ARBAMINCH UNIVERSITY
FACULTY OF BUSINESS AND ECONOMICS
DEPARTMENT OF ECONOMICS

MODULE
INTRODUCTION TO ECONOMETRICS
COURSE CODE:
CREDIT HOUR: 3

PREPARED BY: TSEGAYE TAGESSE
EDITED BY: FIKRU DEBELE

JANUARY, 2010

TABLE OF CONTENTS

Contents

CHAPTER I.....ERROR! BOOKMARK NOT DEFINED.

INTRODUCTION.....ERROR! BOOKMARK NOT DEFINED.

1.1 WHAT IS ECONOMETRICS? .....**ERROR! BOOKMARK NOT DEFINED.**
1.2 ECONOMETRICS AND OTHER BRANCHES OF SCIENCE.....**ERROR! BOOKMARK NOT DEFINED.**
1.2.1 Econometrics and mathematical economics ..... ***Error! Bookmark not defined.***
1.2.2 Econometrics and statistics..... ***Error! Bookmark not defined.***
1.3 GOALS OF ECONOMETRICS .....**ERROR! BOOKMARK NOT DEFINED.**
1.4 METHODOLOGY OF ECONOMETRIC RESEARCH.....**ERROR! BOOKMARK NOT DEFINED.**
EXERCISES.....**ERROR! BOOKMARK NOT DEFINED.**

CHAPTER 2 .....ERROR! BOOKMARK NOT DEFINED.

SIMPLE LINEAR REGRESSION MODEL .....ERROR! BOOKMARK NOT DEFINED.

OVERVIEW .....ERROR! BOOKMARK NOT DEFINED.

2.1 THE MEANING OF REGRESSION ANALYSIS .....**ERROR! BOOKMARK NOT DEFINED.**
2.2 POPULATION AND SAMPLE REGRESSION FUNCTIONS.....**ERROR! BOOKMARK NOT DEFINED.**
2.3 THE ORDINARY LEAST SQUARES (OLS) METHOD.....**ERROR! BOOKMARK NOT DEFINED.**
2.4 ASSUMPTIONS UNDERLYING THE LEAST SQUARES (OLS) METHOD .....**ERROR! BOOKMARK NOT DEFINED.**
2.5 THE DISTRIBUTION OF THE DEPENDENT VARIABLE, Y .....**ERROR! BOOKMARK NOT DEFINED.**
2.6 THE LEAST SQUARES CRITERION AND THE OLS ESTIMATES.....**ERROR! BOOKMARK NOT DEFINED.**
2.7 ESTIMATION OF A FUNCTION WHOSE INTERCEPT IS ZERO.....**ERROR! BOOKMARK NOT DEFINED.**
2.8 FUNCTIONAL FORMS OF REGRESSION MODELS .....**ERROR! BOOKMARK NOT DEFINED.**
2.8.1 The Leg – linear model ..... ***Error! Bookmark not defined.***
2.8.2 Semi-log Models..... ***Error! Bookmark not defined.***
EXERCISES.....**ERROR! BOOKMARK NOT DEFINED.**

CHAPTER 3 .....ERROR! BOOKMARK NOT DEFINED.

PROPERTIES OF OLS ESTIMATORS.....ERROR! BOOKMARK NOT DEFINED.

OVERVIEW .....ERROR! BOOKMARK NOT DEFINED.

3.1 THE MEAN, VARIANCE AND STANDARD ERRORS OF OLS ESTIMATES AND THE ERROR TERM.. **ERROR! BOOKMARK NOT DEFINED.**
3.1.1 The Mean of $\hat{\beta}_1$ ..... ***Error! Bookmark not defined.***
3.1.2 The Variance of $\hat{\beta}_1$ ..... ***Error! Bookmark not defined.***
3.1.3 The mean of $\hat{\beta}_0$ ..... ***Error! Bookmark not defined.***
3.1.4 The Variance of $\hat{\beta}_0$ ..... ***Error! Bookmark not defined.***
3.1.5 The Variance of the Random Variable U ..... ***Error! Bookmark not defined.***
3.1.6 Standard errors of least squares estimates ..... ***Error! Bookmark not defined.***
3.2 THE GAUSS – MARKOV THEOREM.....**ERROR! BOOKMARK NOT DEFINED.**
3.2.1 Linearity..... ***Error! Bookmark not defined.***
3.2.2 Unbiasedness ..... ***Error! Bookmark not defined.***
3.2.3 Minimum Variance..... ***Error! Bookmark not defined.***
EXERCISE .....**ERROR! BOOKMARK NOT DEFINED.**

CHAPTER 4 ..... ERROR! BOOKMARK NOT DEFINED.

STATISTICAL TESTS OF THE REGRESSION MODEL ..... ERROR! BOOKMARK NOT DEFINED.

OVERVIEW ..... ERROR! BOOKMARK NOT DEFINED.

4.1 THE COEFFICIENT OF DETERMINATION (r^2): A MEASURE OF "GOODNESS OF FIT" ..... **ERROR! BOOKMARK NOT DEFINED.**

4.2 HYPOTHESIS TESTING ..... **ERROR! BOOKMARK NOT DEFINED.**

4.2.1 The Test of Significance Approach ..... ***Error! Bookmark not defined.***

4.2.2.1 The standard error test of the least square estimates **Error! Bookmark not defined.**

4.2.2.2 The Z – Test of the Least – Square Estimates ..... **Error! Bookmark not defined.**

4.2.2.3 The Student's t test ..... **Error! Bookmark not defined.**

4.2.2 Confidence Interval Approach to Hypotheses Testing ..... ***Error! Bookmark not defined.***

4.2.2.1 Confidence Intervals for β_0 *and* β_1 ..... **Error! Bookmark not defined.**

4.2.2.2 Confidence interval from standard normal distribution .. **Error! Bookmark not defined.**

4.2.2.3 Confidence interval from the student's t distribution ..... **Error! Bookmark not defined.**

EXERCISE ..... **ERROR! BOOKMARK NOT DEFINED.**

CHAPTER FIVE ..... ERROR! BOOKMARK NOT DEFINED.

MULTIPLE LINEAR REGRESSION MODELS ..... ERROR! BOOKMARK NOT DEFINED.

OVERVIEW ..... ERROR! BOOKMARK NOT DEFINED.

5.1 THE THREE VARIABLES MODELS ..... **ERROR! BOOKMARK NOT DEFINED.**

5.2 ESTIMATION OF THE PARAMETERS OF MULTIPLE LINEAR REGRESSION MODEL ... **ERROR! BOOKMARK NOT DEFINED.**

5.3 THE MULTIPLE COEFFICIENT OF DETERMINATION, R^2 ..... **ERROR! BOOKMARK NOT DEFINED.**

5.4 SIMPLE, PARTIAL AND MULTIPLE CORRELATION COEFFICIENTS ... **ERROR! BOOKMARK NOT DEFINED.**

5.5 THE MEAN AND VARIANCE OF THE PARAMETER ESTIMATES ..... **ERROR! BOOKMARK NOT DEFINED.**

5.6 INTERPRETATION OF REGRESSION COEFFICIENTS ..... **ERROR! BOOKMARK NOT DEFINED.**

5.7 STATISTICAL INFERENCE IN MULTIPLE REGRESSION MODEL ..... **ERROR! BOOKMARK NOT DEFINED.**

5.7.1 Statistical significance of individual coefficients in multiple regression .... ***Error! Bookmark not defined.***

5.7.2 Testing the overall significance of the sample regression function ..... ***Error! Bookmark not defined.***

5.8 MATRIX APPROACH TO MULTIPLE LINEAR REGRESSION MODELS **ERROR! BOOKMARK NOT DEFINED.**

EXERCISE ..... **ERROR! BOOKMARK NOT DEFINED.**

CHAPTER SIX ..... ERROR! BOOKMARK NOT DEFINED.

