



BONGA UNIVERSITY
COLLEGE OF BUSINESS AND ECONOMICS
DEPARTMENT OF ECONOMIC

ECONOMETRICS I
Prepared by Firehun Jemal
Compiled by Fikadu Abera

March 2023

Bonga, Ethiopia

Table of Contents

CHAPTER ONE	3
Definition and scope of econometrics	3
1.1 WHAT IS ECONOMETRICS?	3
1.2 Econometrics vs. mathematical economics	4
1.3 Econometrics vs. statistics	5
1.4 Economic models vs. econometric models	5
1.5 Methodology of econometrics	6
1.6 Desirable properties of an econometric model	9
1.7 Goals of Econometrics	9
CHAPTER TWO	10
THE CLASSICAL REGRESSION ANALYSIS.....	10
2.1. Stochastic and Non-stochastic Relationships.....	10
2.2. Simple Linear Regression model.	11
2.2.1 Assumptions of the Classical Linear Stochastic Regression Model.	11
2.2.2 Methods of estimation.....	13
2.3 Tests of the ‘Goodness of Fit’ With R^2	18
2.4 Testing the Significance Of oLS Parameters	18
CHAPTER THREE.....	22
THE CLASSICAL REGRESSION ANALYSIS.....	22
3.1 Introduction.....	22
3.2 Assumptions of Multiple Regression Model	23
3.3 A Model With Two Explanatory Variables	23
3.3.1 Estimation of parameters of two-explanatory variables model.....	23
3.3.2 The coefficient of determination (R^2):two explanatory variables case	24
3.4.2. Statistical Properties of the Parameters.....	25
3.5. Hypothesis Testing in Multiple Regression Model.....	26
3.5.1. Tests of individual significance	26
3.5.2 Computing p-values for t-tests.....	31
3.5.3 Confidence intervals (CI) and hypothesis testing	32
3.5.4 Test of Overall Significance	34
CHAPTER FOUR:	39
VIOLATIONS OF CLASSICAL ASSUMPTION.....	39
4.1 The Assumption of Zero Expected Disturbances	39

4.2 The Nature of Heteroscedasticity	40
4.3 The Nature of Autocorrelation.....	41
4.4 Multicollinearity	43
4.5 Application examples on violation of CLR	45

CHAPTER ONE

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 carry out such an evaluation of economic theories in empirical terms is econometrics.

1.1 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.

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.

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.

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.

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 the 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 biasedness, 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 behavior 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.

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. A simple linear regression model, i.e. a relationship between two variables related in a linear form. i.e. *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.

The derivation of the observation from non-stochastic 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'.

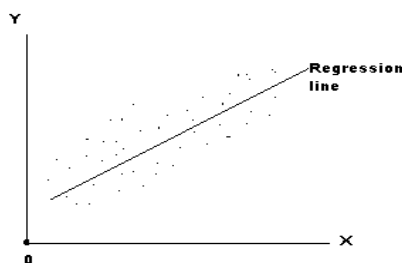


Fig 2.b the scatter diagram and the regression line

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

1. *The model is linear in parameters.*

The classical 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.

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.

3. *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. Mathematically, $E(U_i) = 0$

4. *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. This constant variance is called **homoscedasticity** assumption and the constant variance itself is called homoscedastic variance.

5. *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.

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

6. *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.

7. *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 .

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

This means there is no correlation between the random variable and the explanatory variable.

9. *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.

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:

$$e_i = Y_i - (\hat{\alpha} + \hat{\beta}X_i) \dots\dots\dots(2.3)$$

$$\sum e_i^2 = \sum (Y_i - \hat{\alpha} - \hat{\beta}X_i)^2 \dots\dots\dots(2.4)$$

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.

$$1. \frac{\partial \sum e_i^2}{\partial \hat{\alpha}} = -2 \sum (Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0 \dots\dots\dots(2.5)$$

Rearranging this expression we will get: $\boxed{\sum Y_i = n\alpha + \hat{\beta}\sum X_i} \dots\dots(2.6)$

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

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \dots\dots\dots(2.7)$$

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

Note: $e = Y_i - \hat{\alpha} - \hat{\beta}X_i$. Hence it is possible to rewrite as $-2 \sum e_i = 0$ and $-2 \sum X_i e_i = 0$. It follows that;

$$\sum e_i = 0 \text{ and } \sum X_i e_i = 0 \dots\dots\dots(2.9)$$

$$\boxed{\sum Y_i X_i = \hat{\alpha} \sum X_i + \hat{\beta} \sum X_i^2} \dots\dots\dots(2.10)$$

Equations are called the **Normal Equations**. After certain step we get:

$$\begin{aligned} \sum Y_i X_i &= \sum X_i (\bar{Y} - \hat{\beta}\bar{X}) + \hat{\beta} \sum X_i^2 \\ &= \bar{Y} \sum X_i - \hat{\beta} \bar{X} \sum X_i + \hat{\beta} \sum X_i^2 \end{aligned}$$

$$\begin{aligned}\sum Y_i X_i - \bar{Y} \sum X_i &= \hat{\beta} (\sum X_i^2 - \bar{X} \sum X_i) \\ \sum XY - n\bar{X}\bar{Y} &= \hat{\beta} (\sum X_i^2 - n\bar{X}^2) \\ \boxed{\hat{\beta} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2}} &\dots\dots\dots(2.11)\end{aligned}$$

Equation (2.11) can be rewritten in somewhat different way and arrived final steps are :-

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

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

$$\boxed{\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}} \dots\dots\dots (2.12)$$

The expression in (2.12) 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.

