19)}$$

$$\Rightarrow \hat{\beta} = K_1 Y_1 + K_2 Y_2 + K_3 Y_3 + \text{-----} + K_n Y_n$$

$$\therefore \hat{\beta} \text{ is linear in Y}$$

Check yourself question:

Show that $\hat{\alpha}$ is linear in Y? Hint: $\hat{\alpha} = \sum (\frac{1}{n} - \bar{X}k_i)Y_i$. Derive this relationship between $\hat{\alpha}$ and Y.

b. Unbiasedness:

Proposition: $\hat{\alpha}$ & $\hat{\beta}$ are the unbiased estimators of the true parameters α & β

From your statistics course, you may recall that if $\hat{\theta}$ is an estimator of θ then $E(\hat{\theta}) - \theta = \text{the amount of bias}$ and if $\hat{\theta}$ is the unbiased estimator of θ then bias = 0 i.e. $E(\hat{\theta}) - \theta = 0 \Rightarrow E(\hat{\theta}) = \theta$

In our case, $\hat{\alpha}$ & $\hat{\beta}$ are estimators of the true parameters α & β . To show that they are the unbiased estimators of their respective parameters means to prove that:

$$E(\hat{\beta}) = \beta \quad \text{and} \quad E(\hat{\alpha}) = \alpha$$

- **Proof (1):** Prove that $\hat{\beta}$ is unbiased i.e. $E(\hat{\beta}) = \beta$.

We know that $\hat{\beta} = \sum k_i Y_i = \sum k_i (\alpha + \beta X_i + U_i)$

$$= \alpha \sum k_i + \beta \sum k_i X_i + \sum k_i u_i,$$

$$\text{but } \sum k_i = 0 \quad \text{and} \quad \sum k_i X_i = 1$$

$$\sum k_i = \frac{\sum x_i}{\sum x_i^2} = \frac{\sum (X - \bar{X})}{\sum x_i^2} = \frac{\sum X - n\bar{X}}{\sum x_i^2} = \frac{n\bar{X} - n\bar{X}}{\sum x_i^2} = 0$$

$$\Rightarrow \sum k_i = 0 \dots\dots\dots (2.20)$$

$$\begin{aligned} \sum k_i X_i &= \frac{\sum x_i X_i}{\sum x_i^2} = \frac{\sum (X - \bar{X}) X_i}{\sum x_i^2} \\ &= \frac{\sum X^2 - \bar{X} \sum X}{\sum X^2 - n\bar{X}^2} = \frac{\sum X^2 - n\bar{X}^2}{\sum X^2 - n\bar{X}^2} = 1 \end{aligned}$$

$$\Rightarrow \sum k_i X_i = 1 \dots\dots\dots (2.21)$$

$$\hat{\beta} = \beta + \sum k_i u_i \Rightarrow \hat{\beta} - \beta = \sum k_i u_i \dots\dots\dots (2.22)$$

$$E(\hat{\beta}) = E(\beta) + \sum k_i E(u_i), \text{ Since } k_i \text{ are fixed}$$

$$E(\hat{\beta}) = \beta, \quad \text{since } E(u_i) = 0$$

Therefore, $\hat{\beta}$ is unbiased estimator of β .

- **Proof(2): prove that $\hat{\alpha}$ is unbiased i.e.:** $E(\hat{\alpha}) = \alpha$

From the proof of linearity property under 2.2.2.3 (a), we know that:

$$\begin{aligned}\hat{\alpha} &= \Sigma(\frac{1}{n} - \bar{X}k_i)Y_i \\ &= \Sigma\left[\left(\frac{1}{n} - \bar{X}k_i\right)(\alpha + \beta X_i + U_i)\right], \text{ Since } Y_i = \alpha + \beta X_i + U_i \\ &= \alpha + \beta \frac{1}{n} \Sigma X_i + \frac{1}{n} \Sigma U_i - \alpha \bar{X} \Sigma k_i - \beta \bar{X} \Sigma k_i X_i - \bar{X} \Sigma k_i U_i \\ &= \alpha + \frac{1}{n} \Sigma U_i - \bar{X} \Sigma k_i U_i, \quad \Rightarrow \hat{\alpha} - \alpha = \frac{1}{n} \Sigma U_i - \bar{X} \Sigma k_i U_i \\ &= \Sigma(\frac{1}{n} - \bar{X}k_i)u_i \dots\dots\dots(2.23)\end{aligned}$$

$$E(\hat{\alpha}) = \alpha + \frac{1}{n} \Sigma E(u_i) - \bar{X} \Sigma k_i E(u_i)$$

$$E(\hat{\alpha}) = \alpha \text{-----} (2.24)$$

$\therefore \hat{\alpha}$ is an unbiased estimator of α .

c. Minimum variance of $\hat{\alpha}$ and $\hat{\beta}$

Now, we have to establish that out of the class of linear and unbiased estimators of α and β , $\hat{\alpha}$ and $\hat{\beta}$ possess the smallest sampling variances. For this, we shall first obtain variance of $\hat{\alpha}$ and $\hat{\beta}$ and then establish that each has the minimum variance in comparison of the variances of other linear and unbiased estimators obtained by any other econometric methods than OLS.

a. Variance of $\hat{\beta}$

$$\text{var}(\hat{\beta}) = E(\hat{\beta} - E(\hat{\beta}))^2 = E(\hat{\beta} - \beta)^2 \dots\dots\dots(2.25)$$

Substitute (2.22) in (2.25) and we get

$$\begin{aligned}\text{var}(\hat{\beta}) &= E(\Sigma k_i u_i)^2 \\ &= E[k_1^2 u_1^2 + k_2^2 u_2^2 + \dots\dots\dots + k_n^2 u_n^2 + 2k_1 k_2 u_1 u_2 + \dots\dots + 2k_{n-1} k_n u_{n-1} u_n]\end{aligned}$$

$$\begin{aligned}
 &= E[k_1^2 u_1^2 + k_2^2 u_2^2 + \dots + k_n^2 u_n^2] + E[2k_1 k_2 u_1 u_2 + \dots + 2k_{n-1} k_n u_{n-1} u_n] \\
 &= E\left(\sum k_i^2 u_i^2\right) + E(\Sigma k_i k_j u_i u_j) \quad i \neq j \\
 &= \Sigma k_i^2 E(u_i^2) + 2\Sigma k_i k_j E(u_i u_j) = \sigma^2 \Sigma k_i^2 \quad (\text{Since } E(u_i u_j) = 0)
 \end{aligned}$$

$$\Sigma k_i = \frac{\Sigma x_i}{\Sigma x_i^2}, \text{ and therefore, } \Sigma k_i^2 = \frac{\Sigma x_i^2}{(\Sigma x_i^2)^2} = \frac{1}{\Sigma x_i^2}$$

$$\boxed{\therefore \text{var}(\hat{\beta}) = \sigma^2 \Sigma k_i^2 = \frac{\sigma^2}{\Sigma x_i^2}} \dots \dots \dots (2.26)$$

b. Variance of $\hat{\alpha}$

$$\begin{aligned}
 \text{var}(\hat{\alpha}) &= E((\hat{\alpha} - E(\alpha))^2) \\
 &= E(\hat{\alpha} - \alpha)^2 \dots \dots \dots (2.27)
 \end{aligned}$$

Substituting equation (2.23) in (2.27), we get

$$\begin{aligned}
 \text{var}(\hat{\alpha}) &= E\left[\Sigma \left(\frac{1}{n} - \bar{X} k_i\right)^2 u_i^2\right] \\
 &= \Sigma \left(\frac{1}{n} - \bar{X} k_i\right)^2 E(u_i)^2 \\
 &= \sigma^2 \Sigma \left(\frac{1}{n} - \bar{X} k_i\right)^2 \\
 &= \sigma^2 \Sigma \left(\frac{1}{n^2} - \frac{2}{n} \bar{X} k_i + \bar{X}^2 k_i^2\right) \\
 &= \sigma^2 \Sigma \left(\frac{1}{n} - \frac{2\bar{X}}{n} \Sigma k_i + \bar{X}^2 \Sigma k_i^2\right), \text{ Since } \Sigma k_i = 0 \\
 &= \sigma^2 \left(\frac{1}{n} + \bar{X}^2 \Sigma k_i^2\right) \\
 &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\Sigma x_i^2}\right), \text{ Since } \Sigma k_i^2 = \frac{\Sigma x_i^2}{(\Sigma x_i^2)^2} = \frac{1}{\Sigma x_i^2}
 \end{aligned}$$

Again:

$$\frac{1}{n} + \frac{\bar{X}^2}{\Sigma x_i^2} = \frac{\Sigma x_i^2 + n\bar{X}^2}{n\Sigma x_i^2} = \left(\frac{\Sigma X^2}{n\Sigma x_i^2}\right)$$

$$\therefore \text{var}(\hat{\alpha}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\Sigma x_i^2}\right) = \sigma^2 \left(\frac{\Sigma X_i^2}{n\Sigma x_i^2}\right) \dots \dots \dots (2.28)$$

Dear student! We have computed the variances OLS estimators. Now, it is time to check whether these variances of OLS estimators do possess minimum variance property compared to the variances other estimators of the true α and β , other than $\hat{\alpha}$ and $\hat{\beta}$.

To establish that $\hat{\alpha}$ and $\hat{\beta}$ possess minimum variance property, we compare their variances with that of the variances of some other alternative linear and unbiased estimators of α and β , say α^* and β^* . Now, we want to prove that any other linear and unbiased estimator of the true population parameter obtained from any other econometric method has larger variance than that OLS estimators.

Lets first show minimum variance of $\hat{\beta}$ and then that of $\hat{\alpha}$.

1. Minimum variance of $\hat{\beta}$

Suppose: β^* an alternative linear and unbiased estimator of β and;

$$\text{Let } \beta^* = \sum w_i Y_i \dots\dots\dots(2.29)$$

where, $w_i \neq k_i$; but: $w_i = k_i + c_i$

$$\begin{aligned}\beta^* &= \sum w_i (\alpha + \beta X_i + u_i) \text{ Since } Y_i = \alpha + \beta X_i + U_i \\ &= \alpha \sum w_i + \beta \sum w_i X_i + \sum w_i u_i\end{aligned}$$

$$\therefore E(\beta^*) = \alpha \sum w_i + \beta \sum w_i X_i, \text{ since } E(u_i) = 0$$

Since β^* is assumed to be an unbiased estimator, then for β^* to be an unbiased estimator of β , there must be true that $\sum w_i = 1$ and $\sum w_i X_i = X$ in the above equation.

$$\text{But, } w_i = k_i + c_i$$

$$\sum w_i = \sum (k_i + c_i) = \sum k_i + \sum c_i$$

$$\text{Therefore, } \sum c_i = 0 \text{ since } \sum k_i = \sum w_i = 1$$

$$\text{Again } \sum w_i X_i = \sum (k_i + c_i) X_i = \sum k_i X_i + \sum c_i X_i$$

$$\text{Since } \sum w_i X_i = X \text{ and } \sum k_i X_i = X \Rightarrow \sum c_i X_i = 0.$$

From these values we can drive $\sum c_i x_i = 0$, where $x_i = X_i - \bar{X}$

$$\sum c_i x_i = \sum c_i (X_i - \bar{X}) = \sum c_i X_i - \bar{X} \sum c_i$$

$$\text{Since } \sum c_i x_i = 1 \quad \sum c_i = 0 \Rightarrow \sum c_i x_i = 0$$

Thus, from the above calculations we can summarize the following results.

$$\sum w_i = 0, \quad \sum w_i x_i = 1, \quad \sum c_i = 0, \quad \sum c_i X_i = 0$$

To prove whether $\hat{\beta}$ has minimum variance or not let's compute $\text{var}(\beta^*)$ to compare with $\text{var}(\hat{\beta})$.

$$\text{var}(\beta^*) = \text{var}(\sum w_i Y_i)$$

$$= \sum w_i^2 \text{var}(Y_i)$$

$$\therefore \text{var}(\beta^*) = \sigma^2 \sum w_i^2 \quad \text{since } \text{Var}(Y_i) = \sigma^2$$

$$\text{But, } \sum w_i^2 = \sum (k_i + c_i)^2 = \sum k_i^2 + 2 \sum k_i c_i + \sum c_i^2$$

$$\Rightarrow \sum w_i^2 = \sum k_i^2 + \sum c_i^2 \quad \text{Since } \sum k_i c_i = \frac{\sum c_i x_i}{\sum x_i^2} = 0$$

$$\text{Therefore, } \text{var}(\beta^*) = \sigma^2 (\sum k_i^2 + \sum c_i^2) \Rightarrow \sigma^2 \sum k_i^2 + \sigma^2 \sum c_i^2$$

$$\boxed{\text{var}(\beta^*) = \text{var}(\hat{\beta}) + \sigma^2 \sum c_i^2}$$

Given that c_i is an arbitrary constant, $\sigma^2 \sum c_i^2$ is a positive i.e it is greater than zero.

Thus $\text{var}(\beta^*) > \text{var}(\hat{\beta})$. This proves that $\hat{\beta}$ possesses minimum variance property.

In the similar way we can prove that the least square estimate of the constant intercept ($\hat{\alpha}$) possesses minimum variance.

2. Minimum Variance of $\hat{\alpha}$

We take a new estimator α^* , which we assume to be a linear and unbiased estimator of function of α . The least square estimator $\hat{\alpha}$ is given by:

$$\hat{\alpha} = \sum (\frac{1}{n} - \bar{X} k_i) Y_i$$

By analogy with that the proof of the minimum variance property of $\hat{\beta}$, let's use the weights $w_i = c_i + k_i$ Consequently;

$$\alpha^* = \sum (\frac{1}{n} - \bar{X} w_i) Y_i$$

Since we want α^* to be an unbiased estimator of the true α , that is, $E(\alpha^*) = \alpha$, we substitute for $Y = \alpha + \beta x_i + u_i$ in α^* and find the expected value of α^* .

$$\begin{aligned}\alpha^* &= \Sigma(\frac{1}{n} - \bar{X}w_i)(\alpha + \beta X_i + u_i) \\ &= \Sigma(\frac{\alpha}{n} + \frac{\beta X}{n} + \frac{u_i}{n} - \bar{X}w_i\alpha - \beta\bar{X}X_iw_i - \bar{X}w_iu_i) \\ \alpha^* &= \alpha + \beta\bar{X} + \sum ui/n - \alpha\bar{X}\Sigma w_i - \beta\bar{X}\Sigma w_iX_i - \bar{X}\Sigma w_iu_i\end{aligned}$$

For α^* to be an unbiased estimator of the true α , the following must hold.

$$\sum (w_i) = 0, \quad \Sigma(w_iX_i) = 1 \quad \text{and} \quad \sum (w_iu_i) = 0$$

i.e., if $\Sigma w_i = 0$, and $\Sigma w_iX_i = 1$. These conditions imply that $\Sigma c_i = 0$ and $\Sigma c_iX_i = 0$.