VIOLATIONS OF BASIC ASSUMPTIONS OF LINEAR REGRESSION MODELS ..... ERROR! BOOKMARK NOT DEFINED.

OVERVIEW ..... ERROR! BOOKMARK NOT DEFINED.

6.1 AUTOCORRELATION ..... **ERROR! BOOKMARK NOT DEFINED.**

6.1.1 Introduction ..... ***Error! Bookmark not defined.***

6.1.2 Sources of autocorrelation ..... ***Error! Bookmark not defined.***

6.1.3 The first-order autoregressive scheme ..... ***Error! Bookmark not defined.***

6.1.4 Consequences of Autocorrelation ..... ***Error! Bookmark not defined.***
6.1.5 Tests for Autocorrelation ..... ***Error! Bookmark not defined.***
6.1.6 Solutions for Autocorrelation..... ***Error! Bookmark not defined.***
6.2 HETEROSCEDASTICITY: WHAT HAPPENS IF THE ERROR VARIANCE IS NOT CONSTANT? ..... **ERROR! BOOKMARK NOT DEFINED.**
6.2.1 Introduction..... ***Error! Bookmark not defined.***
6.2.2 Plausibility of the Assumption of Homoscedasticity ..... ***Error! Bookmark not defined.***
6.2.3 Consequences of Heteroscedasticity ..... ***Error! Bookmark not defined.***
6.2.4 Tests for Heteroscedasticity ..... ***Error! Bookmark not defined.***
6.2.5 Remedial Measures for Heteroscedasticity..... ***Error! Bookmark not defined.***
6.3 MULTI-COLLINEARITY: WHAT HAPPENS IF THE REGRESSORS ARE CORRELATED?. **ERROR! BOOKMARK NOT DEFINED.**
6.3.1 Definition ..... ***Error! Bookmark not defined.***
6.3.2 Sources of multicollinearity. ..... ***Error! Bookmark not defined.***
6.3.3 Consequences of Multicollinearity..... ***Error! Bookmark not defined.***
6.3.4 Detection of multicollinearity ..... ***Error! Bookmark not defined.***
EXERCISE ..... **ERROR! BOOKMARK NOT DEFINED.**

CHAPTER SEVEN ..... ERROR! BOOKMARK NOT DEFINED.

MODEL SPECIFICATION ..... ERROR! BOOKMARK NOT DEFINED.

OVER VIEW ..... ERROR! BOOKMARK NOT DEFINED.

7.1 ATTRIBUTES OF A GOOD MODEL ..... **ERROR! BOOKMARK NOT DEFINED.**
7.2. TYPES OF SPECIFICATION ERRORS ..... **ERROR! BOOKMARK NOT DEFINED.**
7.2.1. Omission of a relevant variable(s)..... ***Error! Bookmark not defined.***
7.2.2. Inclusion of an unnecessary variable(s)..... ***Error! Bookmark not defined.***
7.2.3 Adopting the wrong functional form ..... ***Error! Bookmark not defined.***
7.2.4. Errors of measurement..... ***Error! Bookmark not defined.***
7.2.5. Incorrect specification of the stochastic error term..... ***Error! Bookmark not defined.***
7.3. CONSEQUENCES OF MODEL SPECIFICATION ERRORS ... **ERROR! BOOKMARK NOT DEFINED.**
7.3.1 Consequences of omitting relevant variables..... ***Error! Bookmark not defined.***
7.3.2. Consequences of Inclusion of an Irrelevant Variable ..... ***Error! Bookmark not defined.***
7.4. TESTS OF SPECIFICATION ERRORS ..... **ERROR! BOOKMARK NOT DEFINED.**
7.4.1. Detecting the Presence of Unnecessary Variables..... ***Error! Bookmark not defined.***
7.4.2 Tests for Omitted Variables and Incorrect Functional Form ..... ***Error! Bookmark not defined.***
EXERCISE ..... **ERROR! BOOKMARK NOT DEFINED.**

CHAPTER EIGHT ..... ERROR! BOOKMARK NOT DEFINED.

DUMMY VARIABLE REGRESSION MODELS..... ERROR! BOOKMARK NOT DEFINED.

OVERVIEW ..... ERROR! BOOKMARK NOT DEFINED.

8.1 DEFINITION OF DUMMY VARIABLES ..... **ERROR! BOOKMARK NOT DEFINED.**
8.2 DUMMY VARIABLES FOR DIFFERENCES IN INTERCEPT TERMS..... **ERROR! BOOKMARK NOT DEFINED.**
8.3 DUMMY VARIABLES FOR CHANGES IN SLOPE COEFFICIENTS ..... **ERROR! BOOKMARK NOT DEFINED.**
8.4 DUMMY VARIABLES IN SEASONAL ANALYSIS ..... **ERROR! BOOKMARK NOT DEFINED.**
EXERCISE ..... **ERROR! BOOKMARK NOT DEFINED.**

PREFACE

In recent times, it has become common to see econometric applications in economic planning, research, and polices. Econometrics has also become an integral part of social, agricultural, health and other researches. Particularly, the tremendous advancement in computer technology has made econometrics a handy tool of economists to explain the complex realities of the actual world. Hence, this module is prepared with the objective of introducing students of economics to the basic and introductory knowledge of econometrics. Succinctly, the material is designed to enable students to know how to measure economic variables (both quantitative and qualitative) so as to tell the exact extent of the effect that some variables may have on some other variables. Furthermore, the module is prepared with the contention of enabling students to develop strong background about the application of econometrics to policy analysis, decision making, and so on.

This material assumes that readers have sufficient background of statistics, calculus and algebra. As the module is mainly meant for students of economics, it requires solid knowledge of the basic principles and theories of economics as a sin qua none. On the other hand, as the material is presumed for beginners, it gives special emphasis to the elementary and introductory components of econometrics without clutter the exposition with too much algebra.

The module is organized in such a way that the first chapter introduces students to the definition and scope of econometrics, and gives a brief explanation of the methodological stages of econometric research. The second chapter deals with the simple linear regression models and the Ordinary Least Squares (OLS) method used to estimate the models, leaving the explanation of the statistical properties of the OLS estimates chapter three and the explanation of statistical significance of the estimates to chapter four. While, chapter five extends the lessons of simple linear regression models to multiple linear regression models, chapter six relaxes the assumptions of linear

regression models. Finally, chapter seven deals with model misspecification and chapter eight deals with dummy variables.

CHAPTER 2

SIMPLE LINEAR REGRESSION MODEL

Overview

Dear learners in chapter 1, we have learnt the definition and scope of econometrics, and how it differs from different branches of science. Furthermore, we have discussed the stages involved in the methodologies of econometric research.

In this chapter we will give emphasis to linear regression models and the application of ordinary least squares (OLS) method to obtain the estimates of the parameters of the true economic relationships. This means, in this chapter we will learn how to develop formula for the estimates of the parameters by using the method of OLS. Finally, the chapter will emphasize the different ways of developing econometric models based on economic theories.