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

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

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

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

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

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

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

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

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

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

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 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). Sometimes the theorem referred as the BLUE theorem i.e. Best, Linear, and Unbiased Estimator. An estimator is called BLUE if:

- a. **Linear:** a linear function of the 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

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:

$$\hat{\sigma}_u^2 = \frac{\sum e_i^2}{n-2} \dots\dots\dots 2.14$$

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:

- 1) 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).
- 2) The standard error tests of the estimators. This test is used for judging the statistical reliability of the estimates of the regression coefficients.

2.3 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

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 .

2.4 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.

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 relationship between Y and X is in fact $Y = \alpha + (0)x = \alpha$, i.e. there is no relationship between X and Y.

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:

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 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.

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 customarily 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 $\frac{\alpha}{2}$ at n-2 degree of freedom.

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

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 .

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.

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 in multiple 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: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U_i \dots\dots\dots(3.3)$

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

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

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:

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

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

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

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

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:

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

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_t . 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_t 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 loses 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.

3.4.2. Statistical Properties of the Parameters

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 regressions, the OLS estimators also satisfy the BLUE property. Now we proceed to examine the desired properties of the estimators are:

1. Linearity

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

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

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

2. Unbiased ness

$$\hat{\beta} = (X'X)^{-1}X'Y \\ \hat{\beta} = (X'X)^{-1}X'(X\beta + U) \\ \hat{\beta} = \beta + (X'X)^{-1}X'U \dots\dots\dots (3.34) \\ = \beta, \text{ since } E(U) = 0$$

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.

3.5. Hypothesis Testing in Multiple Regression Model

In multiple regression models we will undertake different 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;

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

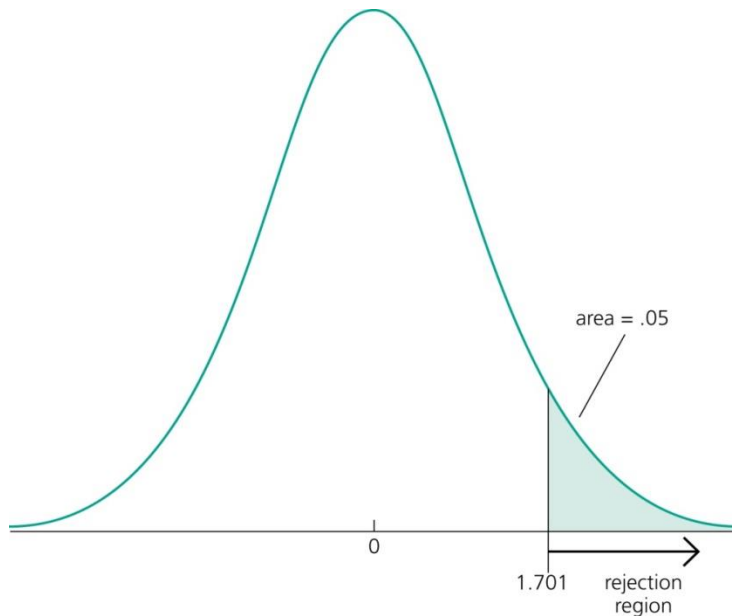
In our null hypothesis $\beta_2 = 0$, the t^* becomes:

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

- 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.

Application example

a) Testing against one-sided alternatives (greater than zero)



Reject the null hypothesis in favour of the alternative hypothesis if the estimated coefficient is “**too large**” (i.e. larger than a critical value).

Construct the critical value so that, if the null hypothesis is true, it is rejected in, for example, 5% of the cases.

In the given example, this is the point of the t-distribution with 28 degrees of freedom that is exceeded in 5% of the cases.