As in the case of $\hat{\beta}$, we need to compute $Var(\alpha^*)$ to compare with $var(\hat{\alpha})$

$$\begin{aligned}var(\alpha^*) &= var(\Sigma(\frac{1}{n} - \bar{X}w_i)Y_i) \\ &= \Sigma(\frac{1}{n} - \bar{X}w_i)^2 var(Y_i) \\ &= \sigma^2 \Sigma(\frac{1}{n} - \bar{X}w_i)^2 \\ &= \sigma^2 \Sigma(\frac{1}{n^2} + \bar{X}^2 w_i^2 - 2\frac{1}{n}\bar{X}w_i) \\ &= \sigma^2 (\frac{1}{n^2} + \Sigma \bar{X}^2 w_i^2 - 2\bar{X}\frac{1}{n}\Sigma w_i) \\ var(\alpha^*) &= \sigma^2 (\frac{1}{n} + \bar{X}^2 \Sigma w_i^2) \quad , \text{Since } \Sigma w_i = 0 \\ \text{but } \Sigma w_i^2 &= \Sigma k_i^2 + \Sigma c_i^2 \\ \Rightarrow var(\alpha^*) &= \sigma^2 (\frac{1}{n} + \bar{X}^2 (\Sigma k_i^2 + \Sigma c_i^2)) \\ var(\alpha^*) &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\Sigma x_i^2} \right) + \sigma^2 \bar{X}^2 \Sigma c_i^2 \\ &= \sigma^2 \left(\frac{\Sigma X_i^2}{n \Sigma x_i^2} \right) + \sigma^2 \bar{X}^2 \Sigma c_i^2\end{aligned}$$

The first term in the bracket is $var(\hat{\alpha})$, hence

$$\boxed{var(\alpha^*) = var(\hat{\alpha}) + \sigma^2 \bar{X}^2 \Sigma c_i^2}$$

$$\Rightarrow var(\alpha^*) > var(\hat{\alpha}) \quad , \text{Since } \sigma^2 \bar{X}^2 \Sigma c_i^2 > 0$$

Therefore, we have proved that the least square estimators of linear regression model are best, linear and unbiased (BLU) estimators.

The variance of the random variable (U_i)

Dear student! You may observe that the variances of the OLS estimates involve σ^2 , which is the population variance of the random disturbance term. But it is difficult to obtain the population data of the disturbance term because of technical and economic reasons. Hence it is difficult to compute σ^2 ; this implies that variances of OLS estimates are also difficult to compute. But we can compute these variances if we take the unbiased estimate of σ^2 which is $\hat{\sigma}^2$ computed from the sample value of the disturbance term e_i from the expression:

$$\hat{\sigma}_u^2 = \frac{\sum e_i^2}{n-2} \dots\dots\dots 2.30$$

To use $\hat{\sigma}^2$ in the expressions for the variances of $\hat{\alpha}$ and $\hat{\beta}$, we have to prove

whether $\hat{\sigma}^2$ is the unbiased estimator of σ^2 , i.e., $E(\hat{\sigma}^2) = E\left(\frac{\sum e_i^2}{n-2}\right) = \sigma^2$

To prove this we have to compute $\sum e_i^2$ from the expressions of Y , $\hat{Y}$, y , $\hat{y}$ and e_i .

Proof:

$$\left. \begin{array}{l} Y_i = \hat{\alpha} + \hat{\beta}X_i + e_i \\ \hat{Y} = \hat{\alpha} + \hat{\beta}x \end{array} \right\} \Rightarrow Y = \hat{Y} + e_i \dots\dots\dots (2.31)$$

$$\Rightarrow e_i = Y_i - \hat{Y} \dots\dots\dots (2.32)$$

Summing (2.31) will result the following expression

$$\begin{aligned} \sum Y_i &= \sum y_i + \sum e_i \\ \sum Y_i &= \sum \hat{Y}_i \quad \text{since } (\sum e_i) = 0 \end{aligned}$$

Dividing both sides the above by 'n' will give us

$$\frac{\sum Y}{n} = \frac{\sum \hat{Y}_i}{n} \rightarrow \bar{Y} = \bar{\hat{Y}} \text{-----}(2.33)$$

Putting (2.31) and (2.33) together and subtract

$$\left. \begin{array}{l} Y = \hat{Y} + e \\ \bar{Y} = \bar{\hat{Y}} \end{array} \right\} \\ \Rightarrow (Y - \bar{Y}) = (\hat{Y} - \bar{\hat{Y}}) + e \\ \Rightarrow y_i = \hat{y}_i + e \text{.....}(2.34)$$

From (2.34):

$$e_i = y_i - \hat{y}_i \text{.....}(2.35)$$

Where the y's are in deviation form.

Now, we have to express y_i and $\hat{y}_i$ in other expression as derived below.

$$\left. \begin{array}{l} \text{From: } Y_i = \alpha + \beta X_i + U_i \\ \bar{Y} = \alpha + \beta \bar{X} + \bar{U} \end{array} \right\}$$

We get, by subtraction

$$\begin{aligned} y_i &= (Y_i - \bar{Y}) = \beta_i(X_i - \bar{X}) + (U_i - \bar{U}) = \beta x_i + (U - \bar{U}) \\ \Rightarrow y_i &= \beta x + (U - \bar{U}) \text{.....}(2.36) \end{aligned}$$

Note that we assumed earlier that $E(u) = 0$, i.e in taking a very large number samples we expect U to have a mean value of zero, but in any particular single sample $\bar{U}$ is not necessarily zero.

Similarly: From;

$$\left. \begin{array}{l} \hat{Y} = \hat{\alpha} + \hat{\beta}x \\ \bar{\hat{Y}} = \hat{\alpha} + \hat{\beta}\bar{x} \end{array} \right\}$$

We get, by subtraction

$$\begin{aligned} \hat{Y} - \bar{\hat{Y}} &= \hat{\beta}(X - \bar{X}) \\ \Rightarrow \hat{y} &= \hat{\beta}x \text{.....}(2.37) \end{aligned}$$

Substituting (2.36) and (2.37) in (2.35) we get

$$\begin{aligned} e_i &= \beta x_i + (u_i - \bar{u}) - \hat{\beta} x_i \\ &= (u_i - \bar{u}) - (\hat{\beta} - \beta) x_i \end{aligned}$$

The summation over the n sample values of the squares of the residuals over the 'n' samples yields:

$$\begin{aligned} \Sigma e_i^2 &= \Sigma [(u_i - \bar{u}) - (\hat{\beta} - \beta) x_i]^2 \\ &= \Sigma [(u_i - \bar{u})^2 + (\hat{\beta} - \beta)^2 x_i^2 - 2(u_i - \bar{u})(\hat{\beta} - \beta) x_i] \\ &= \Sigma (u_i - \bar{u})^2 + (\hat{\beta} - \beta)^2 \Sigma x_i^2 - 2[(\hat{\beta} - \beta) \Sigma x_i (u_i - \bar{u})] \end{aligned}$$

Taking expected values we have:

$$E(\Sigma e_i^2) = E[\Sigma (u_i - \bar{u})^2] + E[(\hat{\beta} - \beta)^2 \Sigma x_i^2] - 2E[(\hat{\beta} - \beta) \Sigma x_i (u_i - \bar{u})] \dots\dots\dots (2.38)$$

The right hand side terms of (2.38) may be rearranged as follows

$$\begin{aligned} \text{a. } E[\Sigma (u - \bar{u})^2] &= E(\Sigma u_i^2 - \bar{u} \Sigma u_i) \\ &= E\left(\Sigma u_i^2 - \frac{(\Sigma u_i)^2}{n}\right) \\ &= \Sigma E(u_i^2) - \frac{1}{n} E(\Sigma u)^2 \\ &= n\sigma^2 - \frac{1}{n} E(u_1 + u_2 + \dots + u_i)^2 \text{ since } E(u_i^2) = \sigma_u^2 \\ &= n\sigma^2 - \frac{1}{n} E(\Sigma u_i^2 + 2\Sigma u_i u_j) \\ &= n\sigma^2 - \frac{1}{n} (\Sigma E(u_i^2) + 2\Sigma u_i u_j) \quad i \neq j \\ &= n\sigma^2 - \frac{1}{n} n\sigma_u^2 - \frac{2}{n} \Sigma E(u_i u_j) \\ &= n\sigma_u^2 - \sigma_u^2 \text{ (given } E(u_i u_j) = 0) \\ &= \sigma_u^2 (n-1) \dots\dots\dots (2.39) \end{aligned}$$

$$\text{b. } E[(\hat{\beta} - \beta)^2 \Sigma x_i^2] = \Sigma x_i^2 E(\hat{\beta} - \beta)^2$$

Given that the X's are fixed in all samples and we know that

$$E(\hat{\beta} - \beta)^2 = \text{var}(\hat{\beta}) = \sigma_u^2 \frac{1}{\Sigma x^2}$$

$$\text{Hence } \sum x_i^2 \cdot E(\hat{\beta} - \beta)^2 = \sum x_i^2 \cdot \sigma_u^2 \frac{1}{\sum x_i^2}$$

$$\boxed{\sum x_i^2 \cdot E(\hat{\beta} - \beta)^2 = \sigma_u^2} \dots \dots \dots (2.40)$$

$$\begin{aligned} \text{c. } -2 E[(\hat{\beta} - \beta) \sum x_i (u_i - \bar{u})] &= -2 E[(\hat{\beta} - \beta)(\sum x_i u_i - \bar{u} \sum x_i)] \\ &= -2 E[(\hat{\beta} - \beta)(\sum x_i u_i)] , \text{ since } \sum x_i = 0 \end{aligned}$$

But from (2.22) , $(\hat{\beta} - \beta) = \sum k_i u_i$ and substitute it in the above expression, we will get:

$$\begin{aligned} -2 E[(\hat{\beta} - \beta) \sum x_i (u_i - \bar{u})] &= -2 E(\sum k_i u_i)(\sum x_i u_i)] \\ &= -2 E \left[\left(\frac{\sum x_i u_i}{\sum x_i^2} \right) (\sum x_i u_i) \right] , \text{ since } k_i = \frac{x_i}{\sum x_i^2} \\ &= -2 E \left[\frac{(\sum x_i u_i)^2}{\sum x_i^2} \right] \\ &= -2 E \left[\frac{\sum x_i^2 u_i^2 + 2 \sum x_i x_j u_i u_j}{\sum x_i^2} \right] \\ &= -2 \left[\frac{\sum x_i^2 E(u_i^2)}{\sum x_i^2} + \frac{2 \sum (x_i x_j) E(u_i u_j)}{\sum x_i^2} i \neq j \right] \\ &= -2 \frac{\sum x_i^2 E(u_i^2)}{\sum x_i^2} \quad (\text{given } E(u_i u_j) = 0) \end{aligned}$$

$$\boxed{= -2 E(u_i^2) = -2 \sigma_u^2} \dots \dots \dots (2.41)$$

Consequently, Equation (2.38) can be written in terms of (2.39), (2.40) and (2.41) as follows: $E(\sum e_i^2) = (n-1)\sigma_u^2 + \sigma^2 - 2\sigma_u^2 = (n-2)\sigma_u^2 \dots \dots \dots (2.42)$

From which we get

$$E \left(\frac{\sum e_i^2}{n-2} \right) = E(\hat{\sigma}_u^2) = \sigma_u^2 \dots \dots \dots (2.43)$$

$$\text{Since } \hat{\sigma}_u^2 = \frac{\sum e_i^2}{n-2}$$

Thus, $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$ is unbiased estimate of the true variance of the error term (σ^2).

Dear student! The conclusion that we can drive from the above proof is that we can substitute $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$ for (σ^2) in the variance expression of $\hat{\alpha}$ and $\hat{\beta}$, since $E(\hat{\sigma}^2) = \sigma^2$. Hence the formula of variance of $\hat{\alpha}$ and $\hat{\beta}$ becomes;

$$Var(\hat{\beta}) = \frac{\hat{\sigma}^2}{\sum x_i^2} = \frac{\sum e_i^2}{(n-2)\sum x_i^2} \dots\dots\dots(2.44)$$

$$Var(\hat{\alpha}) = \hat{\sigma}^2 \left(\frac{\sum X_i^2}{n\sum x_i^2} \right) = \frac{\sum e_i^2 \sum X_i^2}{n(n-2)\sum x_i^2} \dots\dots\dots(2.45)$$

Note: $\sum e_i^2$ can be computed as $\sum e_i^2 = \sum y_i^2 - \hat{\beta} \sum x_i y_i$.

Dear Student! Do not worry about the derivation of this expression! we will perform the derivation of it in our subsequent subtopic.

2.2.2.4. Statistical test of Significance of the OLS Estimators (First Order tests)

After the estimation of the parameters and the determination of the least square regression line, we need to know how ‘good’ is the fit of this line to the sample observation of Y and X, that is to say we need to measure the dispersion of observations around the regression line. This knowledge is essential because the closer the observation to the line, the better the goodness of fit, i.e. the better is the explanation of the variations of Y by the changes in the explanatory variables.

We divide the available criteria into three groups: *the theoretical a priori criteria*, *the statistical criteria*, and *the econometric criteria*. Under this section, our focus is on statistical criteria (first order tests). The two most commonly used first order tests in econometric analysis are:

- i. The coefficient of determination (the square of the correlation coefficient i.e. R^2). This test is used for judging the explanatory power of the independent variable(s).
- ii. The standard error tests of the estimators. This test is used for judging the statistical reliability of the estimates of the regression coefficients.

1. TESTS OF THE 'GOODNESS OF FIT' WITH R^2

r^2 shows the percentage of total variation of the dependent variable that can be explained by the changes in the explanatory variable(s) included in the model. To elaborate this let's draw a horizontal line corresponding to the mean value of the dependent variable $\bar{Y}$. (see figure 'd' below). By fitting the line $\hat{Y} = \hat{\alpha}_0 + \hat{\beta}_1 X$ we try to obtain the explanation of the variation of the dependent variable Y produced by the changes of the explanatory variable X.

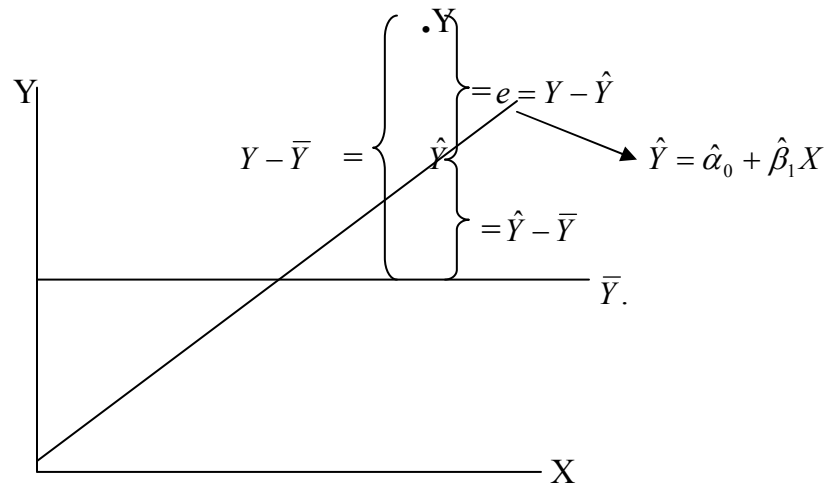


Figure 'd'. Actual and estimated values of the dependent variable Y.

As can be seen from fig.(d) above, $Y - \bar{Y}$ represents measures the variation of the sample observation value of the dependent variable around the mean. However the variation in Y that can be attributed the influence of X, (i.e. the regression line) is given by the vertical distance $\hat{Y} - \bar{Y}$. The part of the total variation in Y about

$\bar{Y}$ that can't be attributed to X is equal to $e = Y - \hat{Y}$ which is referred to as the residual variation.