At the end of this chapter students will be able to:

- Understand regression analysis and how to differentiate it from correlation analysis.
- Know population regression functions and sample regression functions
- Know the Ordinary Least square (OLS) method of estimation and apply it to estimate the parameters of economic functions
- Differentiate among different functional forms of econometric models such as **log–linear models and semi-log models,** and use them in appropriate situations.

2.1 The Meaning of Regression Analysis

Regression analysis is concerned with the study of the dependence of one variable (the dependant variable) on one or more other variables (the explanatory variable(s)) (Gujrati, 2004). The objective of regression analysis could be to estimate and/or predict the (population) mean value of the dependant variable in terms of the known or fixed (in repeated sampling) values of the explanatory variables.

To illustrate the concept of regression analysis, suppose that a researcher collected data on monthly income (Y) and consumption expenditure (C) of 40 families from a hypothetical community. As shown in Table 2.1 the data collected from these 40 families are divided into seven income groups and the monthly expenditures of each family in the seven groups are as shown in the table.

Table 2.1: Monthly Income and Consumption Expenditure of 40 Families

Monthly Income (Y) →	800	1000	1200	1400	1800	2300	3400
Monthly Consumption Expenditure ↓	500	850	790	900	1020	1100	1500
	550	700	800	930	1050	1200	1750
	650	800	940	1030	1400	1250	2500
	700	650	980	1000	1500	1800	2600
	-	880	1000	1100	1550	2000	3000
	-	900	1100	1300	1700	2010	3100

It is evident from the above table that there are seven *fixed* values of *Y* and the corresponding *C* values against each of the fixed Y values. As shown in table, for fixed values of monthly income there can be different values of monthly consumption expenditure. Succinctly, having the same monthly income it's possible for families to have different consumption expenditures. Hence, when

we are taking repeated samples, it is possible that we can generate different samples of the same size with the same data for monthly income but different values for consumption expenditure; this is the essence of fixed values for the explanatory variables in repeated sampling*.

Here, it is worthy to note that these average values of C are conditional on the fixed values of the Y i.e., $E\left(C/Y_i\right)$. This implies that regression analysis is studying the dependence of the monthly expenditure of the families on their monthly income and is used to predict the average monthly expenditures associated with different levels of monthly income.

In a nutshell, regression analysis deals with statistical dependence among variables; but not with functional or deterministic dependence among variables. In statistical relationships we essentially deal with random (stochastic) variables; i.e., variables that have probability distributions.

Stochastic relationship is a relationship wherein for a particular value of the independent variable, there is a probability distribution of the values of the dependent variable. In such a case for any given value of the independent variable (Y in the above example), the dependent variable (C) assumes some specific value only with some probability. In contrary, deterministic relationship is a relationship wherein, for each value of the independent variable there is one and only one corresponding value of the dependent variable.

In econometrics we exclusively deal with stochastic relation ships. The model that describes the relation ship between only two variables is called simple linear regression model. The term linear regression implies that the regression is linear in parameters; but it may or may not be linear in explanatory variables. Although

* Repeated sampling is only hypothetical; in practice we take only one sample and base our regression on this observed sample.

regression analysis deals with the dependence of one variable on other variables, it does not necessarily imply causation. The determination of the direction of causation should come from outside of statistics; for example, from economic theory. In other words, statistical relationships by themselves can not logically imply causation. To ascribe causality, one must appeal to 'a priori' or theoretical considerations.

In addition, regression analysis is closely related to correlation analysis but conceptually there is huge difference between the two analyses. The primary objective of correlation analysis is to measure the strength or degree of linear association between two variables. However, in regression analysis (as already noted), we try to predict the average value of the dependent variable on the basis of fixed values of the explanatory variables. We call these average (mean) values conditional expected values, say $E\left(C/Y_i\right)$, since they are obtained on the basis of the fixed values of the conditioning variable (Y). It is important to distinguish these conditional expected values from the unconditional expected values, $E(C)$; which are calculated simply by summing all the values of C and dividing the result by total number of families. The latter means are so called because in arriving at them we have disregarded the incomes (Y) of the families. In general, the conditional and unconditional mean values are different.

2.2 Population and Sample Regression Functions

As noted above, each conditional mean value of any dependent variable (say Y) is a function of an explanatory variable (say X), where X_i is a given value of X. Symbolically,

$$E\left(Y/X_i\right) = f\left(X_i\right) \quad \text{.....} (2.1)$$

Where $f\left(X_i\right)$ denotes some function of the explanatory variable X. Equation 2.1 is known as the conditional expectation function (CEF) or population regression function (PRF) or population regression (PR), which merely implies that the

expected value of the distribution of Y (given X_i) is functionally related to X_i. In other words, it tells how the average values of Y vary with the values of X.

An important question that should be addressed at this juncture is about the form of the function $f(X_i)$, as in real situations we may not have the entire population available for examination of $f(X_i)$. However, the functional form of PRF is not beyond empirical question, although in specific cases theory may have something to say about it. For example, if we assume (perhaps from theory) that Y and X are linearly related, as a first approximation, the PRF $E\left(Y/X_i\right)$ may be represented as a linear function of X_i as given below:

$$E\left(Y/X\right) = \beta_0 + \beta_1 X_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.2)$$

Where β_0 *and* β_1 are unknown but fixed parameters and known as the regression coefficients.

Therefore, the stochastic specification of PRF is given as:

$$E\left(Y/X_i\right) = Y_i - U_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.3)$$

$$\Rightarrow Y_i = E\left(Y_i/X_i\right) + U_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.3a)$$

$$\Rightarrow \quad Y_i = \beta_0 + \beta_1 X_i + U_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.3b)$$

Thus, in regression analysis we are interested in estimating the PRF, that is, estimating the values of the unknowns β_1 *and* β_2 on the basis of observations on Y and X. However, the challenge is to obtain data on all possible values of Y and X, as in most piratical situations what we have would be sample values of Y associated with fixed X's. Therefore, the usual practice is to estimate the PRF on the basis of the sample information. Nonetheless, the difficulty is that for a fixed value of X, we can have different samples on the values of Y. For example, from

the population of Y values for fixed values of X, we can have the following two samples which are only two of the many possible samples.

Sample: - 1

Y	X
70	80
65	100
90	120
95	240

Sample:-2

Y	X
55	80
88	100
90	120
80	240

Now the question is how to estimate PRF from the observed sample data. This is because the PRF can be estimated on the basis of sample information, though not accurately since sampling always involves sampling fluctuation.

The regression functions based on sample information are called sample regression functions (SRF); for instance, from the above two samples we can have two regression functions to represent the sample regression line. $\left(SRF_1 \quad and \quad SRF_2 \right)$

In equation 2.2 we have noted that $E\left(Y / X_i \right) = \beta_0 + \beta_1 X_i$. Hence, the sample estimator of this relationship would be given as

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad \text{..... 2.4}$$

Where $\hat{Y}_i$ is read as Y – hat and is an estimator of $E\left(Y / X_i \right)$; $\hat{\beta}_0 \quad and \quad \hat{\beta}_1$ are sample estimators of $\beta_0 \quad and \quad \beta_1$ respectively. Equation 2.4 is called sample regression function and its stochastic form is given as:

$$Y_i = \hat{Y}_i + \hat{U}_i$$

Thus,

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{U}_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 2.5$$

Where $\hat{U}_i$ is the estimate of U_i and it defines the sample residual term.