! Reject if t-statistic greater than 1.701

Example: Wage equation

Test whether, after controlling for education and tenure, higher work experience leads to higher hourly wages

$$\widehat{\log(wage)} = .284 + .092 \text{ educ} + .0041 \text{ exper} + .022 \text{ tenure}$$

$$(.104) \quad (.007) \quad (.0017) \quad (.003)$$

Where, standard errors are in parenthesis. $n = 526$, $R^2 = .316$

$H_0 : \beta_{\text{exper}} = 0$ Against $H_1 : \beta_{\text{exper}} > 0$

One would either expect a positive effect of experience on hourly wage or no effect at all.

$$t_{\text{exper}} = .0041 / .0017 \approx 2.41$$

$$df = n - k - 1 = 526 - 3 - 1 = 522 \quad \text{is the Degrees of freedom}$$

Here the standard normal approximation applies. Critical values for the 5% and the 1% significance level (these are conventional significance levels).

$$t_c = t_{\frac{\alpha}{2}, df} = t_{\frac{0.05}{2}, 526-3} = t_{0.025, 523} = 1.645$$

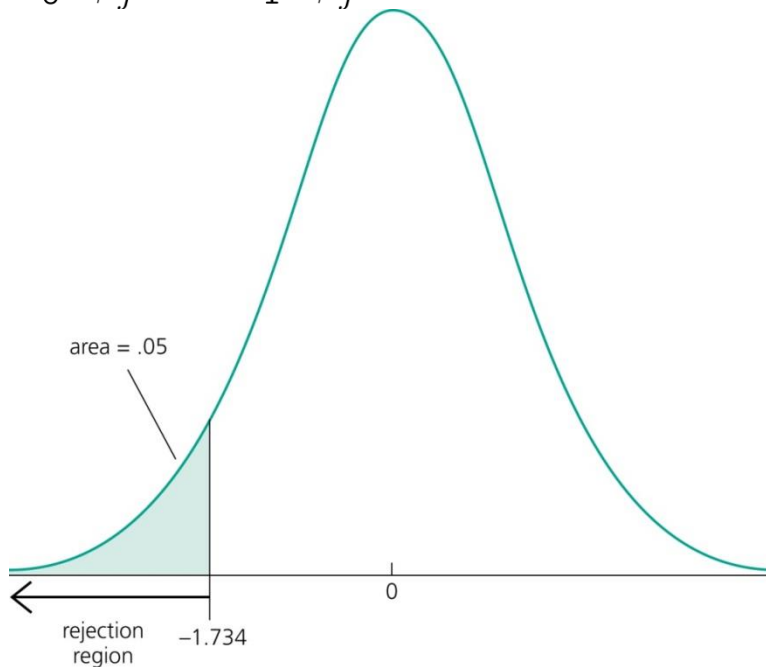
$$t_c = t_{\frac{\alpha}{2}, df} = t_{\frac{0.01}{2}, 526-3} = t_{0.005, 523} = 2.326$$

The null hypothesis is rejected because the t-statistic exceeds the critical value. "The effect of experience on hourly wage is statistically greater than zero at the 5% (and even at the 1%) significance level."

b) Testing against one-sided alternatives (less than zero)

Reject the null hypothesis in favour of the alternative hypothesis if the estimated coefficient is too small“ (i.e. smaller than a critical value).

$$H_0 : \beta_j = 0 \quad H_1 : \beta_j < 0$$



Example: Student performance and school size

Test whether smaller school size leads to better student performance

$$\widehat{math10} = + 2.274 + .00046 \text{ totcomp} + .048 \text{ staff} - .00020 \text{ enroll}$$

$$(6.113) \quad (.00010) \quad (.040) \quad (.00022)$$

Where

Math10=Percentage of students passing maths test

Totcomp=Average annual teacher compensation

Staff=Staff per one thousand students

Enroll=School enrollment (= school size)

$$n = 408, \quad R^2 = .0541$$

$$H_0 : \beta_{enroll} = 0 \quad H_1 : \beta_{enroll} < 0$$

Do larger schools hamper student performance or is there no such effect?

$$t_{enroll} = -.00020 / .00022 \approx -.91$$

$$df = n - k - 1 = 408 - 3 - 1 = 404$$

$$t_c = t_{\frac{\alpha}{2}, df} = t_{\frac{0.05}{2}, 404-3} = t_{0.025, 523} = -1.96$$

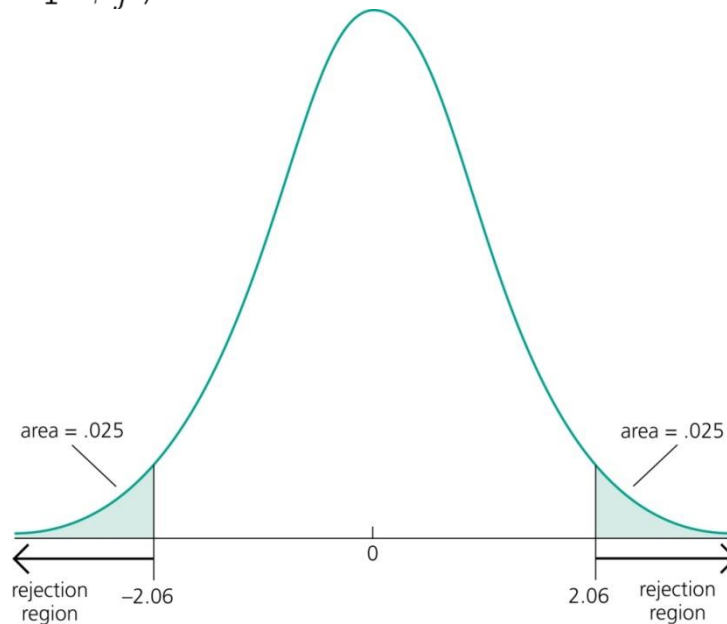
$$t_c = t_{\frac{\alpha}{2}, df} = t_{\frac{0.1}{2}, 404-3} = t_{0.05, 523} = -1.645$$