In summary:

$e_i = Y_i - \hat{Y} =$ deviation of the observation Y_i from the regression line.

$y_i = Y - \bar{Y} =$ deviation of Y from its mean.

$\hat{y} = \hat{Y} - \bar{Y} =$ deviation of the regressed (predicted) value ($\hat{Y}$) from the mean.

Now, we may write the observed Y as the sum of the predicted value ($\hat{Y}$) and the residual term (e_i).

$$\underbrace{Y_i}_{\text{Observed } Y_i} = \underbrace{\hat{Y}}_{\text{predicted } Y_i} + \underbrace{e_i}_{\text{Residual}}$$

From equation (2.34) we can have the above equation but in deviation form

$y = \hat{y} + e$. By squaring and summing both sides, we obtain the following expression:

$$\Sigma y^2 = \Sigma (\hat{y}^2 + e^2)$$

$$\Sigma y^2 = \Sigma (\hat{y}^2 + e_i^2 + 2\hat{y}e_i)$$

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

$$\text{But } \Sigma \hat{y}e_i = \Sigma e(\hat{Y} - \bar{Y}) = \Sigma e(\hat{\alpha} + \hat{\beta}x_i - \bar{Y})$$

$$= \hat{\alpha}\Sigma e_i + \hat{\beta}\Sigma ex_i - \hat{Y}\Sigma e_i$$

$$(\text{but } \Sigma e_i = 0, \Sigma ex_i = 0)$$

$$\Rightarrow \Sigma \hat{y}e = 0 \dots\dots\dots(2.46)$$

Therefore;

$$\boxed{\underbrace{\Sigma y_i^2}_{\text{Total variation}} = \underbrace{\Sigma \hat{y}^2}_{\text{Explained variation}} + \underbrace{\Sigma e_i^2}_{\text{Unexplained variation}}} \dots\dots\dots(2.47)$$

OR,

$$\underbrace{\text{Total sum of square}}_{TSS} = \underbrace{\text{Explained sum of square}}_{ESS} + \underbrace{\text{Residual sum of square}}_{RSS}$$

i.e

$$TSS = ESS + RSS \dots\dots\dots(2.48)$$

Mathematically; the explained variation as a percentage of the total variation is explained as:

$$\frac{ESS}{TSS} = \frac{\Sigma \hat{y}^2}{\Sigma y^2} \dots\dots\dots(2.49)$$

From equation (2.37) we have $\hat{y} = \hat{\beta}x$. Squaring and summing both sides give us

$$\Sigma \hat{y}^2 = \hat{\beta}^2 \Sigma x^2 \text{-----} (2.50)$$

We can substitute (2.50) in (2.49) and obtain:

$$ESS / TSS = \frac{\hat{\beta}^2 \Sigma x^2}{\Sigma y^2} \dots\dots\dots(2.51)$$

$$= \left(\frac{\Sigma xy}{\Sigma x^2} \right)^2 \frac{\Sigma x_i^2}{\Sigma y^2}, \text{ Since } \hat{\beta} = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$$

$$= \frac{\Sigma xy}{\Sigma x^2} \frac{\Sigma xy}{\Sigma y^2} \dots\dots\dots(2.52)$$

Comparing (2.52) with the formula of the correlation coefficient:

$$r = \text{Cov}(X,Y) / \sigma_x^2 \sigma_y^2 = \Sigma xy / n \sigma_x^2 \sigma_y^2 = \Sigma xy / (\Sigma x^2 \Sigma y^2)^{1/2} \dots\dots\dots(2.53)$$

$$\text{Squaring (2.53) will result in: } r^2 = (\Sigma xy)^2 / (\Sigma x^2 \Sigma y^2) \dots\dots\dots(2.54)$$

Comparing (2.52) and (2.54), we see exactly the expressions. Therefore:

$$ESS/TSS = \frac{\Sigma xy}{\Sigma x^2} \frac{\Sigma xy}{\Sigma y^2} = r^2$$

From (2.48), $RSS=TSS-ESS$. Hence R^2 becomes;

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\Sigma e_i^2}{\Sigma y^2} \dots\dots\dots(2.55)$$

From equation (2.55) we can drive;

$$RSS = \sum e_i^2 = \sum y_i^2 (1 - R^2) \text{------(2.56)}$$

The limit of R^2 : The value of R^2 falls between zero and one. i.e. $0 \leq R^2 \leq 1$.

• **Interpretation of R^2**

Suppose $R^2 = 0.9$, this means that the regression line gives a good fit to the observed data since this line explains 90% of the total variation of the Y value around their mean. The remaining 10% of the total variation in Y is unaccounted for by the regression line and is attributed to the factors included in the disturbance variable u_i .

Check yourself question:

- Show that $0 \leq R^2 \leq 1$.
- Show that the square of the coefficient of correlation is equal to ESS/TSS.

Exercise:

Suppose r_{xy} = is the correlation coefficient between Y and X and is give by:

$$= \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \sqrt{\sum y_i^2}}$$

And let $r_{y\hat{y}}^2$ = the square of the correlation coefficient between Y and $\hat{Y}$, and is

given by:
$$r_{y\hat{y}}^2 = \frac{(\sum y\hat{y})^2}{\sum y^2 \sum \hat{y}^2}$$

Show that: i) $r_{y\hat{y}}^2 = R^2$ ii) $r_{yy} = r_{yx}$

2. TESTING THE SIGNIFICANCE OF OLS PARAMETERS

To test the significance of the OLS parameter estimators we need the following:

- Variance of the parameter estimators
- Unbiased estimator of σ^2

- The assumption of normality of the distribution of error term.

We have already derived that:

- $\text{var}(\hat{\beta}) = \frac{\hat{\sigma}^2}{\sum x^2}$
- $\text{var}(\hat{\alpha}) = \frac{\hat{\sigma}^2 \sum X^2}{n \sum x^2}$
- $\hat{\sigma}^2 = \frac{\sum e^2}{n-2} = \frac{RSS}{n-2}$

For the purpose of estimation of the parameters the assumption of normality is not used, but we use this assumption to test the significance of the parameter estimators; because the testing methods or procedures are based on the assumption of the normality assumption of the disturbance term. Hence before we discuss on the various testing methods it is important to see whether the parameters are normally distributed or not.

We have already assumed that the error term is normally distributed with mean zero and variance σ^2 , i.e. $U_i \sim N(0, \sigma^2)$. Similarly, we also proved that $Y_i \sim N[(\alpha + \beta x), \sigma^2]$. Now, we want to show the following:

1. $\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum x^2}\right)$
2. $\hat{\alpha} \sim N\left(\alpha, \frac{\sigma^2 \sum X^2}{n \sum x^2}\right)$

To show whether $\hat{\alpha}$ and $\hat{\beta}$ are normally distributed or not, we need to make use of one property of normal distribution. “..... any linear function of a normally distributed variable is itself normally distributed.”

$$\hat{\beta} = \sum k_i Y_i = k_1 Y_1 + k_2 Y_2 + \dots + k_n Y_n$$

$$\hat{\alpha} = \sum w_i Y_i = w_1 Y_1 + w_2 Y_2 + \dots + w_n Y_n$$

Since $\hat{\alpha}$ and $\hat{\beta}$ are linear in Y , it follows that

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum x^2}\right) \quad ; \quad \hat{\alpha} \sim N\left(\alpha, \frac{\sigma^2 \sum X^2}{n \sum x^2}\right)$$

The OLS estimates $\hat{\alpha}$ and $\hat{\beta}$ are obtained from a sample of observations on Y and X. Since sampling errors are inevitable in all estimates, it is necessary to apply test of significance in order to measure the size of the error and determine the degree of confidence in order to measure the validity of these estimates. This can be done by using various tests. The most common ones are:

i) Standard error test ii) Student's t-test iii) Confidence interval

All of these testing procedures reach on the same conclusion. Let us now see these testing methods one by one.

i) Standard error test

This test helps us decide whether the estimates $\hat{\alpha}$ and $\hat{\beta}$ are significantly different from zero, i.e. whether the sample from which they have been estimated might have come from a population whose true parameters are zero. $\alpha = 0$ and / or $\beta = 0$.

Formally we test the null hypothesis

$$H_0 : \beta_i = 0 \text{ against the alternative hypothesis } H_1 : \beta_i \neq 0$$

The standard error test may be outlined as follows.

First: Compute standard error of the parameters.

$$SE(\hat{\beta}) = \sqrt{\text{var}(\hat{\beta})}$$

$$SE(\hat{\alpha}) = \sqrt{\text{var}(\hat{\alpha})}$$

Second: compare the standard errors with the numerical values of $\hat{\alpha}$ and $\hat{\beta}$.

Decision rule:

- If $SE(\hat{\beta}_i) > \frac{1}{2}\hat{\beta}_i$, accept the null hypothesis and reject the alternative hypothesis. We conclude that $\hat{\beta}_i$ is statistically insignificant.

- If $SE(\hat{\beta}_i) < \frac{1}{2}\hat{\beta}_i$, reject the null hypothesis and accept the alternative hypothesis. We conclude that $\hat{\beta}_i$ is statistically significant.

The acceptance or rejection of the null hypothesis has definite economic meaning. Namely, the acceptance of the null hypothesis $\beta = 0$ (the slope parameter is zero) implies that the explanatory variable to which this estimate relates does not in fact influence the dependent variable Y and should not be included in the function, since the conducted test provided evidence that changes in X leave Y unaffected. In other words acceptance of H_0 implies that the relation ship between Y and X is in fact $Y = \alpha + (0)x = \alpha$, i.e. there is no relationship between X and Y.

Numerical example: Suppose that from a sample of size $n=30$, we estimate the following supply function.

$$Q = 120 + 0.6p + e_i$$
$$SE : (1.7) \quad (0.025)$$

Test the significance of the slope parameter at 5% level of significance using the standard error test.

$$SE(\hat{\beta}) = 0.025$$

$$(\hat{\beta}) = 0.6$$

$$\frac{1}{2}\hat{\beta} = 0.3$$

This implies that $SE(\hat{\beta}_i) < \frac{1}{2}\hat{\beta}_i$. The implication is $\hat{\beta}$ is statistically significant at 5% level of significance.

Note: The standard error test is an approximated test (which is approximated from the z-test and t-test) and implies a two tail test conducted at 5% level of significance.

ii) **Student's t-test**

Like the standard error test, this test is also important to test the significance of the parameters. From your statistics, any variable X can be transformed into t using the general formula:

$$t = \frac{X - \mu}{s_x}, \text{ with } n-1 \text{ degree of freedom.}$$

Where μ_i = value of the population mean

s_x = sample estimate of the population standard deviation

$$s_x = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$$

n = sample size

We can derive the t-value of the OLS estimates

$$\left. \begin{aligned} t_{\hat{\beta}} &= \frac{\hat{\beta}_i - \beta}{SE(\hat{\beta})} \\ t_{\hat{\alpha}} &= \frac{\hat{\alpha} - \alpha}{SE(\hat{\alpha})} \end{aligned} \right\} \text{ with } n-k \text{ degree of freedom.}$$

Where:

SE = is standard error

k = number of parameters in the model.

Since we have two parameters in simple linear regression with intercept different from zero, our degree of freedom is $n-2$. Like the standard error test we formally test the hypothesis: $H_0 : \beta_i = 0$ against the alternative $H_1 : \beta_i \neq 0$ for the slope parameter; and $H_0 : \alpha = 0$ against the alternative $H_1 : \alpha \neq 0$ for the intercept.

To undertake the above test we follow the following steps.

Step 1: Compute t^* , which is called the computed value of t , by taking the value of β in the null hypothesis. In our case $\beta = 0$, then t^* becomes:

$$t^* = \frac{\hat{\beta} - 0}{SE(\hat{\beta})} = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

Step 2: Choose level of significance. Level of significance is the probability of making ‘wrong’ decision, i.e. the probability of rejecting the hypothesis when it is actually true or the probability of committing a type I error. It is customary in

econometric research to choose the 5% or the 1% level of significance. This means that in making our decision we allow (tolerate) five times out of a hundred to be 'wrong' i.e. reject the hypothesis when it is actually true.

Step 3: Check whether there is one tail test or two tail test. If the inequality sign in the alternative hypothesis is $\neq$, then it implies a two tail test and divide the chosen level of significance by two; decide the critical reject or critical value of t called t_c . But if the inequality sign is either $>$ or $<$ then it indicates one tail test and there is no need to divide the chosen level of significance by two to obtain the critical value of t from the t-table.

Example:

If we have $H_0 : \beta_i = 0$

against: $H_1 : \beta_i \neq 0$

Then this is a two tail test. If the level of significance is 5%, divide it by two to obtain critical value of t from the t-table.

Step 4: Obtain critical value of t, called t_c at $\alpha/2$ and n-2 degree of freedom for two tail test.

Step 5: Compare t^* (the computed value of t) and t_c (critical value of t)

- If $t^* > t_c$, reject H_0 and accept H_1 . The conclusion is $\hat{\beta}$ is statistically significant.
- If $t^* < t_c$, accept H_0 and reject H_1 . The conclusion is $\hat{\beta}$ is statistically insignificant.

Numerical Example:

Suppose that from a sample size $n=20$ we estimate the following consumption function:

$$C = 100 + 0.70Y + e$$

(75.5) (0.21)

The values in the brackets are standard errors. We want to test the null hypothesis: $H_0 : \beta_i = 0$ against the alternative $H_1 : \beta_i \neq 0$ using the t-test at 5% level of significance.

a. the t-value for the test statistic is:

$$t^* = \frac{\hat{\beta} - 0}{SE(\hat{\beta})} = \frac{\hat{\beta}}{SE(\hat{\beta})} = \frac{0.70}{0.21} \cong 3.3$$

b. Since the alternative hypothesis (H_1) is stated by inequality sign ($\neq$), it is a two tail test, hence we divide $\alpha/2 = 0.05/2 = 0.025$ to obtain the critical value of 't' at $\alpha/2 = 0.025$ and 18 degree of freedom (df) i.e. ($n-2=20-2$). From the t-table 'tc' at 0.025 level of significance and 18 df is 2.10.

c. Since $t^*=3.3$ and $t_c=2.1$, $t^*>t_c$. It implies that $\hat{\beta}$ is statistically significant.

iii) Confidence interval

Rejection of the null hypothesis doesn't mean that our estimate $\hat{\alpha}$ and $\hat{\beta}$ is the correct estimate of the true population parameter α and β . It simply means that our estimate comes from a sample drawn from a population whose parameter β is different from zero.

In order to define how close the estimate to the true parameter, we must construct confidence interval for the true parameter, in other words we must establish limiting values around the estimate within which the true parameter is expected to lie within a certain "degree of confidence". In this respect we say that with a given probability the population parameter will be within the defined confidence interval (confidence limits).

We choose a probability in advance and refer to it as confidence level (interval coefficient). It is customary in econometrics to choose the 95% confidence level. This means that in repeated sampling the confidence limits, computed from the

sample, would include the true population parameter in 95% of the cases. In the other 5% of the cases the population parameter will fall outside the confidence interval.