The primary objective of regression analysis is to estimate PRF given as $Y_i = \beta_1 + \beta_2 X_i + U_i$ on the basis of SRF: $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{U}_i$.

Graphically, equation 2.5 is given as

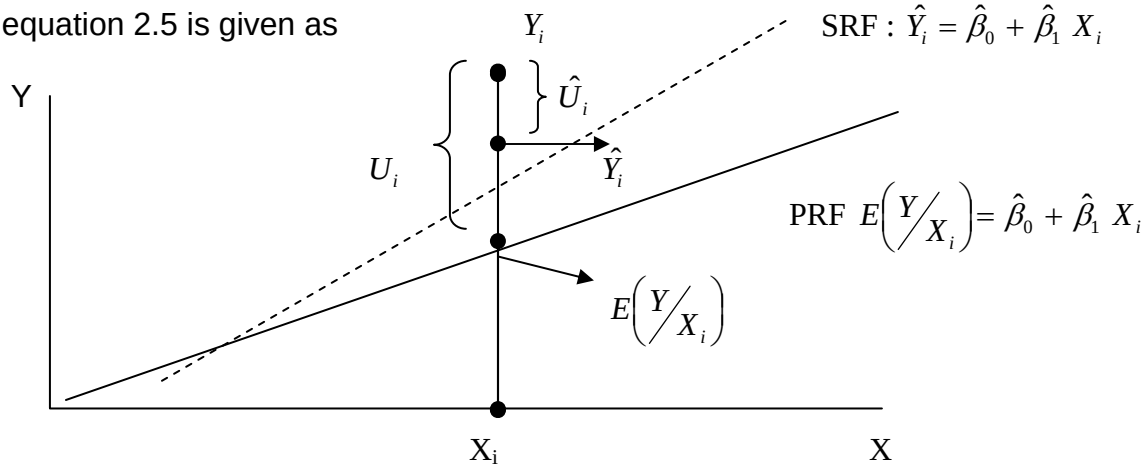


Figure 2.1 Sample and Population Regression Lines.

Note that SRF given in figure 2.2 is only one of the several possible SRFs. So how can we choose the one that best approximates PRF? In other woods, how can we obtain best estimators of the parameters β_0 *and* β_1 based on sample information?

To address this question econometricians developed different techniques; one of which is the **Ordinary Least Squares (OLS) Method**.

2.3 The Ordinary Least Squares (OLS) Method

The OLS method is the most extensively used method of estimation in regression analysis. Under certain assumptions, the least squares method has some attractive statistical properties.

To illustrate the ordinary least squares (OLS) method, think of the theory of supply in economics. In its simplest form, the theory postulates that there is a positive relationship between quantity supplied of a commodity (Y) and its price (X), other things remaining constant.

From Equation 2.3a, we know that the PRF of this relationship is given as:

$$Y_i = E\left(Y/X_i\right) + U_i$$

Assuming linearity, this can be rewritten as

$$\Rightarrow Y_i = \beta_0 + \beta_1 X_i + U_i \quad \text{. ..... (2.6)}$$

Where U_i is a stochastic term and is responsible for different factors that affect the dependent variable(Y_i), but can not explicitly be taken into account by an investigator.

Dear students, "Why do you think an investigator is not in a position to take into account all the factors that affect the dependant variable? ----------

---

Some of the reasons for not taking all the factors that affect the dependent variable into account are discussed as follows:

i) Omissions of variables from the function

In real world, economic variables may be influenced by a very large number of other variables. However, the researcher may not include all of them explicitly in his/her model; which may be attributed to the following reasons:

a) Some of the variables may be unknown to the researcher him/herself
b) Even if all variables are known to the investigator, the available data most often are not adequate to measure all variables that influence the dependent variable.
c) Some of the variables though they are known to be relevant, may not be measured statistically (e.g. tastes, religions, gender etc)
d) Some variables may have, each individually, insignificant influence on the dependent variable
e) Randomness of some variables such as epidemics, earthquakes, war etc, which may make them unpredictable.

Thus, in most cases only a few most important variables would explicitly be included in the model; where the effect of others on the dependent variable is taken in to account by U_i.

ii) Intrinsic Randomness in human behavior

Even if the researcher succeeds in including all the relevant variables into the model, there would be some “intrinsic” randomness in the dependent variable that can not be explained no matter how hard the researcher tries, which may be due to the erratic behavior of human beings.

iii) Misspecification of the model

Albeit the economic phenomena are much more complex than a single equation may reveal, some times researchers may use single equation models.

Furthermore s/he may use linearity to represent the relationship between the dependent and explanatory variables, though the relationship should have been studied by using non–linear models. In either of these cases the researcher ends up with miss specified model and this is one of the reasons why U_i is introduced in econometric models.

iv) Aggregation errors

We often use aggregate data, in which we add magnitudes referring to individuals whose behaviors are dissimilar. Hence, in the process of aggregation attributes expressing individual peculiarities would be lost.

Therefore, in order to take into account the above sources of errors, we introduce a random variable in econometric models, which is usually dented by U and is **called error term or random disturbance term or stochastic term.** U is so called because it is supposed to disturb the exact linear relationship supposed to exist between Y and X.

Having studied the relevance of U_i to economic relationships, the economic theory of supply in its simplest form can be modeled as:

$$Y_i = \beta_0 + \beta_1 X_i + U_i \quad \text{.....2.7}$$

Where U_i represents all other variables than the price of the commodity that affect the quantity supplied (Y). However, the relationship represented in Equation 2.7 is not directly observable and hence, we have to estimate it on the basis of sample information. To estimate $\beta_0 \ and \ \beta_1$ we have to collect data on Y, X and U. Nonetheless, we can not get data on U as it is stochastic and can never be observed. Therefore, in order to estimate the parameters and make we should guess the values of $U_{i\,i}$ i.e., make some plausible assumptions about the shape and distribution of U.

2.4 Assumptions Underlying the Least Squares (OLS) method

The major objectives of regression analysis include estimation of and inferences about the population parameters $\beta_0 \ and \ \beta_1$ based on sample observations. For example, we would like to know how close the estimates, say, $\hat{\beta}_0 \ and \ \hat{\beta}_1$ are to the parameters $\beta_0 \ and \ \beta_1$, respectively. In other words, we want to know how close $\hat{Y}_i$ is to the true $E\left(Y/X_i\right)$. Hence, beyond specifying the functional form of the model, we have to make certain assumptions about the manner in which $Y_i's$ are generated.

Form equation 2.7, we have noted that Y_i depends on both $X_i \ and \ U_i$. Therefore, unless we are specific about how $X_i \ and \ U_i$ are generated, there is no way we can make any statistical inference about Y_i and about the estimates $\hat{\beta}_0 \ and \ \hat{\beta}_1$. Therefore, the assumptions made about X_i-variable(s) and the error term are very critical to make valid interpretation of the regression estimates.

The Gaussian or classical linear regression model is based on the following ten assumptions.

Assumption 1. Linear regression model

The regression model is linear in parameters, as show below

$$Y_i = \beta_0 + \beta_1 X_i + U_i$$

However, this assumption does not exclude models that are non–linear in variables such as $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i + ... + U_i$

Assumption 2: U_i is a random real variable with zero mean value.