Critical values for the 5% and the 10% significance level. The null hypothesis is **not rejected** One cannot reject the hypothesis that there is no effect of school size on student performance (not even for a lax significance level of 15%).

c) **Testing against two-sided alternatives**

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$



Reject the null hypothesis in favor of the alternative hypothesis if the absolute value of the estimated coefficient is too large. Construct the critical value so that, if the null hypothesis is true, it is rejected in, for example, 5% of the cases. In the given example, these are the points of the t-distribution so that 5% of the cases lie in the two tails. ! Reject if absolute value of t-statistic is less than -2.06 or greater than 2.06

Example: Student performance and school size

Questions: Test whether smaller school size leads to better student performance

Percentage of students passing maths test

Average annual teacher compensation

Staff per one thousand students

School enrollment (= school size)

$$\widehat{math10} = + 2.274 + .00046 \text{ totcomp} + .048 \text{ staff} - .00020 \text{ enroll}$$

(6.113) (.00010) (.040) (.00022)

$$n = 408, R^2 = .0541$$

$$\text{Test } H_0 : \beta_{enroll} = 0 \text{ against } H_1 : \beta_{enroll} < 0$$

Do larger schools hamper student performance or is there no such effect?

$$t_{enroll} = -.00020 / .00022 \approx -.91$$

$$df = n - k - 1 = 408 - 3 - 1 = 404$$

$$c_{0.05} = -1.65$$

$$c_{0.15} = -1.04$$

Degrees of freedom;
here the standard normal
approximation applies

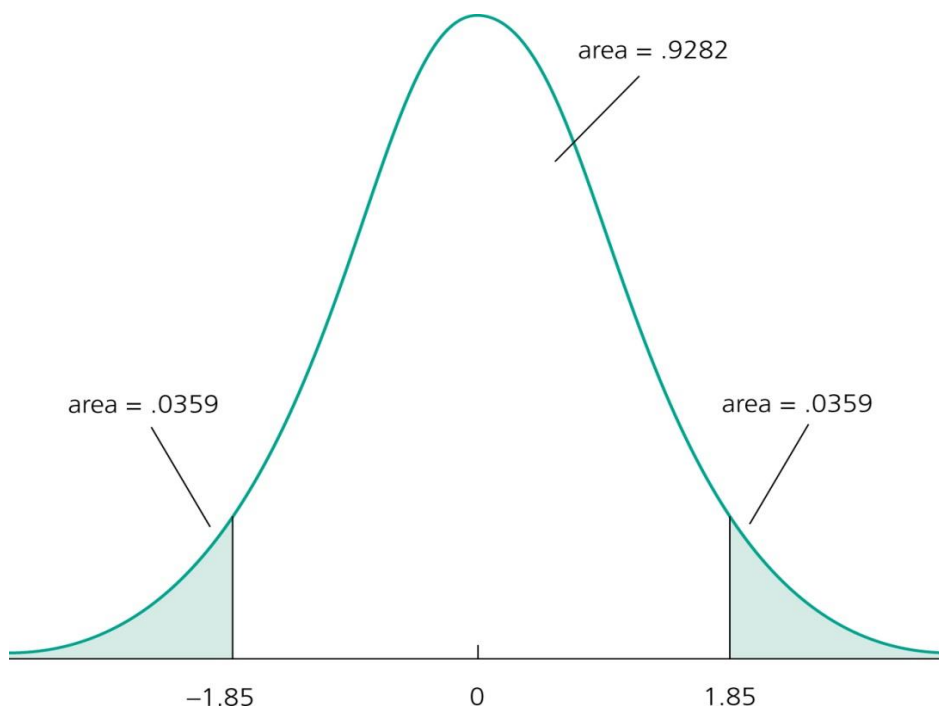
Critical values for the 5% and the 15% significance level.

The null hypothesis is not rejected

3.5.2 Computing p-values for t-tests

A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true.

If the significance level is made **smaller and smaller**, there will be a point where the null hypothesis **cannot** be rejected anymore. The reason is that, by lowering the significance level, one wants to avoid more and more to make the error of rejecting a correct H_0 . The smallest significance level at which the null hypothesis is **still** rejected, is called the p-value of the hypothesis test. A small p-value is evidence **against** the null hypothesis because one would reject the null hypothesis even at small significance levels. A large **p-value** is evidence in **favor** of the null hypothesis. P-values are more informative than tests at fixed significance levels



In the above figure, -1.85/and or 1.85 would be the critical values for a 5% significance level. The **p-value** is the significance level at which one is indifferent between **rejecting** and **not rejecting** the null hypothesis.