In a two-tail test at α level of significance, the probability of obtaining the specific t-value either $-t_c$ or t_c is $\alpha/2$ at n-2 degree of freedom. The probability of obtaining

any value of t which is equal to $\frac{\hat{\beta} - \beta}{SE(\hat{\beta})}$ at n-2 degree of freedom is

$1 - (\alpha/2 + \alpha/2)$ i.e. $1 - \alpha$.

$$\text{i.e. } \Pr\{-t_c < t^* < t_c\} = 1 - \alpha \quad \dots\dots\dots(2.57)$$

$$\text{but } t^* = \frac{\hat{\beta} - \beta}{SE(\hat{\beta})} \quad \dots\dots\dots(2.58)$$

Substitute (2.58) in (2.57) we obtain the following expression.

$$\Pr\left\{-t_c < \frac{\hat{\beta} - \beta}{SE(\hat{\beta})} < t_c\right\} = 1 - \alpha \quad \dots\dots\dots(2.59)$$

$$\Pr\{-SE(\hat{\beta})t_c < \hat{\beta} - \beta < SE(\hat{\beta})t_c\} = 1 - \alpha \quad \text{--- by multiplying } SE(\hat{\beta})$$

$$\Pr\{-\hat{\beta} - SE(\hat{\beta})t_c < -\beta < -\hat{\beta} + SE(\hat{\beta})t_c\} = 1 - \alpha \quad \text{--- by subtracting } \hat{\beta}$$

$$\Pr\{+\hat{\beta} + SE(\hat{\beta}) > \beta > \hat{\beta} - SE(\hat{\beta})t_c\} = 1 - \alpha \quad \text{--- by multiplying by } -1$$

$$\Pr\{\hat{\beta} - SE(\hat{\beta})t_c < \beta < \hat{\beta} + SE(\hat{\beta})t_c\} = 1 - \alpha \quad \text{--- interchanging}$$

The limit within which the true β lies at $(1 - \alpha)\%$ degree of confidence is:

$[\hat{\beta} - SE(\hat{\beta})t_c, \hat{\beta} + SE(\hat{\beta})t_c]$; where t_c is the critical value of t at $\alpha/2$ confidence interval and n-2 degree of freedom.

The test procedure is outlined as follows.

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

Decision rule: If the hypothesized value of β in the null hypothesis is within the confidence interval, accept H_0 and reject H_1 . The implication is that $\hat{\beta}$ is

statistically insignificant; while if the hypothesized value of β in the null hypothesis is outside the limit, reject H_0 and accept H_1 . This indicates $\hat{\beta}$ is statistically significant.

Numerical Example:

Suppose we have estimated the following regression line from a sample of 20 observations.

$$Y = 128.5 + 2.88X + e$$
$$(38.2) \quad (0.85)$$

The values in the bracket are standard errors.

- a. Construct 95% confidence interval for the slope of parameter
- b. Test the significance of the slope parameter using constructed confidence interval.

Solution:

- a. The limit within which the true β lies at 95% confidence interval is:

$$\hat{\beta} \pm SE(\hat{\beta})t_c$$

$$\hat{\beta} = 2.88$$

$$SE(\hat{\beta}) = 0.85$$

t_c at 0.025 level of significance and 18 degree of freedom is 2.10.

$$\Rightarrow \hat{\beta} \pm SE(\hat{\beta})t_c = 2.88 \pm 2.10(0.85) = 2.88 \pm 1.79.$$

The confidence interval is:

$$(1.09, 4.67)$$

- b. The value of β in the null hypothesis is zero which implies it is out side the confidence interval. Hence β is statistically significant.

2.2.3 Reporting the Results of Regression Analysis

The results of the regression analysis derived are reported in conventional formats. It is not sufficient merely to report the estimates of β 's. In practice we report

regression coefficients together with their standard errors and the value of R^2 . It has become customary to present the estimated equations with standard errors placed in parenthesis below the estimated parameter values. Sometimes, the estimated coefficients, the corresponding standard errors, the p-values, and some other indicators are presented in tabular form.

These results are supplemented by R^2 on (to the right side of the regression equation).

Example:
$$\begin{array}{c} Y = 128.5 + 2.88 X \\ (38.2) \quad (0.85) \end{array}, \quad R^2 = 0.93.$$
 The numbers in the

parenthesis below the parameter estimates are the standard errors. Some econometricians report the t-values of the estimated coefficients in place of the standard errors.

Review Questions

Review Questions

1. Econometrics deals with the measurement of economic relationships which are stochastic or random. The simplest form of economic relationships between two variables X and Y can be represented by:

$$Y_i = \beta_0 + \beta_1 X_i + U_i ; \text{ where } \beta_0 \text{ and } \beta_1 = \text{are regression parameters and } U_i = \text{the stochastic disturbance term}$$

What are the reasons for the insertion of U-term in the model?

2. The following data refers to the demand for money (M) and the rate of interest (R) in for eight different economies:

M (In billions)	56	50	46	30	20	35	37	61
R%	6.3	4.6	5.1	7.3	8.9	5.3	6.7	3.5

- a. Assuming a relationship $M = \alpha + \beta R + U_i$, obtain the OLS estimators of α and β
- b. Calculate the coefficient of determination for the data and interpret its value
- c. If in a 9th economy the rate of interest is $R=8.1$, predict the demand for money(M) in this economy.

3. The following data refers to the price of a good 'P' and the quantity of the good supplied, 'S'.

P	2	7	5	1	4	8	2	8
S	15	41	32	9	28	43	17	40

- Estimate the linear regression line $E(S) = \alpha + \beta P$
 - Estimate the standard errors of $\hat{\alpha}$ and $\hat{\beta}$
 - Test the hypothesis that price influences supply
 - Obtain a 95% confidence interval for α
4. The following results have been obtained from a sample of 11 observations on the values of sales (Y) of a firm and the corresponding prices (X).

$$\bar{X} = 519.18$$

$$\bar{Y} = 217.82$$

$$\sum X_i^2 = 3,134,543$$

$$\sum X_i Y_i = 1,296,836$$

$$\sum Y_i^2 = 539,512$$

- Estimate the regression line of sale on price and interpret the results
 - What is the part of the variation in sales which is not explained by the regression line?
 - Estimate the price elasticity of sales.
5. The following table includes the GNP(X) and the demand for food (Y) for a country over ten years period.

year	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
Y	6	7	8	10	8	9	10	9	11	10
X	50	52	55	59	57	58	62	65	68	70

- Estimate the food function
 - Compute the coefficient of determination and find the explained and unexplained variation in the food expenditure.
 - Compute the standard error of the regression coefficients and conduct test of significance at the 5% level of significance.
6. A sample of 20 observation corresponding to the regression model $Y_i = \alpha + \beta X_i + U_i$ gave the following data.

$$\sum Y_i = 21.9$$

$$\sum (Y_i - \bar{Y})^2 = 86.9$$

$$\sum X_i = 186.2$$

$$\sum (X_i - \bar{X})^2 = 215.4$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 106.4$$

- a. Estimate α and β
 - b. Calculate the variance of our estimates
 - c. Estimate the conditional mean of Y corresponding to a value of X fixed at X=10.
7. Suppose that a researcher estimates a consumption function and obtains the following results:

$$C = 15 + 0.81Y_d$$

(3.1) (18.7)

$$n = 19$$

$$R^2 = 0.99$$

where C=Consumption, Y_d =disposable income, and numbers in the parenthesis are the 't-ratios'

- a. Test the significance of Y_d statistically using t-ratios
 - b. Determine the estimated standard deviations of the parameter estimates
8. State and prove Gauss-Markov theorem
9. Given the model:
- $$Y_i = \beta_0 + \beta_1 X_i + U_i$$
- with usual OLS assumptions. Derive the expression for the error variance.

Chapter Three

THE CLASSICAL REGRESSION ANALYSIS

[The Multiple Linear Regression Model]

3.1 Introduction

In simple regression we study the relationship between a dependent variable and a single explanatory (independent variable). But it is rarely the case that economic relationships involve just two variables. Rather a dependent variable Y can depend on a whole series of explanatory variables or regressors. For instance, in demand studies we study the relationship between quantity demanded of a good and price of the good, price of substitute goods and the consumer's income. The model we assume is:

$$Y_i = \beta_0 + \beta_1 P_1 + \beta_2 P_2 + \beta_3 X_i + u_i \text{-----} (3.1)$$

Where Y_i = quantity demanded, P_1 is price of the good, P_2 is price of substitute goods, X_i is consumer's income, and β 's are unknown parameters and u_i is the disturbance.

Equation (3.1) is a multiple regression with three explanatory variables. In general for K -explanatory variable we can write the model as follows:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + u_i \text{-----} (3.2)$$

Where $X_{ki} (i = 1, 2, 3, \dots, K)$ are explanatory variables, Y_i is the dependent variable and $\beta_j (j = 0, 1, 2, \dots, (k+1))$ are unknown parameters and u_i is the disturbance term. The disturbance term is of similar nature to that in simple regression, reflecting:

- the basic random nature of human responses
- errors of aggregation
- errors of measurement
- errors in specification of the mathematical form of the model

and any other (minor) factors, other than x_i that might influence Y.

In this chapter we will first start our discussion with the assumptions of the multiple regressions and we will proceed our analysis with the case of two explanatory variables and then we will generalize the multiple regression model in the case of k-explanatory variables using matrix algebra.

3.2 Assumptions of Multiple Regression Model

In order to specify our multiple linear regression model and proceed our analysis with regard to this model, some assumptions are compulsory. But these assumptions are the same as in the single explanatory variable model developed earlier except the assumption of no perfect multicollinearity. These assumptions are:

1. *Randomness of the error term*: The variable u is a real random variable.
2. *Zero mean of the error term*: $E(u_i) = 0$
3. *Homoscedasticity*: The variance of each u_i is the same for all the x_i values.

$$\text{i.e. } E(u_i^2) = \sigma_u^2 \text{ (constant)}$$

4. *Normality of u* : The values of each u_i are normally distributed.

$$\text{i.e. } U_i \sim N(0, \sigma^2)$$

5. *No auto or serial correlation*: The values of u_i (corresponding to X_i) are independent from the values of any other u_i (corresponding to X_j) for $i \neq j$.

$$\text{i.e. } E(u_i u_j) = 0 \text{ for } x_i \neq j$$

6. *Independence of u_i and X_i* : Every disturbance term u_i is independent of the explanatory variables. i.e. $E(u_i X_{1i}) = E(u_i X_{2i}) = 0$

This condition is automatically fulfilled if we assume that the values of the X's are a set of fixed numbers in all (hypothetical) samples.

7. *No perfect multicollinearity*: The explanatory variables are not perfectly linearly correlated.

We can't exclusively list all the assumptions but the above assumptions are some of the basic assumptions that enable us to proceed our analysis.

3.3 A Model With Two Explanatory Variables

In order to understand the nature of multiple regression model easily, we start our analysis with the case of two explanatory variables, then extend this to the case of k-explanatory variables.

3.3.1 Estimation of parameters of two-explanatory variables model

The model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U_i$ (3.3)

is multiple regression with two explanatory variables. The expected value of the above model is called population regression equation i.e.

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \text{ Since } E(U_i) = 0. \text{(3.4)}$$

where β_i is the population parameters. β_0 is referred to as the intercept and β_1 and β_2 are also some times known as regression slopes of the regression. Note that, β_2 for example measures the effect on $E(Y)$ of a unit change in X_2 when X_1 is held constant.

Since the population regression equation is unknown to any investigator, it has to be estimated from sample data. Let us suppose that the sample data has been used to estimate the population regression equation. We leave the method of estimation unspecified for the present and merely assume that equation (3.4) has been estimated by sample regression equation, which we write as:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 \text{(3.5)}$$

Where $\hat{\beta}_j$ are estimates of the β_j and $\hat{Y}$ is known as the predicted value of Y.

Now it is time to state how (3.3) is estimated. Given sample observation on Y, X_1 & X_2 , we estimate (3.3) using the method of least square (OLS).

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + e_i \dots\dots\dots(3.6)$$

is sample relation between Y, X_1 & X_2 .

$$e_i = Y_i - \hat{Y} = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} \dots\dots\dots(3.7)$$

To obtain expressions for the least square estimators, we partially differentiate $\sum e_i^2$ with respect to $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$ and set the partial derivatives equal to zero.

$$\frac{\partial [\sum e_i^2]}{\partial \hat{\beta}_0} = -2 \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}) = 0 \dots\dots\dots(3.8)$$

$$\frac{\partial [\sum e_i^2]}{\partial \hat{\beta}_1} = -2 \sum X_{1i} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}) = 0 \dots\dots\dots(3.9)$$

$$\frac{\partial [\sum e_i^2]}{\partial \hat{\beta}_2} = -2 \sum X_{2i} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}) = 0 \dots\dots\dots(3.10)$$

Summing from 1 to n, the multiple regression equation produces three **Normal Equations**:

$$\sum Y = n\hat{\beta}_0 + \hat{\beta}_1 \sum X_{1i} + \hat{\beta}_2 \sum X_{2i} \dots\dots\dots(3.11)$$

$$\sum X_{1i} Y_i = \hat{\beta}_0 \sum X_{1i} + \hat{\beta}_1 \sum X_{1i}^2 + \hat{\beta}_2 \sum X_{1i} X_{2i} \dots\dots\dots(3.12)$$

$$\sum X_{2i} Y_i = \hat{\beta}_0 \sum X_{2i} + \hat{\beta}_1 \sum X_{1i} X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 \dots\dots\dots(3.13)$$

From (3.11) we obtain $\hat{\beta}_0$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 \text{-----} (3.14)$$

Substituting (3.14) in (3.12), we get:

$$\begin{aligned} \sum X_{1i} Y_i &= (\bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2) \sum X_{1i} + \hat{\beta}_1 \sum X_{1i}^2 + \hat{\beta}_2 \sum X_{2i} \\ \Rightarrow \sum X_{1i} Y_i - \hat{Y} \sum X_{1i} &= \hat{\beta}_1 (\sum X_{1i}^2 - \bar{X}_1 \sum X_{2i}) + \hat{\beta}_2 (\sum X_{1i} X_{2i} - \bar{X}_2 \sum X_{2i}) \\ \Rightarrow \sum X_{1i} Y_i - n \bar{Y} \bar{X}_1 &= \hat{\beta}_1 (\sum X_{1i}^2 - n \bar{X}_1^2) + \hat{\beta}_2 (\sum X_{1i} X_{2i} - n \bar{X}_1 \bar{X}_2) \text{-----} (3.15) \end{aligned}$$

We know that

$$\sum (X_i - Y_i)^2 = (\sum X_i Y_i - n \bar{X}_i \bar{Y}_i) = \sum x_i y_i$$

$$\sum (X_i - \bar{X}_i)^2 = (\sum X_i^2 - n \bar{X}_i^2) = \sum x_i^2$$

Substituting the above equations in equation (3.14), the normal equation (3.12) can be written in deviation form as follows:

$$\sum x_1 y = \hat{\beta}_1 \sum x_1^2 + \hat{\beta}_2 \sum x_1 x_2 \dots\dots\dots(3.16)$$

Using the above procedure if we substitute (3.14) in (3.13), we get

$$\sum x_2 y = \hat{\beta}_1 \sum x_1 x_2 + \hat{\beta}_2 \sum x_2^2 \dots\dots\dots(3.17)$$