This means that the values that U_i may assume in any particular period or instance depend on chance. It may assume positive, negative or zero values. Furthermore, given the value of X, the mean value of the random disturbance term U_i is zero. Technically, this means that the conditional mean value of U_i is zero. Symbolically,

$$E\left(U_i / X_i\right) = 0 \quad \text{.....2.8}$$

In a nutshell, this assumption implies that the factors not explicitly included in the model and therefore subsumed in U_i , do not systematically affect the mean value of Y; i.e., the positive U_i values cancel out the negative U_i values so that their average effect on Y is zero. Given $Y_i = \beta_0 + \beta_1 X_i + U_i$, this assumption leads to the fact that:

$$E\left(Y / X_i\right) = \beta_0 + \beta_1 X_i \cdot \text{.....2.9}$$

Assumption 3: The disturbance term U_i has a normal distribution

This assumption is an extension of assumption 3. It suggests that the values of U_i (for each X_i) have a normal distribution, which is bell–shaped and symmetrical about the zero mean of U_i

It is customary that there is gap between individual values of Y_i and the average value of Y_i associated with the fixed value of X (see figure 2.1). This gap is represented by U_i , which could be positive or negative. Furthermore, the values of U_i associated with a given value of X is symmetrically distributed around its mean value zero and it has a normal distribution.

Assumption 4: Homoscedasticity of U_i

The variance of U_i about its mean is constant at all values of X. In other words, for all values of X, the U_i values will show the same dispersion around their mean. Furthermore, given the value of X, the variance if U_i is the same (constant) for all observations.

Symbolically,

$$Var\left(U_i / X_i\right) = E\left[\left(U_i / X_i\right) - E\left(U_i / X_i\right)\right]^2$$

$$= E\left[U_i^2 / X_i\right] \quad \sin ce \quad E\left(U_i / X_i\right) = 0$$

$$\Rightarrow Var\left(U_i / X_i\right) = \sigma^2 \text{ , ……..2.10}$$

Note that the variance in equation 2.10 is constant..

Note: Assumption 4 implies that the values of Y corresponding to various values of X have constant variance.

Assumptions 5. No Autocorrelation between the values of the disturbance term, U_i

This means that the values of U_i associated with one value of X are independent of its values associated with other values of X. That means the covariance of any U_i with other U_j is equal to zero. In other words, the value that the disturbance term U assumes in any one period does not depend on its value in other periods. Shortly, given any two X values, $X_i \text{ and } X_j$ (where $i \neq j$), the correlation between any two U_i and U_j ($i \neq j$) is zero.

Symbolically,

$$Cov\left(U_i, U_j \;/\; X_i, X_j\right) = E\left\{\left[U_i - E\left(U_i\right)\right] /\; X_i\right\}\left\{\left[U_j - E\left(U_j\right)\right] / \; X_j\right\}$$

$$= E\left(U_i / X_i\right)\left(U_j / X_j\right) \quad ....\sin ceE\left(U_i / X_i\right) = E\left(U_j / X_j\right) = 0$$

$$\Rightarrow Cov\left(U_i, U_j \;/\; X_i, X_j\right) = 0 \quad .....\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 2.11$$

Assumption 6. The values of X are fixed in repeated samples.

This means that in taking a large number of samples on Y and X, the X values are the same on all samples but the values of Y do differ from sample to sample; i.e., X is assumed to be non stochastic.

For example, as we discussed in section 2.1 above, when we collect data on family income (Y) and consumption expenditure (C) from a certain community, keeping the value of Y fixed, say, at level ET Birr 1000 we may perhaps draw at random a family with monthly (or weekly) consumption expenditure (C) of, say ET Birr 600. Still keeping X at ET Birr 1000, we may draw at random another family with C value of ET Birr 800 and so on.

Assumption 7: U_i is independent of the explanatory variable (X).

This means that the disturbance U_i and the explanatory variable X are uncorrelated. The values of *U* and X do not tend to vary together; i.e., their covariance is zero.

Symbolically,

$$Cov\left(U_i, X_i\right) = E\left[U_i - E\left(U_i\right)\right]\left[X_i - E\left(X_i\right)\right]$$

$$= E\left[U_i\left[X_i - E\left(X_i\right)\right] - Since \;\; E\left(U_i\right) = 0\right]$$

$$= E\left[U_i X_i - U_i E\left(X_i\right)\right] = E\left[U_i X_i\right] - E\left[U_i E\left(X_i\right)\right]$$

Since $E(X_i)$ is non stochastic, $E(U_i\ E(X_i)) = E[U_i]\ E(X_i)$,

Thus,

$$Cov\ (U_i\ ,\ X_i) = E(U_i\ X_i) - E(U_i)\ E(X_i)$$

$$= E(u_i\ x_i)\ ,\ since\ E(u_i) = 0$$

$$\Rightarrow Cov\ (U_i\ ,\ X_i) = 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 2.12$$

Equation 2.12 is the implication of assumption 6.

Assumption 8: No perfect multicollinearity among explanatory variables

The explanatory variables are not perfectly correlated with each other. In other words, there is no perfect linear relationship among the explanatory variables. This assumption however, does not exclude non-linear relationships among the explanatory variables.

Assumption 9: Variability in X values

The X values in a given sample must not all be the same. Technically, $Var(X)$ must be a finite positive number. This means that x assumes different values in a given sample; but it assumes fixed values in a hypothetical repeated samples.

Assumption 9 is very critical since without this assumption it would be impossible to estimate the parameters and hence, regression analysis would fail. For example, if there is little variation in family income, we will not be able to explain much of the variation in the consumption expenditure of the families.

Activity

Dear readers, what do you think is the difference between assumptions 2 and 9?----------

----

Assumption 10: The regression model is correctly specified

This means that the mathematical form of the model is correctly specified and all important explanatory variables are included in it. In other words, there is no specification bias or error in the model used in empirical analysis. Unfortunately, in practice one rarely specifies the correct model. Hence, an econometrician would use some judgment in choosing the correct model, i.e., in determining the number of variables entering the model, assumptions about the distribution of the variables and functional form of the model s/he has to utilize some a priori or theoretical grounds.

2.5 The Distribution of the Dependent Variable, Y

So far we have determined the distribution of the explanatory variables and the stochastic term. In this section, we will determine the distribution of the dependent variable. Based on the assumptions we discussed so far about the distributions of X and U, we can establish that Y is normally distributed with:

1. Mean

$E(Y_i) = \beta_0 + \beta_1 X_i$ · and

2. Variance

$Var(Y_i) = Var(U_i) = \sigma_u^2$.

Proof:

1. By definition, the expected value of Y is equal to its mean value. Therefore, the mean of Y is given as

$E\ (Y_i\ )\ =\ E\ (\beta_0\ +\ \beta_1\ X_i\ .+\ U_i\ )$, since $Y_i = \beta_0\ +\ \beta_1\ X_i\ .+U_i$

$= E(\beta_0\ +\ \beta_1\ X_i).+E(U_i)$

We know that β_0 and β_1 are parameters and hence, they are constant. Furthermore, by Assumption 6, the values of X are a set of fixed numbers and by Assumption 2 $E\left(U_i \middle/ X_i\right)\ =\ 0$.

Therefore,

$E\ (Y_i\ )\ =\ E\ (\beta_0\ +\ \beta_1\ X_i\ ).+\ 0$

$=\beta_0\ +\ \beta_1\ X_i$,

Since Assumption 6 implies that $E(\beta_0\ +\ \beta_1\ X_i) = \beta_0\ +\ \beta_1\ X_i$

2. The variance of Y is given as

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

Substituting

$Y_i = \beta_0\ +\ \beta_1\ X_i\ .+U_i$ and $E(Y_i) = \beta_0\ +\ \beta_1\ X_i$

$Var\left(Y_i\right) = E(\beta_0\ +\ \beta_1\ X_i\ .+U_i - \beta_0\ -\ \beta_1\ X_i\ )^2$

$= E\left[U_i\right]^2$

$=\sigma_u^2$

Since by Assumption 4, the homoscedastic variance of U is given as $Var(U_i) = E(U_i)^2$

Therefore, we can conclude that the variance of Y is the same as the variance of the stochastic term.