In the two-sided case, the p-value is thus the probability that the **t-distributed** variable takes on a larger absolute value than the realized value of the test statistic, e.g.:

$$P(|t - ratio| > 1.85) = 2(.0359) = .0718$$

From this, it is clear that a null hypothesis is rejected if and only if the corresponding p-value is smaller than the significance level. For example, for a significance level of 5% the t-statistic would not lie in the rejection region.

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.

One cannot reject the hypothesis that there is no effect of school size on student performance (not even for a lax significance level of 15%).

3.5.3 Confidence intervals (CI) and hypothesis testing

Under the CLM assumptions, we can easily construct a confidence interval (CI) for the population parameter β_j . Confidence intervals are also called interval estimates because they provide a range of likely values for the population parameter, and not just a point estimate.

Simple manipulation of the result in Theorem 3.2 implies that

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

has a t distribution with n –k-1 degrees of freedom

Using the fact that simple manipulation leads to a CI for the unknown β_j : a 95% confidence interval, given by

$$P(\hat{\beta}_j - t_{0.05} \cdot se(\hat{\beta}_j) < \hat{\beta}_j < \hat{\beta}_j + t_{0.05} \cdot se(\hat{\beta}_j) = 0.95$$

lower bound of the
Confidence interval

upper bound of the
Confidence interval

Where

$t_{0.05}$ =Critical value of two-sided test

Interpretation of the confidence interval

- The bounds of the interval are random. In repeated samples, the interval that is constructed in the above way will cover the population regression coefficient in 95% of the cases.

Confidence intervals for typical confidence levels

$$P(\hat{\beta}_j - t_{0.001} \cdot se(\hat{\beta}_j) < \beta_j < \hat{\beta}_j + t_{0.001} \cdot se(\hat{\beta}_j) = 0.99$$

$$P(\hat{\beta}_j - t_{0.05} \cdot se(\hat{\beta}_j) < \beta_j < \hat{\beta}_j + t_{0.05} \cdot se(\hat{\beta}_j) = 0.95$$

$$P(\hat{\beta}_j - t_{0.01} \cdot se(\hat{\beta}_j) < \beta_j < \hat{\beta}_j + t_{0.01} \cdot se(\hat{\beta}_j) = 0.90$$

Relationship between confidence intervals and hypotheses tests

$$t_{0.01}=2.576, t_{0.05}=1.96 \quad 0.1=1.645$$

Decision Criteria: We reject the null hypothesis if the interval does not include “zero”

Example

Suppose that you have estimated the following model;

$$\widehat{\log(rd)} = -4.38 + 1.084 \log(sales) + .0218 \text{ profmarg}$$

$$(.47) \quad (.060) \quad (.0217)$$

Where rd= Spending on R&D, sales= Annual sales and profmarg=Profits as percentage of sales

$$n = 32, R^2 = .918, df = 32 - 2 - 1 = 29 \Rightarrow c_{0.05} = 2.045$$

- Construct the 95% confidence interval for sales and profit.
- Examine whether sales and profit affect spending on research and Development?

Solution

95% CI for sales

$$= (.961, 1.21)$$

95% CI for Profit

$$= (-.0045, .0479)$$

Interpretation : The effect of **sales** on R&D is relatively precisely estimated as the interval is narrow. Moreover, the effect is significantly different from zero because **zero** is outside the interval.

Interpretation: This effect is imprecisely estimated as the interval is **very wide**. It is not even statistically significant because **zero** lies in the interval.

3.5.4 Test of Overall Significance

Throughout 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. A single (individual) hypothesis is not equivalent to testing that same hypothesis. The intuitive reason for this is that in a joint test of several hypotheses any single hypothesis is affected by the information in the other hypothesis."

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 compliant 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 $RRSS - URSS \geq 0$.

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:

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

$$H_1 : \text{at least one of the } \beta_k \text{ is different from zero.}$$

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.

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

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

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

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

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

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).

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 $\sum y^2 = TSS$ 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.

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

Suppose that you estimated the demand for meat (q1) which depend on price of meat and income of the consumer(y). If critical the critical value of $F_{0.05, 2, 27}=3.35$, examine whether the result is valid or not.

Source	SS	df	MS	Number of obs	=	30
				F(2, 27)	=	27.48
Model	105297.059	2	52648.5293	Prob > F	=	0.0000
Residual	51734.3635	27	1916.08754	R-squared	=	0.6705
				Adj R-squared	=	0.6461
Total	157031.422	29	5414.87662	Root MSE	=	43.773

q1	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
p1	-.7085007	.2391904	-2.96	0.006	-1.199279	-.2177225
y	2.262142	.3342771	6.77	0.000	1.576262	2.948022
_cons	14.90447	36.43313	0.41	0.686	-59.85013	89.65907

Note: SS=sum of square, df=degree freedom, MS=mean sum, q1=quantity demanded of meat, p1=price of meat, y=income of the consumer. “cons” refers to β_0 , coef. Refers to estimated coefficients, std.error refers to standard error, t=statistic, p>|t| refers to probability of rejecting H0 and 95% is the confidence interval for population parameter.