Let's bring (2.17) and (2.18) together

$$\sum x_1 y = \hat{\beta}_1 \sum x_1^2 + \hat{\beta}_2 \sum x_1 x_2 \dots\dots\dots(3.18)$$

$$\sum x_2 y = \hat{\beta}_1 \sum x_1 x_2 + \hat{\beta}_2 \sum x_2^2 \dots\dots\dots(3.19)$$

$\hat{\beta}_1$ and $\hat{\beta}_2$ can easily be solved using matrix

We can rewrite the above two equations in matrix form as follows.

$$\begin{pmatrix} \sum x_1^2 & \sum x_1 x_2 \\ \sum x_1 x_2 & \sum x_2^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \sum x_1 y \\ \sum x_2 y \end{pmatrix} \dots\dots\dots(3.20)$$

If we use Cramer's rule to solve the above matrix we obtain

$$\hat{\beta}_1 = \frac{\sum x_1 y \cdot \sum x_2^2 - \sum x_1 x_2 \cdot \sum x_2 y}{\sum x_1^2 \cdot \sum x_2^2 - (\sum x_1 x_2)^2} \dots\dots\dots(3.21)$$

$$\hat{\beta}_2 = \frac{\sum x_2 y \cdot \sum x_1^2 - \sum x_1 x_2 \cdot \sum x_1 y}{\sum x_1^2 \cdot \sum x_2^2 - (\sum x_1 x_2)^2} \dots\dots\dots(3.22)$$

We can also express $\hat{\beta}_1$ and $\hat{\beta}_2$ in terms of covariance and variances of Y , X_1 and X_2

$$\hat{\beta}_1 = \frac{\text{Cov}(X_1, Y) \cdot \text{Var}(X_2) - \text{Cov}(X_1, X_2) \cdot \text{Cov}(X_2, Y)}{\text{Var}(X_1) \cdot \text{Var}(X_2) - [\text{Cov}(X_1, X_2)]^2} \dots\dots\dots(3.23)$$

$$\hat{\beta}_2 = \frac{\text{Cov}(X_2, Y) \cdot \text{Var}(X_1) - \text{Cov}(X_1, X_2) \cdot \text{Cov}(X_1, Y)}{\text{Var}(X_1) \cdot \text{Var}(X_2) - [\text{Cov}(X_1, X_2)]^2} \dots\dots\dots(3.24)$$

3.3.2 The coefficient of determination (R^2): two explanatory variables case

In the simple regression model, we introduced R^2 as a measure of the proportion of variation in the dependent variable that is explained by variation in the explanatory variable. In multiple regression model the same measure is relevant, and the same formulas are valid but now we talk of the proportion of variation in the dependent variable explained by all explanatory variables included in the model. The coefficient of determination is:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum y_i^2} \text{-----} (3.25)$$

In the present model of two explanatory variables:

$$\begin{aligned} \sum e_i^2 &= \sum (y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2 \\ &= \sum e_i (y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}) \\ &= \sum e_i y_i - \hat{\beta}_1 \sum x_{1i} e_i - \hat{\beta}_2 \sum x_{2i} e_i \\ &= \sum e_i y_i \text{ since } \sum e_i x_{1i} = \sum e_i x_{2i} = 0 \\ &= \sum y_i (y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}) \\ \text{i.e } \sum e_i^2 &= \sum y_i^2 - \hat{\beta}_1 \sum x_{1i} y_i - \hat{\beta}_2 \sum x_{2i} y_i \\ \Rightarrow \underbrace{\sum y_i^2}_{\text{Total sum of square (Total variation)}} &= \underbrace{\hat{\beta}_1 \sum x_{1i} y_i + \hat{\beta}_2 \sum x_{2i} y_i}_{\text{Explained sum of square (Explained variation)}} + \underbrace{\sum e_i^2}_{\text{Residual sum of squares (unexplained variation)}} \text{-----} (3.26) \end{aligned}$$

$$\boxed{\therefore R^2 = \frac{ESS}{TSS} = \frac{\hat{\beta}_1 \sum x_{1i} y_i + \hat{\beta}_2 \sum x_{2i} y_i}{\sum y_i^2}} \text{-----} (3.27)$$

As in simple regression, R^2 is also viewed as a measure of the prediction ability of the model over the sample period, or as a measure of how well the estimated regression fits the data. The value of R^2 is also equal to the squared sample correlation coefficient between $\hat{Y}$ & Y_i . Since the sample correlation coefficient measures the linear association between two variables, if R^2 is high, that means there is a close association between the values of Y_i and the values of predicted by

the model, $\hat{Y}_t$. In this case, the model is said to “fit” the data well. If R^2 is low, there is no association between the values of Y_t and the values predicted by the model, $\hat{Y}_t$ and the model does not fit the data well.

3.3.3 Adjusted Coefficient of Determination ($\bar{R}^2$)

One difficulty with R^2 is that it can be made large by adding more and more variables, even if the variables added have no economic justification. Algebraically, it is the fact that as the variables are added the sum of squared errors (RSS) goes down (it can remain unchanged, but this is rare) and thus R^2 goes up. If the model contains $n-1$ variables then $R^2=1$. The manipulation of model just to obtain a high R^2 is not wise. An alternative measure of goodness of fit, called the adjusted R^2 and often symbolized as $\bar{R}^2$, is usually reported by regression programs. It is computed as:

$$\bar{R}^2 = 1 - \frac{\sum e_i^2 / n - k}{\sum y^2 / n - 1} = 1 - (1 - R^2) \left(\frac{n-1}{n-k} \right) \text{-----(3.28)}$$

This measure does not always goes up when a variable is added because of the degree of freedom term $n-k$ is the numerator. As the number of variables k increases, RSS goes down, but so does $n-k$. The effect on $\bar{R}^2$ depends on the amount by which R^2 falls. While solving one problem, this corrected measure of goodness of fit unfortunately introduces another one. It losses its interpretation; $\bar{R}^2$ is no longer the percent of variation explained. This modified $\bar{R}^2$ is sometimes used and misused as a device for selecting the appropriate set of explanatory variables.

3.4.General Linear Regression Model and Matrix Approach

So far we have discussed the regression models containing one or two explanatory variables. Let us now generalize the model assuming that it contains k variables. It will be of the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + U$$

There are k parameters to be estimated. The system of normal equations consist of k+1 equations, in which the unknowns are the parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ and the known terms will be the sums of squares and the sums of products of all variables in the structural equations.

Least square estimators of the unknown parameters are obtained by minimizing the sum of the squared residuals.

$$\Sigma e_i^2 = \Sigma (y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \dots - \hat{\beta}_k X_{ki})^2$$

With respect to $\beta_j (j = 0, 1, 2, \dots, (k+1))$

The partial derivations are equated to zero to obtain normal equations.

$$\frac{\partial \Sigma e_i^2}{\partial \hat{\beta}_0} = -2 \Sigma (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \dots - \hat{\beta}_k X_{ki}) = 0$$

$$\frac{\partial \Sigma e_i^2}{\partial \hat{\beta}_1} = -2 \Sigma (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \dots - \hat{\beta}_k X_{ki}) (x_{1i}) = 0$$

.....

$$\frac{\partial \Sigma e_i^2}{\partial \hat{\beta}_k} = -2 \Sigma (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \dots - \hat{\beta}_k X_{ki}) (x_{ki}) = 0$$

The general form of the above equations (except first) may be written as:

$$\frac{\partial \Sigma e_i^2}{\partial \hat{\beta}_j} = -2 \Sigma (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \dots - \hat{\beta}_k X_{ki}) = 0; \quad \text{where } (j = 1, 2, \dots, k)$$

The normal equations of the general linear regression model are

$$\begin{aligned} \Sigma Y_i &= n\hat{\beta}_0 + \hat{\beta}_1 \Sigma X_{1i} + \hat{\beta}_2 \Sigma X_{2i} + \dots + \hat{\beta}_k \Sigma X_{ki} \\ \Sigma Y_i X_{1i} &= \hat{\beta}_0 \Sigma X_{1i} + \hat{\beta}_1 \Sigma X_{1i}^2 + \dots + \hat{\beta}_k \Sigma X_{1i} X_{ki} \\ \Sigma Y_i X_{2i} &= \hat{\beta}_0 \Sigma X_{2i} + \hat{\beta}_1 \Sigma X_{1i} X_{2i} + \hat{\beta}_2 \Sigma X_{2i}^2 + \dots + \hat{\beta}_k \Sigma X_{2i} X_{ki} \\ &\vdots \\ &\vdots \\ \Sigma Y_i X_{ki} &= \hat{\beta}_0 \Sigma X_{ki} + \hat{\beta}_1 \Sigma X_{1i} X_{ki} + \dots + \hat{\beta}_k \Sigma X_{ki}^2 \end{aligned}$$

To derive the OLS estimators of β , under the usual (classical) assumptions mentioned earlier, we define two vectors $\hat{\beta}$ and 'e' as:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ . \\ . \\ \hat{\beta}_k \end{bmatrix} \quad \text{and} \quad e = \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ e_n \end{bmatrix}$$

Thus we can write: $Y = X\hat{\beta} + e$ and $e = Y - X\hat{\beta}$

We have to minimize:

$$\begin{aligned} \sum_{i=1}^n e_i^2 &= e_1^2 + e_2^2 + e_3^2 + \dots + e_n^2 \\ &= [e_1, e_2, \dots, e_n] \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ e_n \end{bmatrix} = e'e \\ &= \sum e_i^2 = e'e \\ e'e &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) \\ &= YY' - \hat{\beta}' X' Y - Y' X \hat{\beta} + \hat{\beta}' X' X \hat{\beta} \dots \dots \dots (3.30) \end{aligned}$$

Since $\hat{\beta}' X' Y$ is scalar (1x1), it is equal to its transpose;

$$\begin{aligned} \hat{\beta}' X' Y &= Y' X \hat{\beta} \\ e'e &= Y' Y - 2\hat{\beta}' X' Y + \hat{\beta}' X' X \hat{\beta} \dots \dots \dots (3.31) \end{aligned}$$

Minimizing $e'e$ with respect to the elements in $\hat{\beta}$

$$\frac{\partial \sum e_i^2}{\partial \hat{\beta}} = \frac{\partial (e'e)}{\partial \hat{\beta}} = -2X'Y + 2X'X\hat{\beta}$$

$$\text{Since } \frac{\partial (X'AX)}{\partial \hat{\beta}} = 2AX \text{ and also too } 2X'A$$

Equating the expression to null vector 0, we obtain:

$$-2X'Y + 2X'X\hat{\beta} = 0 \Rightarrow X'X\hat{\beta} = X'Y$$

$$\boxed{\hat{\beta} = (X'X)^{-1}X'Y} \dots\dots\dots (3.32)$$

Hence $\hat{\beta}$ is the vector of required least square estimators, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$.

3.4.2. Statistical Properties of the Parameters (Matrix) Approach

We have seen, in simple linear regression that the OLS estimators ($\hat{\alpha}$ & $\hat{\beta}$) satisfy the small sample property of an estimator i.e. BLUE property. In multiple regression, the OLS estimators also satisfy the BLUE property. Now we proceed to examine the desired properties of the estimators in matrix notations:

1. Linearity

We know that: $\hat{\beta} = (X'X)^{-1}X'Y$

$$\begin{aligned} \text{Let } C &= (X'X)^{-1}X' \\ \Rightarrow \hat{\beta} &= CY \dots\dots\dots (3.33) \end{aligned}$$

Since C is a matrix of fixed variables, equation (3.33) indicates us $\hat{\beta}$ is linear in Y.

2. Unbiased ness

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'Y \\ \hat{\beta} &= (X'X)^{-1}X'(X\beta + U) \\ \hat{\beta} &= \beta + (X'X)^{-1}X'U \dots\dots\dots (3.34) \\ \text{since } [(X'X)^{-1}X'X &= I] \\ E(\hat{\beta}) &= E\{\beta + (X'X)^{-1}X'U\} \\ &= E(\beta) + E[(X'X)^{-1}X'U] \\ &= \beta + E(X'X)^{-1}X'E(U) \\ &= \beta, \text{ since } E(U) = 0 \end{aligned}$$

Thus, least square estimators are unbiased.

3. Minimum variance

Before showing all the OLS estimators are best (possess the minimum variance property), it is important to derive their variance.

We know that, $\text{var}(\hat{\beta}) = E[(\hat{\beta} - \beta)^2] = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)']$

$$E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] =$$

$$\begin{bmatrix} E(\hat{\beta}_1 - \beta_1)^2 & E[(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2)] & \dots & E[(\hat{\beta}_1 - \beta_1)(\hat{\beta}_k - \beta_k)] \\ E[(\hat{\beta}_2 - \beta_2)(\hat{\beta}_1 - \beta_1)] & E(\hat{\beta}_2 - \beta_2)^2 & \dots & E[(\hat{\beta}_2 - \beta_2)(\hat{\beta}_k - \beta_k)] \\ \vdots & \vdots & \ddots & \vdots \\ E[(\hat{\beta}_k - \beta_k)(\hat{\beta}_1 - \beta_1)] & E[(\hat{\beta}_k - \beta_k)(\hat{\beta}_2 - \beta_2)] & \dots & E(\hat{\beta}_k - \beta_k)^2 \end{bmatrix}$$

$$= \begin{bmatrix} \text{var}(\hat{\beta}_1) & \text{cov}(\hat{\beta}_1, \hat{\beta}_2) & \dots & \text{cov}(\hat{\beta}_1, \hat{\beta}_k) \\ \text{cov}(\hat{\beta}_2, \hat{\beta}_1) & \text{var}(\hat{\beta}_2) & \dots & \text{cov}(\hat{\beta}_2, \hat{\beta}_k) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(\hat{\beta}_k, \hat{\beta}_1) & \text{cov}(\hat{\beta}_k, \hat{\beta}_2) & \dots & \text{var}(\hat{\beta}_k) \end{bmatrix}$$

The above matrix is a symmetric matrix containing variances along its main diagonal and covariance of the estimators every where else. This matrix is, therefore, called the *Variance-covariance* matrix of least squares estimators of the regression slopes. Thus,

$$\text{var}(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] \dots \dots \dots (3.35)$$

$$\text{From (3.15)} \quad \hat{\beta} = \beta + (X'X)^{-1}X'U$$

$$\Rightarrow \hat{\beta} - \beta = (X'X)^{-1}X'U \dots \dots \dots (3.36)$$

Substituting (3.17) in (3.16)

$$\text{var}(\hat{\beta}) = E\left\{[(X'X)^{-1}X'U][(X'X)^{-1}X'U]'\right\}$$

$$\text{var}(\hat{\beta}) = E[(X'X)^{-1}X'UU'X(X'X)^{-1}]$$

$$= (X'X)^{-1}X'E(UU')X(X'X)^{-1}$$

$$= (X'X)^{-1}X'\sigma_u^2 I_n X(X'X)^{-1}$$

$$= \sigma_u^2 (X'X)^{-1}X'X(X'X)^{-1}$$

$$\boxed{\text{var}(\hat{\beta}) = \sigma_u^2 (X'X)^{-1}} \dots \dots \dots (3.37)$$

Note: (σ_u^2 being a scalar can be moved in front or behind of a matrix while identity matrix I_n can be suppressed).