3. The shape of the distribution of Y is normal

The distribution of Y is merely determined by the distribution of U . This is due to the fact that β_0 and β_1 are constants and hence, they do not affect the distribution of the dependent variable. Furthermore, by Assumption 6, the values of X are a set of fixed numbers and therefore, do not affect the distribution of Y . Thus, the distribution of Y is normal following the normality of the distribution of U .

2.6 The Least Squares Criterion and the OLS Estimates

Now assume that we have completed the work involved in the first four stages of the econometric methodology discussed in chapter one; namely we have specified the econometric model, stated its assumptions and collected the required data. Then the next step is the estimation of the model.

Recall the two variables PRF given in Equation 2.6.

$$Y_i = \beta_0 + \beta_1 X_i + U_i$$

This relationship holds for the population values of Y and X, so that we could obtain numerical values of β_0 and β_1 only if we could have all the conceivably possible values of Y, X and U, which form the population values of the variables.

Nonetheless, this is impossible in practice. Therefore, we have to obtain a sample of observed values of Y and X, specify the distribution of the U and try to get satisfactory estimates of the true parameters of the relationship. This is done by fitting a regression line (SRF) through the observations of the sample, which would be considered as an approximation to the true line.

In Equation 2.5 we have noted that SRF is given as:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 + \hat{U}_i$$

Therefore,

$$\hat{U}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 = Y_i - \hat{Y}_i \quad \text{.....2.13}$$

The question now is as to how to determine SRF. This means we are mainly interested in determining the SRF in such a way that the line is as close as possible to the actual Y. It is intuitively obvious that the smaller the deviations from the line, the better the fit of the line to the actual observations on Y, i.e., we have to choose the SRF in such a manner that the sum of the residuals $\sum \hat{U}_i = \sum \left(Y_i - \hat{Y}_i\right)$ is as small as possible. This approach is not an appropriate approach, no matter how intuitively appealing it may be. The reason for this is that the minimization of $\sum \hat{U}_i$ gives equal weight to different deviations; no matter how large or small the deviations may be; i.e., it attaches equal importance to all $U_i's$ no matter how close or how widely scattered the individual observation are from the SRF. Consequently, the algebraic sum of the $\hat{U}_i$ is small (even zero) although individual U_i are widely scattered about the SRF. This means that the minimization of the $\sum \hat{U}_i$ doesn't necessarily imply that individual deviations $\left(\hat{U}_i's\right)$ are minimized.

To ease this problem, we adopt the least squares criterion. This criterion requires the regression line to be drawn (its parameters to be chosen) in such a way as to minimize the sum of the squares of the deviations of the observations from it; i.e., it should minimize $\sum \hat{U}_i$ by squaring $\hat{U}_i$. Hence, this approach gives more weight to residuals with wider dispersion than those with closer dispersion around the line.

From Equation 2. 13, we know that $\hat{U}_i = Y_i - \hat{Y}_i$.

Thus, $\sum \hat{U}_i^2 = \sum \left(Y_i - \hat{Y}_i\right)^2 = \sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1\right)^2$ ……..2.14

It is clear from Equation 2.14 that

$$\sum \hat{U}_i^2 = f\left(\hat{\beta}_0 , \hat{\beta}_1\right) \text{ ……..2.15}$$

This implies that for any given set of data, choosing different values for $\hat{\beta}_0$ *and* $\hat{\beta}_1$ will give different $\hat{U}'s$ and hence different values of $\sum \hat{U}_i^2$. This implies that by assigning different values for $\hat{\beta}_0$ *and* $\hat{\beta}_1$ we will have different regression lines (SRFs) for the same sample.

For example, if $\hat{\beta}_0 = 1.5$ *and* $\hat{\beta}_1 = 1.3$, then SRF can be given as

SRF$_1$: $\hat{Y}_i = 1.5 + 1.3X_i$

If, on the other hand, $\hat{\beta}_0 = 3$ *and* $\hat{\beta}_1 = 1$, then SRF can be given as

SRF$_2$: $\hat{Y}_i = 3 + X_i$

Dear readers, which of these two lines do you think will give the best fit to the observed data? Alternatively, which set of $\hat{\beta}$ should be chosen? ----------
---

According to the least squares criterion the one that produces minimum values to $\sum \hat{U}_i^2$ must be chosen. This means, that to choose the best set of $\hat{\beta}'s$, we will assign many more values to $\hat{\beta}'s$ and see what may happen to $\sum U_i^2$. However,

in practice we may not have sufficient time and patience to conduct these trial and error processes. Therefore, we need to look for some short cuts. Fortunately, the method of least squares provides us such a short cut. The principle of least squares chooses $\hat{\beta}_0$ *and* $\hat{\beta}_1$ in such a way that, for a given sample, $\sum \hat{U}_i^2$ is as small as possible.

The mechanism of accomplishing this is straight forward by using differential calculus. Now recall from Equation 2.14 that for n pairs of observations on Y and X,

$$\sum_{i=1}^{n} \hat{U}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1\right)^2 \quad \text{…….. ……2.14*}$$

According to the principle of least squares, we have to minimize Equation 2.14 with respect to $\hat{\beta}_0$ *and* $\hat{\beta}_1$. At the minimum point of $\sum \hat{U}_i^2$, First Order Conditions(FOC) must be satisfied, which are given as:

$$FOC: i) \quad \frac{\partial \sum \hat{U}_i^2}{\partial \hat{\beta}_0} = 0 \quad \text{……2.16}$$

$$ii) \quad \frac{\partial \sum \hat{U}_i^2}{\partial \hat{\beta}_1} = 0 \quad \text{……2.17}$$

By applying the function of a function rule of differentiation to Equations 2.16 and 2.17 will yield:

$$1) \quad \frac{\partial \sum U_i^2}{\partial \hat{\beta}_0} = \frac{\partial \sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2}{\partial \hat{\beta}_0} = 0$$

$$= 2 \sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) . (-1) = 0$$

$$= -2 \sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) = 0$$

$$= \sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) = 0$$

$$= \sum Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum X_i = 0 \text{ ……..2.18}$$

By interchanging the places of terms in Equation 2.18, we obtain:

$$\Rightarrow \sum Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum X_i \text{ ……..2.19}$$

Performing differentiation to Equation 2.17 yields the following result:

$$\frac{\partial \sum \hat{U}_i^2}{\partial \hat{\beta}_1} = 0$$

$$\Rightarrow \frac{\partial \sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2}{\partial \hat{\beta}_1} = 0$$

$$\Rightarrow 2\sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)(-X_i) = 0$$

$$\Rightarrow -2\sum \left(Y_i X_i - \hat{\beta}_0 X_i - \beta_1 X_i^2\right) = 0$$

$$\Rightarrow \sum Y_i X_i - \hat{\beta}_0 \sum X_i - \hat{\beta}_1 \sum X_i^2 = 0$$

$$\Rightarrow \sum Y_i X_i = \hat{\beta}_0 \sum X_i + \hat{\beta}_1 \sum X_i^2 \text{ ..... ……..2.20}$$

Note that equations 2.19 and 2.20 are called **normal equations of OLS**.