The interpretation of the result in the above Table;

Model refers to variation in the dependent variable due to the independent variable. In this table, SS refers to sum of square and the value=105297.059 is the explained sum of square (SSE) and MS refers to mean sum of square (MSE);

$$MSE = \frac{SSE}{k} = \frac{105297.059}{2} = 52648.5293, \text{ where } k = \text{number of independent variables.}$$

In addition, residual is the estimated error. It is source of variation in the dependent variable which is because of error. Under the column SS, the value 51734.3635 is the residual sum of square (SSR) and the value under column MS =1916.08754 is mean residual sum of square which is computed as follows;

$$MSR = \frac{SSR}{n-k-1} = \frac{51734.3635}{30-2-1} = 1916.08754, \text{ where } n = \text{the number of observation and } k = \text{number of independent variables and } n-k-1 \text{ determines the degree freedom in the residual sum of square.}$$

Moreover, in the above result, total refers total variation in the dependent is decomposed into SSE and SSR. Therefore,

- SST=SSE+SSR.

- $MST = MSE + MSR$
- $df(\text{total}) = k(df \text{ for SSE}) + n - k + 1(df \text{ for SSR}) = n - 1 = 30 - 1 = 29$

The first Panel of the above table reports, the result which is used to examine the overall significance of regression coefficients (F-test).

To conduct the overall significance of coefficients;

1. State the hypothesis;

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

H_1 : H_0 is not true or at least one coefficient is not zero

2. Compute F-statistics:

$$F_{stat} = \frac{\frac{SSE}{K}}{\frac{SSR}{n-k-1}} = 27.48, \text{ see the above table}$$

3. Decide level of significance(α)=0.05

4. Determine the critical value of F-test;

$$F_c = F_{\alpha, df1, df2} = F_{0.05, k, n-k-1} = F_{0.05, 2, 27} = 3.35$$

Where $df1$ =degree freedom for SSE and $df2$ =degree freedom for SSR

5. Decision rule;

- Reject H_0 if $F_{stat} > F_c$ and conclude that the overall coefficients are significant meaning that at least one coefficient not zero.
- In our case, $F_{stat} > F_c = 27.48 > 3.35$, so, we reject the null hypothesis.

The lower panel of the above table presents the estimated result of demand for meat. So, how examine whether p_1 and y affect q_1 ? We need to conduct hypothesis about the significance of each coefficients for p_1 and y .

CHAPTER FOUR:

VIOLATIONS OF CLASSICAL ASSUMPTION

Recall that in the classical model we have assumed

- a) Zero mean of the random term
- b) Constant variance of the error term (i.e., the assumption of homoscedasticity)
- c) No autocorrelation of the error term
- d) Normality of the error term
- e) No multicollinearity among the explanatory variable.

4.1 The Assumption of Zero Expected Disturbances

This assumption is imposed by the stochastic nature of economic relationships, which otherwise it would be impossible to estimate with the common rule of mathematics. The assumption implies that the observations of Y and X must be scattered around the line in a random way (and hence the estimated line $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ be a good approximation of the true line.) This defines the relationship connecting Y and X ‘on the average’. The alternative possible assumptions are either $E(U) > 0$ or $E(U) < 0$. Assume that for some reason the U’s had not an average value of zero, but tended most of them to be positive. This would imply that the observation of Y and X would lie above the true line.

It can be shown that by using these observations we would get a bad estimate of the true line. If the true line lies below or above the observations, the estimated line would be biased.

Note that there is no test for the verification of this assumption because the assumption $E(U) = 0$ is forced upon us if we are to establish the true relationship. i.e, we set $E(U) = 0$ at the outset of our estimation procedure. Its plausibility should be examined in each particular case on a priori grounds. In any econometric application we must be sure that the things are fulfilled so as to be safe from violating the assumption of $E(U) = 0$

- i) All the important variables have been included into the function.
- ii) There are no systematically positive or systematically negative errors of measurement in the dependent variable.

4.2 The Nature of Heteroscedasticity

The assumption of homoscedasticity (or constant variance) about the random variable U is that its probability distribution remains the same over all observations of X , and in particular that the variance of each U_i is the same for all values of the explanatory variable. Symbolically we have

$$\text{Var}(U) = E\{(U_i - E(U))^2\} = E(U_i) = \sigma_u^2, \text{ constant}$$

If the above is not satisfied in any particular case, we say that the U 's are heteroscedastic. That is

$\text{Var}(U_i) = \sigma_{ui}^2$ not constant. The meaning of homoscedasticity is that the variation of each U_i around its zero mean does not depend on the value of X . That is $\sigma_{ui}^2 \neq f(X_i)$.

Note that if σ_u^2 is not constant, but its value depends on X , we may write $\sigma_{ui}^2 = f(X_i)$. As shown in the above diagrams there are various forms of heteroscedasticity. For example in figure (c) the variance of U_i decreases as X increases.