Thus we obtain, $\text{var}(\hat{\beta}) = \sigma_u^2 (X'X)^{-1}$

$$\text{Where, } (X'X)^{-1} = \begin{bmatrix} n & \Sigma X_{1n} & \dots & \Sigma X_{kn} \\ \Sigma X_{1n} & \Sigma X_{1n}^2 & \dots & \Sigma X_{1n} X_{kn} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma X_{kn} & \Sigma X_{1n} X_{kn} & \dots & \Sigma X_{kn}^2 \end{bmatrix}$$

We can, therefore, obtain the variance of any estimator say $\hat{\beta}_1$ by taking the i^{th} term from the principal diagonal of $(X'X)^{-1}$ and then multiplying it by σ_u^2 .

Where the X's are in their absolute form. When the x's are in deviation form we can write the multiple regression in matrix form as ;

$$\hat{\beta} = (x'x)^{-1} x'y$$

$$\text{where } \hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} \quad \text{and} \quad (x'x) = \begin{bmatrix} \sum x_1^2 & \Sigma x_1 x_2 & \dots & \Sigma x_1 x_k \\ \Sigma x_2 x_1 & \Sigma x_2^2 & \dots & \Sigma x_2 x_k \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma x_n x_1 & \Sigma x_n x_2 & \dots & \Sigma x_k^2 \end{bmatrix}$$

The above column matrix $\hat{\beta}$ doesn't include the constant term $\hat{\beta}_0$. Under such conditions the variances of slope parameters in deviation form can be written as: $\text{var}(\hat{\beta}) = \sigma_u^2 (x'x)^{-1} \dots \dots \dots (2.38)$

(the proof is the same as (3.37) above). In general we can illustrate the variance of the parameters by taking two explanatory variables.

The multiple regression when written in deviation form that has two explanatory variables is

$$y_1 = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\text{var}(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)']$$

$$\text{In this model; } (\hat{\beta} - \beta) = \begin{bmatrix} (\hat{\beta}_1 - \beta_1) \\ (\hat{\beta}_2 - \beta_2) \end{bmatrix}$$

$$(\hat{\beta} - \beta)' = [(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2)]$$

$$\therefore (\hat{\beta} - \beta)(\hat{\beta} - \beta)' = \begin{bmatrix} (\hat{\beta}_1 - \beta_1) \\ (\hat{\beta}_2 - \beta_2) \end{bmatrix} [(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2)]$$

$$\begin{aligned} \text{and } E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] &= E \begin{bmatrix} (\hat{\beta}_1 - \beta_1)^2 & (\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2) \\ (\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2) & (\hat{\beta}_2 - \beta_2)^2 \end{bmatrix} \\ &= \begin{bmatrix} \text{var}(\hat{\beta}_1) & \text{cov}(\hat{\beta}_1, \hat{\beta}_2) \\ \text{cov}(\hat{\beta}_1, \hat{\beta}_2) & \text{var}(\hat{\beta}_2) \end{bmatrix} \end{aligned}$$

In case of two explanatory variables, x in the deviation form shall be:

$$x = \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ \vdots & \vdots \\ x_{1n} & x_{2n} \end{bmatrix} \text{ and } x' = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \end{bmatrix}$$

$$\therefore \sigma_u^2 (x'x)^{-1} = \sigma_u^2 \begin{bmatrix} \Sigma x_1^2 & \Sigma x_1 x_2 \\ \Sigma x_1 x_2 & \Sigma x_2^2 \end{bmatrix}^{-1}$$

$$\text{Or } \sigma_u^2 (x'x)^{-1} = \frac{\sigma_u^2 \begin{bmatrix} \Sigma x_2^2 & -\Sigma x_1 x_2 \\ -\Sigma x_1 x_2 & \Sigma x_1^2 \end{bmatrix}}{\begin{vmatrix} \Sigma x_1^2 & \Sigma x_1 x_2 \\ \Sigma x_1 x_2 & \Sigma x_2^2 \end{vmatrix}}$$

$$\text{i.e., } \text{var}(\hat{\beta}_1) = \frac{\sigma_u^2 \Sigma x_2^2}{\Sigma x_1^2 \Sigma x_2^2 - (\Sigma x_1 x_2)^2} \dots \dots \dots (3.39)$$

$$\text{and, } \text{var}(\hat{\beta}_2) = \frac{\sigma_u^2 \Sigma x_1^2}{\Sigma x_1^2 \Sigma x_2^2 - (\Sigma x_1 x_2)^2} \dots \dots \dots (3.40)$$

$$\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = \frac{(-)\sigma_u^2 \Sigma x_1 x_2}{\Sigma x_1^2 \Sigma x_2^2 - (\Sigma x_1 x_2)^2} \dots \dots \dots (3.41)$$

The only unknown part in variances and covariance of the estimators is σ_u^2 .

As we have seen in simple regression model $\hat{\sigma}^2 = \left\{ \frac{\sum e_i^2}{n-2} \right\}$. For k-parameters (including the constant parameter) $\hat{\sigma}^2 = \left\{ \frac{\sum e_i^2}{n-k} \right\}$.

In the above model we have three parameters including the constant term and

$$\hat{\sigma}^2 = \left\{ \frac{\sum e_i^2}{n-3} \right\}$$

$$\sum e_i^2 = \sum y_i^2 - \beta_1 \sum x_1 y - \beta_2 \sum x_2 y \dots + \beta_k \sum x_k y \dots \dots \dots (3.42)$$

this is for k explanatory variables. For two explanatory variables

$$\sum e_i^2 = \sum y_i^2 - \beta_1 \sum x_1 y - \beta_2 \sum x_2 y \dots \dots \dots (3.43)$$

This is all about the variance covariance of the parameters. Now it is time to see the minimum variance property.

Minimum variance of $\hat{\beta}$

To show that all the β_i 's in the $\hat{\beta}$ vector are Best Estimators, we have also to prove that the variances obtained in (3.37) are the smallest amongst all other possible linear unbiased estimators. We follow the same procedure as followed in case of single explanatory variable model where, we first assumed an alternative linear unbiased estimator and then it was established that its variance is greater than the estimator of the regression model.

Assume that $\hat{\beta}$ is an alternative unbiased and linear estimator of β . Suppose that

$$\hat{\beta} = [(X'X)^{-1}X' + B]Y$$

Where B is (k x n) matrix of known constants.

$$\therefore \hat{\beta} = [(X'X)^{-1}X' + B][X\beta + U]$$

$$\hat{\beta} = (X'X)^{-1}X'(X\beta + U) + B(X\beta + U)$$

$$\begin{aligned}
 E(\hat{\beta}) &= E[(X'X)^{-1}X'(X\beta + U) + B(X\beta + U)] \\
 &= E[(X'X)^{-1}X'X\beta + (X'X)^{-1}X'U + BX\beta + BU] \\
 &= \beta + BX\beta, \text{ [since } E(U) = 0] \dots \dots \dots (3.44)
 \end{aligned}$$

Since our assumption regarding an alternative $\hat{\beta}$ is that it is to be an unbiased estimator of β , therefore, $E(\hat{\beta})$ should be equal to β ; in other words (βXB) should be a null matrix.

Thus we say, BX should be $=0$ if $(\hat{\beta}) = [(X'X)^{-1}X' + B]Y$ is to be an unbiased estimator. Let us now find variance of this alternative estimator.

$$\begin{aligned}
 \text{var}(\hat{\beta}) &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] \\
 &= E\{[(X'X)^{-1}X' + B]Y - \beta\}[(X'X)^{-1}X' + B]Y - \beta\}' \\
 &= E\{[(X'X)^{-1}X' + B](X\beta + U) - \beta\}[(X'X)^{-1}X' + B](X\beta + U) - \beta\}' \\
 &= E\{(X'X)^{-1}X'X\beta + (X'X)^{-1}X'U + BX\beta + BU - \beta\} \\
 &\quad \{(X'X)^{-1}X'X\beta + (X'X)^{-1}X'U + BX\beta + BU - \beta\}' \\
 &= E\{(X'X)^{-1}X'U + BU\}(X'X)^{-1}X'U + BU\}' \\
 &\quad (\because BX = 0) \\
 &= E\{(X'X)^{-1}X'U + BU\}\{U'X(X'X)^{-1} + U'B'\} \\
 &= E\{(X'X)^{-1}B\}U'U\{X(X'X)^{-1} + U'B'\} \\
 &= [(X'X)^{-1}X' + B]E(UU')\{X(X'X)^{-1} + B'\} \\
 &= \sigma_u^2 I_n [(X'X)^{-1}X' + B]\{X(X'X)^{-1} + B'\} \\
 &= \sigma_u^2 [(X'X)^{-1}X'X(X'X)^{-1} + BX(X'X)^{-1} + (X'X)^{-1}] \\
 &= \sigma_u^2 [(X'X)^{-1}X'X(X'X)^{-1} + BX(X'X)^{-1} + (X'X)^{-1}X'B' + BB'] \\
 &= \sigma_u^2 [(X'X)^{-1} + BB'] (\because BX = 0) \\
 &\boxed{\text{var}(\hat{\beta}) = \sigma_u^2 (X'X)^{-1} + \sigma_u^2 BB'} \dots \dots \dots (3.45)
 \end{aligned}$$

Or, in other words, $\text{var}(\hat{\beta})$ is greater than $\text{var}(\beta)$ by an expression $\sigma_u^2 BB'$ and it proves that $\hat{\beta}$ is the best estimator.

3.4.3. Coefficient of Determination in Matrix Form

The coefficient of determination (R^2) can be derived in matrix form as follows.

We know that $\sum e_i^2 = e'e = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$ since $(X'X)\hat{\beta} = X'Y$ and

$$\sum Y_i^2 = Y'Y$$

$$\therefore e'e = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'Y$$

$$e'e = Y'Y - \hat{\beta}'X'Y \dots\dots\dots(3.46)$$

$$\hat{\beta}'X'Y = e'e - Y'Y \dots\dots\dots(3.47)$$

We know, $y_i = Y_i - \bar{Y}$

$$\therefore \sum y_i^2 = \sum Y_i^2 - \frac{1}{n}(\sum Y_i)^2$$

In matrix notation

$$\sum y_i^2 = Y'Y - \frac{1}{n}(\sum Y_i)^2 \dots\dots\dots(3.48)$$

Equation (3.48) gives the total sum of squares variations in the model.

$$\text{Explained sum of squares} = \sum y_i^2 - \sum e_i^2$$

$$= Y'Y - \frac{1}{n}(\sum y)^2 - e'e$$

$$= \hat{\beta}'X'Y - \frac{1}{n}(\sum Y_i)^2 \dots\dots\dots(3.49)$$

$$\text{Since } R^2 = \frac{\text{Explained sum of squares}}{\text{Total sum of squares}}$$

$$\therefore R^2 = \frac{\hat{\beta}'X'Y - \frac{1}{n}(\sum Y_i)^2}{Y'Y - \frac{1}{n}(\sum Y_i)^2} = \frac{\hat{\beta}'X'Y - n\bar{Y}}{Y'Y - n\bar{Y}^2} \dots\dots\dots(3.50)$$

Dear Students! We hope that from the discussion made so far on multiple regression model, in general, you may make the following summary of results.

(i)	Model:	$Y = X\beta + U$
(ii)	Estimators:	$\hat{\beta} = (X'X)^{-1} X'Y$
(iii)	Statistical properties:	BLUE
(iv)	Variance-covariance:	$\text{var}(\hat{\beta}) = \sigma_u^2 (X'X)^{-1}$
(v)	Estimation of (e'e):	$e'e = Y'Y - \hat{\beta}' X'Y$
(vi)	Coeff. of determination:	$R^2 = \frac{\hat{\beta}' X'Y - \frac{1}{n}(\sum Y_i)^2}{Y'Y - \frac{1}{n}(\sum Y_i)^2} = \frac{\hat{\beta}' X'Y - n\bar{Y}}{Y'Y - n\bar{Y}}$

3.5. Hypothesis Testing in Multiple Regression Model

In multiple regression models we will undertake two tests of significance. One is significance of individual parameters of the model. This test of significance is the same as the tests discussed in simple regression model. The second test is overall significance of the model.

3.5.1. Tests of individual significance

If we invoke the assumption that $U_i \sim N(0, \sigma^2)$, then we can use either the t-test or standard error test to test a hypothesis about any individual partial regression coefficient. To illustrate consider the following example.

$$\text{Let } Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + e_i \dots\dots\dots (3.51)$$

$$\text{A. } H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

$$\text{B. } H_0 : \beta_2 = 0$$

$$H_1 : \beta_2 \neq 0$$

The null hypothesis (A) states that, holding X_2 constant X_1 has no (linear) influence on Y . Similarly hypothesis (B) states that holding X_1 constant, X_2 has no influence on the dependent variable Y_i . To test these null hypothesis we will use the following tests:

- i- **Standard error test:** under this and the following testing methods we test only for $\hat{\beta}_1$. The test for $\hat{\beta}_2$ will be done in the same way.

$$SE(\hat{\beta}_1) = \sqrt{\text{var}(\hat{\beta}_1)} = \sqrt{\frac{\hat{\sigma}^2 \sum x_{2i}^2}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_1 x_2)^2}} \quad ; \text{ where } \hat{\sigma}^2 = \frac{\sum e_i^2}{n-3}$$

- If $SE(\hat{\beta}_1) > \frac{1}{2} \hat{\beta}_1$, we accept the null hypothesis that is, we can conclude that the estimate β_i is not statistically significant.
- If $SE(\hat{\beta}_1) < \frac{1}{2} \hat{\beta}_1$, we reject the null hypothesis that is, we can conclude that the estimate β_i is statistically significant.

Note: The smaller the standard errors, the stronger the evidence that the estimates are statistically reliable.

- ii. **The student's t-test:** We compute the t-ratio for each $\hat{\beta}_i$

$$t^* = \frac{\hat{\beta}_i - \beta}{SE(\hat{\beta}_i)} \sim t_{n-k} \quad , \text{ where } n \text{ is number of observation and } k \text{ is number of}$$

parameters. If we have 3 parameters, the degree of freedom will be $n-3$. So;

$$t^* = \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} \quad ; \text{ with } n-3 \text{ degree of freedom}$$

In our null hypothesis $\beta_2 = 0$, the t^* becomes:

$$t^* = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)}$$

- If $t^* < t$ (tabulated), we accept the null hypothesis, i.e. we can conclude that $\hat{\beta}_2$ is not significant and hence the regressor does not appear to contribute to the explanation of the variations in Y .

- If $t^* > t$ (tabulated), we reject the null hypothesis and we accept the alternative one; $\hat{\beta}_2$ is statistically significant. Thus, the greater the value of t^* the stronger the evidence that β_i is statistically significant.

3.5.2 Test of Overall Significance

Through out the previous section we were concerned with testing the significance of the estimated partial regression coefficients individually, i.e. under the separate hypothesis that each of the true population partial regression coefficient was zero.

In this section we extend this idea to joint test of the relevance of all the included explanatory variables. Now consider the following:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + U_i$$

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$$

$$H_1 : \text{at least one of the } \beta_k \text{ is non-zero}$$

This null hypothesis is a joint hypothesis that $\beta_1, \beta_2, \dots, \beta_k$ are jointly or simultaneously equal to zero. A test of such a hypothesis is called **a test of overall significance** of the observed or estimated regression line, that is, whether Y is linearly related to $X_1, X_2, \dots, X_k$.