Then, to develop formula to compute numerical values for $\hat{\beta}_0$ *and* $\hat{\beta}_1$, we solve these normal equations simultaneously by using **Cramer's rule**:

Let $A = \begin{bmatrix} \sum Y_i \\ \sum Y_i X_i \end{bmatrix}$

$$B = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}$$

$$C = \begin{bmatrix} \sum Y_i & \sum X_i \\ \sum Y_i X_i & \sum X_i^2 \end{bmatrix}$$

$$D = \begin{bmatrix} n & \sum Y_i \\ \sum X_i & \sum Y_i X_i \end{bmatrix}$$

Then,

$$\hat{\beta}_0 = \frac{\det er \min ant\ of\ C}{\det er \min ante\ of\ B}$$

$$= \frac{\left(\sum Y_i\right).\left(\sum X_i^2\right) - \sum x_i \sum Y_i x_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \quad \text{.....2.21}$$

And,

$$\hat{\beta}_1 = \frac{\det er \min ante\ of\ D}{\det er \min ante\ of\ B}$$

$$= \frac{n\sum Y_i x_i - \sum x_i \sum Y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \quad \text{.....2.22}$$

The estimators $\hat{\beta}_0$ *and* $\hat{\beta}_1$ obtained from this process are called **the least squares estimators**, since they are developed via the least squares principle.

In passing, note that equations 2.21 and 2.22 can be expressed in deviation forms as:

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \quad \text{..... .....2.23}$$

And $$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} \quad \text{.....2.24}$$

2.7 Estimation of a Function Whose Intercept is Zero

In some cases economic theory postulates relationships which have a zero constant intercept. For example, linear production functions of manufactured products should normally have zero intercept, since out put is zero when the factor in puts are zero.

In this event, we would estimate the function

$Y_i = \beta_0 + \beta_1 X_i + U_i$, imposing the restriction $\beta_0 = 0$.

In this case we want to fit the line $Y = \beta_0 + \beta_1 X + U$ subject to $\beta_0 = 0$.This is a restricted minimization problem. Thus, we minimize:

$$\sum \hat{U}_i^2 = \sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2$$

subject to

$$\hat{\beta}_0 = 0$$

To solve this problem, we form a composite function called Lagrange function as follows

$$L = \sum \left(Y - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2 - \lambda \hat{\beta}_0 \quad \text{.....2.25}$$

Where λ is called the Lagrange multiplier.

The values of $\hat{\beta}_0$ $and \hat{\beta}_1$ that maximize equation 2.25 can be obtained by taking the partial differential of equation 2.25, which is given as follows;

1) $$\frac{\partial L}{\partial \hat{\beta}_0} = 2 \sum \left(Y - \hat{\beta}_0 - \hat{\beta}_1 X\right)(-1) - \lambda = 0$$

$$= -2 \sum \left(Y - \hat{\beta}_0 - \hat{\beta}_1 X\right) - \lambda = 0 \quad \text{.....2.26}$$

2) $\frac{\partial L}{\partial \hat{\beta}_1} = 2\sum\left(Y - \hat{\beta}_0 - \hat{\beta}_1 X\right)(-X) = 0$

$= -2\sum\left(Y - \hat{\beta}_0 - \hat{\beta}_1 X\right)(X) = 0$ ……..2.27

3) $\frac{\partial L}{\partial \lambda} = -\hat{\beta}_0 = 0$

$\Rightarrow \hat{\beta}_0 = 0$ ……..2.28

Substituting equation 2.28 into equation 2.27, we get

$\Rightarrow 2\sum\left(Y - \hat{\beta}_1 X\right)(x) = 0$

$\Rightarrow \sum YX - \hat{\beta}_1 \sum X^2 = 0$

$\Rightarrow \sum YX = \hat{\beta}_1 \sum X^2$

$\Rightarrow \hat{\beta}_1 = \frac{\sum YX}{\sum X^2}$ ……2.28*

Note that the difference between equations 2.24 and 2.28 is that the former is in deviation form while the latter involves actual values.*

2.8 Functional Forms of Regression Models

So far we primarily dealt with models that are linear in parameter and in variables. Now, in this section we will consider some commonly used regression models that may be non–linear in variables but are linear in the parameters or can be made so by suitable transformations of the variables.

2.8.1 The Leg – linear model

To illustrate this model, assume that you are given the so called exponential regression model given as:

$$Y_i = \beta_0 \; X_i^{B_1} \; e^{Ui} \text{ ……………………. …………………..2.29}$$

Where e = 2.718, which is constant.

Then, taking the natural logarithm of equation 2.29, will yield

$$\ln Y_i = \ln \beta_0 + \beta_1 \; la \; X_i \; U_i \text{ ………………………………2.30}$$

Note: Equation 2.30 is the result of the following properties:

i) $\ln(AB) = \ln A + \ln B$

ii) $\ln A/B = \ln A - \ln B$

iii) $\ln \left(A^K\right) = k \ln A$

Then, letting $\ln \beta_0 = \alpha$, equation 2.30 can be rewritten as:

$$\ln Y_i = \alpha + \beta_1 \ln X_i + u_i \text{ ……………………………………….2.31}$$

From equation 2.31, it is clear that the model is linear in parameters α *and* β_1 and linear in the logarithms of the variables Y and X. Hence, it can be estimated by OLS regression, which is suitable only for linear models. It is this linearity that makes the model to be called as **log-log or double log or log-linear model**.

This model is very popular in applied work. This is mainly due the fact that the slope coefficient (e.g. β_1 in equation 2.31) measures the elasticity of Y with respect to X. i.e., the percentage change in Y for a give (small) percentage change in X.

Note: The log-linear model assumes that the elasticity coefficient between Y and X, remains constant throughout. In other words it shows that the elasticity remains the same no matter at which value of X we measure the elasticity. Due to this reasons the model is also called **constant elasticity model**.

Activity

Let Y represent the quantity supplied of a commodity and X represent the price of the commodity. Then, if you use equation 2.29 to model the relationship between Y and X, interpret the coefficient of X?----------

---

2.8.2 Semi-log Models

i) Log–Lin Models

Some times we may be interested in finding out the rate of growth of certain economic variables. In this case we will use **Log-Lin** models. For example, if we want to find out the growth rate of personal consumption expenditure on services, this model will be applied.

To see how these models develop, let Y_t denote real expenditure on services at time t and Y_0 the initial value of the expenditure on service. As you may recall the formula of compound growth rate is given as

$$Y_t = Y_0 (1 + r)^t \quad \text{.....2.32}$$

Where r is the compound rate of growth of Y.

Taking the natural logarithm of equation 2.32 will give:

$$\ln Y_t = \ln Y_0 + t \ln (1 + r) \quad \text{.....2.33}$$

If we let $\ln Y_0 = \beta_0$ and $\ln(1 + r) = \beta_1$, we will have

$$\ln Y_t = \beta_0 + \beta_1 t \quad \text{.....2.34}$$

To make equation 2.34 stochastic and develop econometric model, we add the disturbance term. Then it becomes,

$$\ln Y_t = \beta_0 + \beta_1 t + U_t \quad \text{.....2.35}$$

As it can be seen from the above equation, the model is linear in the parameters β_0 *and* β_1. The only difference in this model is that the regressand is $\ln Y$ and the regressor is time, t. Models like equation 2.35 are called **Semi-Log models** since it is only one variable that appears in logarithmic form.