Furthermore, suppose we have a cross-section sample of family budget from which we want to measure the savings function. That means $\text{Saving} = f(\text{income})$. In this case the assumption of constant variance of the U 's is not appropriate, because high-income families show a much greater variability in their saving behavior than do low income families. Families with high income tend to stick to a certain standard of living and when their income falls they cut down their savings rather than their consumption expenditure. But this is not the case in low income families. Hence, the variance of U_i 's increase as income increases.

Note, however, that the problem of heteroscedasticity is the problem of cross-sectional data rather than time series data. That is, the problem is more serious on cross section data.

Causes of Heteroscedasticity

Heteroscedasticity can also arise as a result of several cases. The first one is the presence of outliers (i.e., extreme values compared to the majority of a variable). The inclusion or exclusion of such an observation, especially if the sample size is small, can substantially alter the results of regression analysis. With outliers it would be hard to maintain the assumption of homoscedasticity.

Another source of heteroscedasticity arises from violating the assumption that the regression model is correctly specified. Very often what looks like heteroscedasticity may be due to the fact that some important variables are omitted from the model. In such situation the residuals obtained from the regression may give the distinct impression that the error variance may not be constant. But if the omitted variables are included in the model, the impression may disappear.

The consequence of Heteroscedasticity

If the assumption of homoscedastic disturbance is not fulfilled we have the following consequences:

- i) If U is heteroscedastic, the OLS estimates do not have the minimum variance property in the class of unbiased estimators; that is, they are inefficient in small samples. Furthermore, they are inefficient in large samples
- ii) The coefficient estimates would still be statistically unbiased. That is the expected value of the $\hat{\beta}_i$'s will equal to the true parameters, $E(\hat{\beta}_i) = \beta_i$
- iii) The prediction (of Y for a given value of X) would be inefficient because of high variance. This is because the variance of the prediction includes the variances of U and of the parameter estimates, which are not minimum due to the incidence of heteroscedasticity.

In any case how does one detect whether the problem really exists.

4.3 The Nature of Autocorrelation

An important assumption of the classical linear model is that there is no autocorrelation or serial correlation among the disturbances U_i entering into the population regression function. This assumption implies that the covariance of U_i and U_j is equal to zero. That is:

$$\text{Cov}(U_i U_j) = E\{[U_i - E(U_i)][U_j - E(U_j)]$$

$$= E(U_i U_j) = 0 \text{ (for } i \neq j \text{)}$$

But if this assumption is violated, it implies that the disturbances are said to be autocorrelated. This could arise for several reasons.

- i) **Spatial autocorrelation:** In regional cross-section data, a random shock affecting economic activity in one region may cause economic activity in an adjacent region to change because of close economic ties between the regions. Shocks due to

weather similarities might also tend to cause the error terms between adjacent regions to be related.

- ii) ***Prolonged influence of shocks***: In time series data, random shocks (disturbances) have effects that often persist over more than one time period. An earthquake, flood, strike or war, for example, will probably affect the economy's operation in periods.
- iii) ***Inertia***: past action often have a strong effect on current actions, so that a positive disturbance in one period is likely to influence activity in succeeding periods.
- iv) Data manipulation published data often undergo interpolation or smoothing, procedures that average true disturbances over successive time periods.
- v) Misspecification: An omitted relevant independent variable that is autocorrelated will make the disturbance (associated with the misspecified model) autocorrelated. An incorrect functional form or a misspecification of the equation's dynamics could do the same. In these instances the appropriate procedure is to correct the misspecification.

Note that autocorrelation is a special case of correlation. Autocorrelation refers to the relationship not between two (or more) different variables, but between the successive values of the same variable (where in this section we are particularly interested in the autocorrelation of the U 's. Moreover, note that the term autocorrelation and serial correlation are treated synonymously.

Consequences of Autocorrelation

When the disturbance term exhibits serial correlation the values as well as the standard errors of the parameter estimates are affected.

- i) If disturbances are correlated, the prevailing value of the disturbances has some information to convey about the current disturbances. If this information is ignored it is clear that the sample data is not being used with maximum efficiency. However the estimates of the parameters do not have the statistical bias even when the residuals are serially correlated. That is, the parameter of OLS estimates is statistically unbiased in the sense that their expected value is equal to the true parameter.
- ii) The variance of the random term U may be seriously underestimated. In particular, the underestimation of the variance of U will be more serious in the case of positive autocorrelation of the error term (U_t). With positive first-order autocorrelated errors

it implies that fitting an OLS estimating line clearly gives an estimate quite wide of the mark. The high variation in these estimates will cause the variance of β^{OLS} to be greater than it would have been had the errors been distributed randomly.