Can the joint hypothesis be tested by testing the significance of individual significance of $\hat{\beta}_i$'s as the above? The answer is no, and the reasoning is as follows.

In testing the individual significance of an observed partial regression coefficient, we assumed implicitly that each test of significance was based on different (i.e. independent) sample. Thus, in testing the significance of $\hat{\beta}_2$ under the hypothesis that $\beta_2 = 0$, it was assumed tacitly that the testing was based on different sample

from the one used in testing the significance of $\hat{\beta}_3$ under the null hypothesis that $\beta_3 = 0$. But to test the joint hypothesis of the above, we shall be violating the assumption underlying the test procedure.

“.....testing a series of single (individual) hypothesis is not equivalent to testing those same hypothesis. The institutive reason for this is that in a joint test of several hypotheses any single hypothesis is affected by the information in the other hypothesis.”¹

The test procedure for any set of hypothesis can be based on a ***comparison of the sum of squared errors from the original, the unrestricted multiple regression model to the sum of squared errors from a regression model in which the null hypothesis is assumed to be true.*** When a null hypothesis is assumed to be true, we in effect place conditions or constraints, on the values that the parameters can take, and the sum of squared errors increases. The idea of the test is that if these sum of squared errors are substantially different, then the assumption that the joint null hypothesis is true has significantly reduced the ability of the model to fit the data, and the data do not support the null hypothesis.

If the null hypothesis is true, we expect that the data are compliable with the conditions placed on the parameters. Thus, there would be little change in the sum of squared errors when the null hypothesis is assumed to be true.

Let the Restricted Residual Sum of Square (**RRSS**) be the sum of squared errors in the model obtained by assuming that the null hypothesis is true and **URSS** be the sum of the squared error of the original unrestricted model i.e. unrestricted residual sum of square (URSS). It is always true that $RRSS - URSS \geq 0$.

¹ Gujarati, 3rd ed.pp

Consider $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_k X_k + e_i$.

This model is called unrestricted. The test of joint hypothesis is that:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$$

H_1 : at least one of the β_k is different from zero.

We know that: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_k X_{ki}$

$$Y_i = \hat{Y} + e$$

$$e_i = Y_i - \hat{Y}_i$$

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

This sum of squared error is called unrestricted residual sum of square (URSS).

This is the case when the null hypothesis is not true. If the null hypothesis is assumed to be true, i.e. when all the slope coefficients are zero.

$$Y = \hat{\beta}_0 + e_i$$

$$\hat{\beta}_0 = \frac{\sum Y_i}{n} = \bar{Y} \rightarrow \text{(applying OLS)} \dots \dots \dots (3.52)$$

$$e = Y - \hat{\beta}_0 \quad \text{but} \quad \hat{\beta}_0 = \bar{Y}$$

$$e = Y - \bar{Y}$$

$$\sum e_i^2 = \sum (Y_i - \bar{Y})^2 = \sum y^2 = TSS$$

The sum of squared error when the null hypothesis is assumed to be true is called Restricted Residual Sum of Square (RRSS) and this is equal to the total sum of square (TSS).

$$\text{The ratio: } \frac{RRSS - URSS / K - 1}{URSS / n - K} \sim F_{(k-1, n-k)} \dots \dots \dots (3.53);$$

(has an F-distribution with k-1 and n-k degrees of freedom for the numerator and denominator respectively)

$$RRSS = TSS$$

$$URSS = \sum e_i^2 = \sum y^2 - \hat{\beta}_1 \sum yx_1 - \hat{\beta}_2 \sum yx_2 + \dots \dots \dots \hat{\beta}_k \sum yx_k = RSS$$

$$F = \frac{(TSS - RSS) / k - 1}{RSS / n - k}$$

$$F = \frac{ESS / k - 1}{RSS / n - k} \dots\dots\dots (3.54)$$

If we divide the above numerator and denominator by $\Sigma y^2 = TSS$ then:

$$F = \frac{\frac{ESS}{TSS} / k - 1}{\frac{RSS}{TSS} / n - k}$$

$$F = \frac{R^2 / k - 1}{1 - R^2 / n - k} \dots\dots\dots (3.55)$$

This implies the computed value of F can be calculated either as a ratio of ESS & TSS or R^2 & $1-R^2$. If the null hypothesis is not true, then the difference between RRSS and URSS (TSS & RSS) becomes large, implying that the constraints placed on the model by the null hypothesis have large effect on the ability of the model to fit the data, and the value of F tends to be large. Thus, we reject the null hypothesis if the F test static becomes too large. This value is compared with the critical value of F which leaves the probability of α in the upper tail of the F-distribution with k-1 and n-k degree of freedom.

If the computed value of F is greater than the critical value of F (k-1, n-k), then the parameters of the model are jointly significant or the dependent variable Y is linearly related to the independent variables included in the model.

Application of Multiple Regression.

In order to help you understand the working of matrix algebra in the estimation of the regression coefficient, variance of the coefficients and testing of the parameters and the model, consider the following numerical example.

Example 1. Consider the data given in Table 2.1 below to fit a linear function:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + U$$

Table: 2.1. Numerical example for the computation of the OLS estimators.

n	Y	X_1	X_2	X_3	y_i	x_1	x_2	x_3	y_i^2	x_1x_2	x_2x_3	x_1x_3	x_1^2	x_2^2	x_3^2	x_1y_i	x_2y_i	x_3y_i
1	49	35	53	200	-3	-7	-9	0	9	63	0	0	49	81	0	21	27	0
2	40	35	53	212	-12	-7	-9	12	144	63	-108	-84	49	81	144	84	108	-144
3	41	38	50	211	-11	-4	-12	11	121	48	-132	-44	16	144	121	44	132	-121
4	46	40	64	212	-6	-2	2	12	36	-4	24	-24	4	4	144	12	-12	-72
5	52	40	70	203	0	-2	8	3	0	-16	24	-6	4	64	9	0	0	0
6	59	42	68	194	7	0	6	-6	49	0	-36	0	0	36	36	0	42	-42
7	53	44	59	194	1	2	-3	-6	1	-6	18	-12	4	9	36	2	-3	-06
8	61	46	73	188	9	4	11	-12	81	44	-132	-48	16	121	144	36	99	-108
9	55	50	59	196	3	8	-3	-4	9	-24	12	-32	64	9	16	24	-9	-12
1	64	50	71	190	12	8	9	-10	144	72	-90	-80	64	81	100	96	108	-120
0																		
	520	420	620	2000														
					$\sum y_i = 594$	$\sum x_1 = 0$	$\sum x_2 = 0$	$\sum x_3 = 0$	$\sum y_i^2 = 594$	$\sum x_1x_2 = 240$	$\sum x_2x_3 = -420$	$\sum x_1x_3 = -330$	$\sum x_1^2 = 270$	$\sum x_2^2 = 630$	$\sum x_3^2 = 750$	$\sum x_1y_i = 319$	$\sum x_2y_i = 492$	$\sum x_3y_i = -625$

From the table, the means of the variables are computed and given below:

$$\bar{Y} = 52; \quad \bar{X}_1 = 42; \quad \bar{X}_2 = 62; \quad \bar{X}_3 = 200$$

Based on the above table and model answer the following question.

- i. Estimate the parameter estimators using the matrix approach
- ii. Compute the variance of the parameters.
- iii. Compute the coefficient of determination (R^2)
- iv. Report the regression result.

Solution:

In the matrix notation: $\hat{\beta} = (x'x)^{-1}x'y$; (when we use the data in deviation form),

$$\text{Where, } \hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix}, x = \begin{bmatrix} x_{11} & x_{21} & x_{31} \\ x_{12} & x_{22} & x_{32} \\ \vdots & \vdots & \vdots \\ x_{1n} & x_{2n} & x_{3n} \end{bmatrix}; \text{ so that}$$

$$(x'x) = \begin{bmatrix} \sum x_1^2 & \sum x_1 x_2 & \sum x_1 x_3 \\ \sum x_1 x_2 & \sum x_2^2 & \sum x_2 x_3 \\ \sum x_1 x_3 & \sum x_2 x_3 & \sum x_3^2 \end{bmatrix} \text{ and } x'y = \begin{bmatrix} \sum x_1 y \\ \sum x_2 y \\ \sum x_3 y \end{bmatrix}$$

(i) Substituting the relevant quantities from table 2.1 we have;

$$(x'x) = \begin{bmatrix} 270 & 240 & -330 \\ 240 & 630 & -420 \\ -330 & -420 & 750 \end{bmatrix} \text{ and } x'y = \begin{bmatrix} 319 \\ 492 \\ -625 \end{bmatrix}$$

Note: the calculations may be made easier by taking 30 as common factor from all the elements of matrix $(x'x)$. This will not affect the final results.

$$|x'x| = \begin{vmatrix} 270 & 240 & -330 \\ 240 & 630 & -420 \\ -330 & -420 & 750 \end{vmatrix} = 4716000$$

$$(x'x)^{-1} = \begin{bmatrix} 0.0085 & -0.0012 & 0.0031 \\ -0.0012 & 0.0027 & 0.0009 \\ 0.0031 & 0.0009 & 0.0032 \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = (x'x)^{-1} x'y = \begin{bmatrix} 0.0085 & -0.0012 & 0.0031 \\ -0.0012 & 0.0027 & 0.0009 \\ 0.0031 & 0.0009 & 0.0032 \end{bmatrix} \begin{bmatrix} 319 \\ 492 \\ -625 \end{bmatrix} = \begin{bmatrix} 0.2063 \\ 0.3309 \\ -0.5572 \end{bmatrix}$$

And

$$\begin{aligned} \alpha &= \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3 \\ &= 52 - (0.2063)(42) - (0.3309)(62) - (-0.5572)(200) \\ &= 52 - 8.6633 - 20.5139 + 111.4562 = 134.2789 \end{aligned}$$

(ii) The elements in the principal diagonal of $(x'x)^{-1}$ when multiplied σ_u^2 give the variances of the regression parameters, i.e.,

$$\left. \begin{aligned} \text{var}(\hat{\beta}_1) &= \sigma_u^2 (0.0085) \\ \text{var}(\hat{\beta}_2) &= \sigma_u^2 (0.0027) \\ \text{var}(\hat{\beta}_3) &= \sigma_u^2 (0.0032) \end{aligned} \right\} \hat{\sigma}_u^2 = \frac{\sum e_i^2}{n-k} = \frac{17.11}{6} = 2.851$$

$$\text{var}(\hat{\beta}_1) = 0.0243, \quad SE(\hat{\beta}_1) = 0.1560$$

$$\text{var}(\hat{\beta}_2) = 0.0077, \quad SE(\hat{\beta}_2) = 0.0877$$

$$\text{var}(\hat{\beta}_3) = 0.0093, \quad SE(\hat{\beta}_3) = 0.0962$$

$$(iii) \quad R^2 = \frac{\hat{\beta}' X' Y - \frac{1}{n} (\sum Y_i)^2}{Y' Y - \frac{1}{n} (\sum Y_i)^2} = \frac{\hat{\beta}_1 \sum x_1 y + \hat{\beta}_2 \sum x_2 y + \hat{\beta}_3 \sum x_3 y}{\sum y_i^2} = \frac{575.98}{594} = 0.97$$

(iv) The estimated relation may be put in the following form:

$$\hat{Y} = 134.28 + 0.2063X_1 + 0.3309X_2 - 0.5572X_3$$

$SE(\hat{\beta}_i)$	(0.1560)	(0.0877)	(0.0962)	$R^2 = 0.97$
t^*	(1.3221)	(3.7719)	(5.7949)	

The variables X_1, X_2 and X_3 explain 97 percent of total variations.

We can test the significance of individual parameters using the student's t-test.

The computed value of 't' is given above as t^* . these values indicates us only $\hat{\beta}_1$ is insignificant.

Example 2. The following matrix gives the variances and covariance of the of the three variables:

$$\begin{matrix} & y & x_1 & x_2 \\ y & \begin{bmatrix} 7.59 & 3.12 & 26.99 \end{bmatrix} \\ x_1 & \begin{bmatrix} - & 29.16 & 30.80 \end{bmatrix} \\ x_2 & \begin{bmatrix} - & - & 133.00 \end{bmatrix} \end{matrix}$$

The first row and the first column of the above matrix shows $\sum y^2$ and the first row and the second column shows $\sum yx_{1i}$ and so on.

Consider the following model

$$Y_1 = AY_2^{\beta_1} Y_3^{\beta_2} e^{v_i}$$

Where; Y_1 is food consumption per capita

Y_2 is food price

Y_3 is disposable income per capita

And $Y = \ln Y_1$, $X_1 = \ln Y_2$ and $X_2 = \ln Y_3$

$$y = Y - \bar{Y}, x_1 = X - \bar{X}, \text{ and } x_2 = X - \bar{X}$$

Using the values in above matrix answer the following questions.

- Estimate β_1 and β_2
- Compute variance of $\hat{\beta}_1$ and $\hat{\beta}_2$
- Compute coefficient of determination
- Report the regression result.

Solution: It is difficult to estimate the above model as it is, to estimate the above model easily let's take the natural log of the above model;

$$\ln Y_1 = \ln A + \beta_1 \ln Y_2 + \beta_2 \ln Y_3 + V_i$$

And let: $\beta_0 = \ln A$, $Y = \ln Y_1$, $X_1 = \ln Y_2$ and $X_2 = \ln Y_3$ the above model becomes: $Y = \beta_0 + \beta X_1 + \beta X_2 + V_i$

The above matrix is based on the transformed model. Using values in the matrix we can now estimate the parameters of the original model.