4.4 Multicollinearity

One of the assumptions of the classical linear regression model (CLRM) is that there is no perfect multicollinearity among the regressors included in the regression model. Note that although the assumption is said to be violated only in the case of exact multicollinearity (i.e., an exact linear relationship among some of the regressors), the presence of multicollinearity (an approximate linear relationship among some of the regressors) lead to estimating problems important enough to warrant out treating it as a violation of the classical linear regression model. Multicollinearity does not depend on any theoretical or actual linear relationship among any of the regressors; it depends on the existence of an approximate linear relationship in the data set at hand. Unlike most other estimating problems, this problem is caused by the particular sample available. Multicollinearity in the data could arise for several reasons. For example, the independent variables may all share a common time trend, one independent variable might be the lagged value of another that follows a trend, some independent variable may have varied together because the data were not collected from a wide enough base, or there could in fact exist some kind of approximate relationship among some of the regressors.

Note that the existence of multicollinearity will affect seriously the parameter estimates. Intuitively, when any two explanatory variables are changing in nearly the same way, it becomes extremely difficult to establish the influence of each one regressors on the dependent variable separately. That is, if two explanatory variables change by the same proportion, the influence on the dependent variable by one of the explanatory variables may be erroneously attributed to the other. Their effect cannot be sensibly investigated, due to the high inter correlation.

In general, the problem of multicollinearity arises when individual effects of explanatory variables cannot be isolated and the corresponding parameter magnitudes cannot be determined with the desired degree of precision. Though it is quite frequent in cross section data as well, it should be noted that it tends to be more common and more serious problem in time series data.

Consequences of Multicollinearity

In the case of near or high multicollinearity, one is likely to encounter the following consequences

- i) Although BLUE, the OLS estimators have large variances and covariances, making precise estimation difficult. This is clearly seen through the formula of variance of the estimators.
- ii) Because of consequence (i), the confidence interval tend to be much wider, leading to the acceptance of the “Zero null hypothesis” (i.e., the true population coefficient is zero).
- iii) Because of consequence (i), the t-ratio of one or more coefficient’s tend to be statistically insignificant.
- iv) Although the t-ratio of one or more coefficients is statistically insignificant, R^2 , the overall measure of goodness of fit, can be very high. This is the basic symptom of the problem.
- v) The OLS estimators and their standard errors can be sensitive to small changes in the data. That is when few observations are included, the pattern of relationship may change and affect the result.
- vi) Forecasting is still possible if the nature of the collinearity remains the same within the new (future) sample observation. That is, if collinearity exists on the data of the past 15 years sample, and if collinearity is expected to be the same for the future sample period, then forecasting will not be a problem.

4.5 Application examples on violation of CLR

Suppose that you estimated the Model; $q1 = \beta_0 + 1p_1 + \beta_2 y$ with following result

Source	SS	df	MS	Number of obs	=	30
				F(2, 27)	=	27.48
Model	105297.059	2	52648.5293	Prob > F	=	0.0000
Residual	51734.3635	27	1916.08754	R-squared	=	0.6705
				Adj R-squared	=	0.6461
Total	157031.422	29	5414.87662	Root MSE	=	43.773

q1	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
p1	-.7085007	.2391904	-2.96	0.006	-1.199279	-.2177225
y	2.262142	.3342771	6.77	0.000	1.576262	2.948022
_cons	14.90447	36.43313	0.41	0.686	-59.85013	89.65907

Diagnostics testing is the process of examining the violation of classical linear regression. The basic assumption that should be checked in MLR is the following;

1. Specification error test

Ramsey RESET test using powers of the fitted values of q1

Ho: model has no omitted variables

$$F(3, 24) = 2.22$$

$$\text{Prob} > F = 0.1120$$

Specification error is the violation of zero conditional mean assumption. Omitting important variable from regression would leads to correlation between error and that omitted variables. This create specification error. Ramsey RESET test is used to evaluate whether there is specification error or not. It based on the F-test which has Ho: model has no omitted variables (see the above table). The probability value computed from F-statistic is 0.112 which greater 5%. So, we fail to reject Ho and conclude that the model has no omitted variables.

2. Heteroscedasticity test

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity

Ho: Constant variance

Variables: fitted values of q1

$$F(1, 28) = 1.16$$

$$\text{Prob} > F = 0.2916$$

The other important in assumption of classical linear regression is homoscedasticity of the error term. That is error term has constant variance. The violation of this assumption is called heteroscedasticity. The common diagnostic test for heteroscedasticity is called the Breusch-Pagan / Cook-Weisberg test, which based on F-test. It has the H_0 : Constant variance. We need the larger p-value to preserve this hypothesis, which will computed from F-statistic. As we observe the above table, $P=0.29$ which far greater than 5% standard value .So, we fail to reject the H_0 and conclude that error term has constant variance.

3. Multi-Collinearity

Variable	VIF	1/VIF
p1	1	0.99991
y	1	0.99991
Mean VIF	1	

Yet, multi-collinearity is another important assumption in classical linear regression. We use the variance inflation factor (VIF) method to examine the violation of no high correlation between independent variables. From the table, $VIF=1$, which less 5 implying that there is no multi-collinearity