We know that $\hat{\beta} = (x'x)^{-1}x'y$

In the present question:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \text{ and } x = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{1n} & x_{nn} \end{bmatrix}$$

$$\therefore x'x = \begin{bmatrix} \Sigma x_1^2 & \Sigma x_1 x_2 \\ \Sigma x_1 x_2 & \Sigma x_2^2 \end{bmatrix} \text{ and } x'y = \begin{bmatrix} \Sigma x_1 y \\ \Sigma x_2 y \end{bmatrix}$$

Substituting the relevant quantities from the given variance-covariance matrix, we obtain:

$$x'x = \begin{bmatrix} 29.16 & 30.80 \\ 30.80 & 133.00 \end{bmatrix} \text{ and } x'y = \begin{bmatrix} 3.12 \\ 26.99 \end{bmatrix}$$

$$|x'x| = \begin{vmatrix} 29.16 & 30.80 \\ 30.80 & 133.00 \end{vmatrix} = 2929.64$$

$$\therefore (x'x)^{-1} = \frac{1}{2929.64} = \begin{bmatrix} 133.00 & -30.80 \\ -30.80 & 29.16 \end{bmatrix} \approx \begin{bmatrix} 0.0454 & -0.0105 \\ -0.0105 & 0.0099 \end{bmatrix}$$

$$(a) \hat{\beta} = (x'x)^{-1}x'y = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$$

$$\begin{bmatrix} 0.0454 & -0.0105 \\ -0.0105 & 0.0099 \end{bmatrix} \begin{bmatrix} 3.12 \\ 26.99 \end{bmatrix} \approx \begin{bmatrix} -0.1421 \\ 0.2358 \end{bmatrix}$$

(b). The element in the principal diagonal of $(x'x)^{-1}$ when multiplied by σ_u^2 give the variances of the $\hat{\alpha}$ and $\hat{\beta}$.

$$(c). \quad R^2 = \frac{\hat{\alpha}\Sigma x_2 y + \hat{\beta}\Sigma x_3 y}{\Sigma y_i^2}$$

$$= \frac{-(0.1421)(3.12) + (0.2358)(26.99)}{7.59}$$

$$\therefore R^2 = 0.78; \Sigma e_i^2 = (1 - R^2)(\Sigma y_i^2) \approx 1.6680$$

$$\therefore \hat{\sigma}_u^2 = \frac{1.6680}{17} = 0.0981$$

$$\text{var}(\hat{\alpha}) = (0.0981)(0.0454) \approx 0.0045, \therefore (\hat{\alpha})SE = 0.0667$$

$$\text{var}(\hat{\beta}) = (0.0981)(0.0099) \approx 0.0009, \therefore (\hat{\beta})SE = 0.0312$$

(d). The results may be put in the following form:

$$\begin{array}{lll} \hat{Y}_1 = AY_1^{(-0.1421)} Y_3^{(0.2358)} & & \\ SE & (0.0667)(0.0312) & R^2 = 0.78 \\ t^* & (-2.13) (7.55) & \end{array}$$

The (constant) food price elasticity is negative but income elasticity is positive. Also income elasticity is highly significant. About 78 percent of the variations in the consumption of food are explained by its price and income of the consumer.

Example 3:

Consider the model: $Y = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i$

On the basis of the information given below answer the following question

$$\begin{array}{lll} \Sigma X_1^2 = 3200 & \Sigma X_1 X_2 = 4300 & \Sigma X_2 = 400 \\ \Sigma X_2^2 = 7300 & \Sigma X_1 Y = 8400 & \Sigma X_2 Y = 13500 \\ \Sigma Y = 800 & \Sigma X_1 = 250 & n = 25 \\ \Sigma Y_i^2 = 28,000 & & \end{array}$$

- Find the OLS estimate of the slope coefficient
- Compute variance of $\hat{\beta}_2$
- Test the significance of β_2 slope parameter at 5% level of significance
- Compute R^2 and $\bar{R}^2$ and interpret the result
- Test the overall significance of the model

Solution:

- Since the above model is a two explanatory variable model, we can estimate

$\hat{\beta}_1$ and $\hat{\beta}_2$ using the formula in equation (3.21) and (3.22) i.e.

$$\hat{\beta}_1 = \frac{\Sigma x_2 y \Sigma x_2^2 - \Sigma x_2 y \Sigma x_1 x_2}{\Sigma x_1^2 \Sigma x_2^2 - (\Sigma x_1 x_2)^2}$$

$$\hat{\beta}_2 = \frac{\Sigma x_2 y \Sigma x_1^2 - \Sigma x_1 y \Sigma x_1 x_2}{\Sigma x_1^2 \Sigma x_2^2 - (\Sigma x_1 x_2)^2}$$

Since the x's and y's in the above formula are in deviation form we have to find the corresponding deviation forms of the above given values.

We know that:

$$\begin{aligned}\Sigma x_1 x_2 &= \Sigma X_1 X_2 - n \bar{X}_1 \bar{X}_2 \\ &= 4300 - (25)(10)(16) \\ &= \underline{\underline{300}}\end{aligned}$$

$$\begin{aligned}\Sigma x_1 y &= \Sigma X_1 Y - n \bar{X}_1 \bar{Y} \\ &= 8400 - 25(10)(32) \\ &= 400\end{aligned}$$

$$\begin{aligned}\Sigma x_2 y &= \Sigma X_2 Y - n \bar{X}_2 \bar{Y} \\ &= 13500 - 25(16)(32) \\ &= 700\end{aligned}$$

$$\begin{aligned}\Sigma x_1^2 &= \Sigma X_1^2 - n \bar{X}_1^2 \\ &= 3200 - 25(10)^2 \\ &= 700\end{aligned}$$

$$\begin{aligned}\Sigma x_2^2 &= \Sigma X_2^2 - n \bar{X}_2^2 \\ &= 7300 - 25(16)^2 \\ &= 900\end{aligned}$$

Now we can compute the parameters.

$$\begin{aligned}\hat{\beta}_1 &= \frac{\Sigma x_2 y \Sigma x_2^2 - \Sigma x_2 y \Sigma x_1 x_2}{\Sigma x_1^2 \Sigma x_2^2 - (\Sigma x_1 x_2)^2} \\ &= \frac{(400)(900) - (700)(300)}{(900)(700) - (300)^2} \\ &= \underline{\underline{0.278}}\end{aligned}$$

$$\hat{\beta}_2 = \frac{\Sigma x_2 y \Sigma x_1^2 - \Sigma x_1 y \Sigma x_1 x_2}{\Sigma x_1^2 \Sigma x_2^2 - (\Sigma x_1 x_2)^2}$$

$$\begin{aligned} &= \frac{(700)(700) - (400)(300)}{(900)(700) - (300)^2} \\ &= \underline{\underline{0.685}} \end{aligned}$$

The intercept parameter can be computed using the following formula.

$$\begin{aligned} \hat{\alpha} &= \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 \\ &= 32 - (0.278)(10) - (0.685)(16) \\ &= \underline{\underline{18.26}} \end{aligned}$$

$$\begin{aligned} b. \text{var}(\hat{\beta}_1) &= \frac{\hat{\sigma}^2 \Sigma x_2^2}{\Sigma x_1^2 \Sigma x_2^2 - (\Sigma x_1 x_2)^2} \\ \Rightarrow \hat{\sigma}^2 &= \frac{\Sigma e_i^2}{n - k} \text{ Where } k \text{ is the number of parameter} \end{aligned}$$

In our case $k=3$

$$\begin{aligned} \Rightarrow \hat{\sigma}^2 &= \frac{\Sigma e_i^2}{n - 3} \\ \Sigma e_1^2 &= \Sigma y^2 - \hat{\beta}_1 \Sigma x_1 y - \hat{\beta}_2 \Sigma x_2 y \\ &= 2400 - 0.278(400) - (0.685)(700) \\ &= 1809.3 \end{aligned}$$

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\Sigma e_i^2}{n - 3} \\ &= \frac{1809.3}{25 - 3} \\ &= 82.24 \end{aligned}$$

$$\Rightarrow \text{var}(\hat{\beta}_1) = \frac{(82.24)(900)}{540,000} = 0.137$$

$$SE(\hat{\beta}) = \sqrt{\text{var}(\hat{\beta}_1)} = \sqrt{0.137} = 0.370$$

$$\begin{aligned} \text{var}(\hat{\beta}_2) &= \frac{\hat{\sigma}^2 \Sigma x_1^2}{\Sigma x_1^2 \Sigma x_2^2 - (\Sigma x_1 x_2)^2} \\ &= \frac{(82.24)(700)}{540,000} = 0.1067 \end{aligned}$$

$$SE(\hat{\beta}) = \sqrt{\text{var}(\hat{\beta}_2)} = \sqrt{0.1067} = 0.327$$

c. $\hat{\beta}_1$ can be tested using students t-test

This is done by comparing the computed value of t and critical value of t which is obtained from the table at $\alpha/2$ level of significance and n-k degree of freedom.

$$\text{Hence; } \frac{t^*}{SE(\hat{\beta}_1)} = \frac{0.278}{0.370} = 0.751$$

The critical value of t from the t-table at $\alpha/2 = 0.05/2 = 0.025$ level of significance and 22 degree of freedom is 2.074.

$$t_c = 2.074$$

$$t^* = 0.755$$

$$\Rightarrow t^* < t_c$$

The decision rule if $t^* < t_c$ is to reject the alternative hypothesis that says β is different from zero and to accept the null hypothesis that says β is equal to zero.

The conclusion is $\hat{\beta}_1$ is statistically insignificant or the sample we use to estimate $\hat{\beta}_1$ is drawn from the population of Y & X_1 in which there is no relationship between Y and X_1 (i.e. $\beta_1 = 0$).

d. R^2 can be easily using the following equation

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

We know that $RSS = \sum e_i^2$

and $TSS = \sum y^2$ and $ESS = \sum \hat{y}^2 = \hat{\beta}_1 \sum x_1 y + \hat{\beta}_2 \sum x_2 y + \dots + \hat{\beta}_k \sum x_k y$

For two explanatory variable model:

$$\begin{aligned} R^2 &= 1 - \frac{RSS}{TSS} = 1 - \frac{10809.3}{2400} \\ &= 0.24 \end{aligned}$$

$\Rightarrow 24\%$ of the total variation in Y is explained by the regression line $\hat{Y} = 18.26 + 0.278X_1 + 0.685X_2$ or by the explanatory variables (X_1 and X_2).

$$\begin{aligned}\text{Adjusted } R^2 &= 1 - \frac{\Sigma e_i^2 / n - k}{\Sigma y^2 / n - 1} = 1 - \frac{(1 - R^2)(n - 1)}{n - k} \\ &= 1 - \frac{(1 - 0.24)(24)}{22} \\ &= 0.178\end{aligned}$$

e. Let's set first the joint hypothesis as

$$H_0 : \beta_1 = \beta_2 = 0$$

against H_1 : at least one of the slope parameter is different from zero.

The joint test hypothesis is testing using the F-test given below.

$$\begin{aligned}F^*_{[(k-1),(n-k)]} &= \frac{ESS / k - 1}{RSS / n - k} \\ &= \frac{R^2 / k - 1}{1 - R^2 / n - k}\end{aligned}$$

From (d) $R^2 = 0.24$ and $k = 3$

$F^*_{(2,22)} = 3.4736$ this is the computed value of F. Let's compare this with the critical value F at 5% level of significance and (3,.23) numerator and denominator respectively. F (2,22) at 5%level of significance = 3.44.

$$F^*_{(2,22)} = 3.47$$

$$F_c(2,22)=3.44$$

$\Rightarrow F^* > F_c$, the decision rule is to reject H_0 and accept H_1 . We can say that the model is significant i.e. the dependent variable is, at least, linearly related to one of the explanatory variables.

Sample exam questions :

Mid-exam (May, 2004/2005)

- **Instructions:**

Read the following instructions carefully.

- Make sure that your exam paper contains 4 pages
- The exam has four parts. Attempt
 - ✓ All questions of part one
 - ✓ Only two questions from part two
 - ✓ One question from part three
 - ✓ And the question in part four.
- Maximum weight of the exam is 40%

Part One: Attempt all of the following questions (15pts).

1. Discuss briefly the goals of econometrics.
2. Researcher is using data for a sample of 10 observations to estimate the relation between consumption expenditure and income. Preliminary analysis of the sample data produces the following data.

$$\sum xy = 700, \sum x^2 = 1000, \sum X = 100$$
$$\sum Y = 200$$

Where $x = X_i - \bar{X}_i$ and $y = Y_i - \bar{Y}$

- a. Use the above information to compute **OLS** estimates of the intercept and slope coefficients and interpret the result
 - b. Calculate the variance of the slope parameter
 - c. Compute the value **R²** (coefficient of determination) and interpret the result
 - d. Compute 95% confidence interval for the slope parameter
 - e. Test the significance of the slope parameter at 5% level of confidence using t-test
3. If the model $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i$ is to be estimated from a sample of 20 observation using the semi- processed data given in matrix in **deviation** form.

$$(x'x)^{-1} = \begin{pmatrix} 0.5 & -0.08 \\ -0.08 & 0.6 \end{pmatrix}$$

$$x'y = \begin{pmatrix} 100 \\ 250 \end{pmatrix} \quad \bar{X}_1 = 10, \bar{X}_2 = 25 \text{ and } \bar{Y} = 30$$

Obtain the OLS estimate of the above parameters.

4. Linearity is one assumption of classicalist in simple regression analysis. Identify which of the following satisfies this assumption. Discuss why?

- a. $\ln Y = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i$
- b. $Y = \alpha + (1/\beta_1) X_{1i} + \beta_2 X_{2i} + U_i$
- c. $Y = \alpha + \beta_1 X_{1i}^2 + \beta_2 X_{2i} + U_i$
- d. $\ln Y = \alpha + \beta_1 \ln X_{1i} + \beta_2 \ln X_{2i} + U_i$
- e. $Y_i = \alpha + \beta^2 X_i + U_i$

Part Two: Attempt any two of the following questions.(10pts). 1. Consider the model

$$Y_i = \alpha + \beta X_i + U_i$$

Show that the OLS estimate of α is unbiased.

2. Suppose σ^2 is the population variance of the error term and $\hat{\sigma}^2$ is estimator of σ^2 .

Show that $\hat{\sigma}^2$ of **maximum likely hood** is the biased estimator of the true σ^2 for the model $Y_i = \alpha + \beta X_i + U_i$.

3. In the model

$$Y_i = \alpha + \beta X_i + U_i$$

Show that $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$ possesses minimum variance.

4. using the assumptions of simple regression model show that

a. $Y \sim N(\alpha + \beta X_i, \sigma^2)$

b. $\text{Cov}(\hat{\alpha}, \hat{\beta}) = -\bar{X} \text{Var}(\hat{\beta})$

For the model $Y_i = \alpha + \beta X_i + U_i$

Part Three: Attempt any one the Following(10 pts.)

1. The model $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + U_i$ is to be estimated from a sample of 20 observations. Using the information below obtain the OLS estimate of the parameters of the above model.

$$(x'x)^{-1} = \begin{pmatrix} 0.1 & -0.12 & -0.03 \\ -0.12 & 0.04 & 0.02 \\ -0.03 & 0.02 & 0.08 \end{pmatrix} \quad X'Y = \begin{pmatrix} 10,000 \\ 20,300 \\ 10,100 \\ 30,200 \end{pmatrix}$$

$$\sum X_1 = 400, \sum X_2 = 200, \text{ and } \sum X_3 = 600$$

Where $x = X_i - \bar{X}_i$ and $y = Y_i - \bar{Y}$

2. In a study of 100 firms, the total cost(C) was assumed to be dependent on the rate of out put (X_1) and the rate of absenteeism (X_2). The means were: $\bar{C} = 6$, $\bar{X}_1 = 3$ and $\bar{X}_2 = 4$. The matrix showing sums of squares and cross products adjusted for means is

$$\begin{matrix} & c & x_1 & x_2 \\ c & 100 & 50 & 40 \\ x_1 & 50 & 50 & -70 \\ x_2 & 40 & -70 & 900 \end{matrix} \quad \text{where, } x_i = X_i - \bar{X}_i \quad \text{and } c = C_i - \bar{C}$$

Estimate the linear relation ship between C and the other two variables. (10points)

3. Consider the linear regression model

$$Y_i = \alpha + \beta_1 X_i + U_i$$

Suppose that there are no data on X_i but we have data on $Z_i = a_0 + a_1 X_i$, whrer a_0 and a_1 are arbitrary known constants. Using data of variable Z_i we can estimate

$$Y_i = c_0 + c_1 Z_i + U_i$$

Show how from the estimates of $\hat{c}_0$ and $\hat{c}_1$ you can obtain the estimates of the original model.