Semi- log models are called **Log-Lin** models if the regressand is in logarithmic form. In Log-Lin models, the slope coefficient measures a constant proportional or relative change in Y for a given absolute change in the values of the regressor. That means in equation 2.35,

$$\beta_2 = \frac{\text{Re}\,lative\ \ change\ \ in\ \ a\, regressand}{Absolute\ \ change\ \ in\ a\ regressor} \quad \text{.....2.36}$$

If we multiply the relative change in Y by 100, equation 2.36 will give the percentage change or the growth rate in Y for an absolute change in the explanatory variables i.e., 100 times β_2 gives the growth rate in Y and some times it is called the **semi-elasticity of Y** with respect to the explanatory variable.

ii) The Lin–Log model

In this case we are interested in finding out the absolute change in Y for a percentage change in explanatory variable, X. This model can be written as

$$Y_i = \beta_0 + \beta_1 \ln x_i + U_i \quad \text{.....2.37}$$

These types of models are known as Lin-Log models. In this case, β_2 is given as:

$$\beta_2 = \frac{Change\ in\ Y}{Change\ in\ \ln X}$$

$$= \frac{Change\ in\ Y}{\text{Re}lative\ change\ in\ X.}$$

$$= \frac{\Delta Y}{\Delta X / X}$$

Equivalently,

$$\Delta Y = \beta_2 \left(\Delta X / X\right) \text{ ....2.38}$$

This equation states that the absolute change in $Y\ (i.e. \Delta Y)$ is equal to slope times relative change in X. If the terms in the right had side of equation 2.38 is multiplied by 100, then the equation gives the absolute change in Y for a percentage change in X. Thus, if $\Delta X / X$ changes by 0.01 units (1 percent), the absolute change in Y is 0.01 β_2. That means, for example, if we find β_2 = 500, then the absolute change in Y is (0.01) (500) = 5.

Thus, it is noteworthy that when equation 2.38 is estimated by OLS, the value of the estimated slope coefficient must be multiplied by 0.01; otherwise your interpretation will be misleading.

Generally**,** while the choice of a particular functional form may depend on the underlying theory, it is a good practice use a model that enables us to find out the rate of change of the dependent variable with respect to the explanatory variable as well as the elasticity of the regressand with respect to the explanatory variables.

Exercises

1. Why do we need regression analysis?
2. What is the difference between regression and correlation analysis?
3. What is the difference between the population and sample regression functions?
4. What is the role of the stochastic error term U_i in regression analysis?
5. Given the following two models:

Model I: $Y_i = \beta_0 + \beta_1 X_i + U_i$

Model II: $Y_i = \alpha_0 + \alpha_1 (X_i - \overline{X}) + U_i$

   i. Find the estimators of β_0 and α_0, and their variances. Explain the differences between these estimators, if any.
   ii. Find the estimators of β_1 and α_1, and their variances. Explain the differences, if any.
   iii. What is the advantage, if any, of model II over model I?

6. The following results are obtained from a sample of 11 observations on the dependent variable (Y) and explanatory variable (X)

$\overline{X} = 520$

$\overline{Y} = 220$ $\quad\quad$ $\sum X_i Y_i = 1290$

$\sum X_i^2 = 3100$ $\quad\quad$ $\sum Y_i^2 = 539{,}500$

Based on the given information

a. Estimate the coefficients of regression line of Y on X
b. Interpret the coefficients of your model

7. Suppose in question #6 above, on rechecking the data it was found that two pairs of observations were erroneously recorded as

Y	*X*		*Y*	*X*
90	120	instead of	80	110
140	220		150	210

i. Find the OLS estimates of the coefficients of the regression line of Y on X.
ii. Explain the effect of the data recording error on the estimates of the coefficients of the regression model in question #6.

CHAPTER 3

PROPERTIES OF OLS ESTIMATORS

Overview

In chapter 2, we have learnt how to obtain the OLS estimates of population parameters based on sample observations of the dependant and explanatory variables. Furthermore, the chapter dealt with different functional forms of economic relationships and the interpretation of OLS estimates obtained from them.

In the current chapter we will establish the statistical properties of the estimates obtained under chapter 2. In addition, it emphasizes on the procedures of verifying whether OLS estimates do possess the best statistical properties of estimates or not.

At the end of this chapter students will be able to:

- Understand the statistical properties of good estimates
- Show that OLS estimates are Best, Linear and Unbiased (BLU) estimators of the true parameters
- Drive the formula of the means, variances and standard errors of OLS estimates
- Develop a formula to estimate the variance of the unobservable stochastic term based on the sample observations of the dependent and explanatory variables.

3.1 The Mean, Variance and Standard Errors of OLS Estimates and the Error Term

The derivation of a formula from which the means and variances of the estimates would be obtained is a sin qua none for establishing the statistical properties of OLS estimates. Hence, in this part, we will try to drive formula for the means and variances of the OLS estimates.

3.1.1 The Mean of $\hat{\beta}_1$

Assume that we draw repeated samples of size n from the population of Y and X, and for each sample we estimate the parameters $\hat{\beta}_0$ *and* $\hat{\beta}_1$. If all the possible samples are taken, then the mean value of $\hat{\beta}_1$ will be its expected value, i.e. mean of $\left(\hat{\beta}_1\right) = E\left(\hat{\beta}_1\right)$. To find the value of the mean in terms of the observations of our sample of Y and X we work as follows:

Recall that from Equation 2. 24,

$$\hat{\beta}_1 = \frac{\sum x_i\, y_i}{\sum x_i^2} \quad \text{.....2.24*}$$

Substituting $Y_i = Y_i - \overline{Y}$ and $x_{ii} = X - \overline{X}$ in the above equation, we obtain

$$\hat{\beta}_1 = \frac{\sum x_i \left(Y_i - \overline{Y}\right)}{\sum x_i^2} = \frac{\sum x_i\, Y_i}{\sum x_i^2} - \frac{\overline{Y}\sum x_i}{\sum x_i^2}$$

But we know that $\sum x_i = 0$ due to the properties of arithmetic mean.

Therefore,

$$\Rightarrow \hat{\beta}_1 = \frac{\sum x_i Y_i}{\sum x_i^2} = \sum\left[\frac{x_i}{\sum x_i^2}\right] Y_i \quad \text{.....3.1}$$

Nonetheless, in Chapter 2 we have discussed that by Assumption 6 of the linear regression models, the values of X are a set of fixed numbers, which do not change from sample to sample. Consequently, the ratio $\frac{x_i}{\sum x_i^2}$ in equation 3.1 will be constant from sample to sample. Then, if we denote this ratio by k_i we may write the estimate $\hat{\beta}_1$ in Equation 3.1 as follows:

$$\hat{\beta}_1 = \sum k_i Y_i \quad \text{.....3.2}$$

where, $k_i = \frac{x_i}{\sum x_i^2}$

But in equation 2.3b, Y_i was defined as $Y_i = \beta_0 + \beta_1 X_i + U_i$. Hence, equation 3.2 can be rewritten as:

$$\hat{\beta}_1 = \sum k_i \left(\beta_0 + \beta X_i + U_i\right)$$

$$= \beta_0 \sum k_i + \beta_1 \sum k_i X_i + \sum k_i U_i \quad \text{.....3.3}$$

But $\sum k_i = \sum \left[\frac{x_i}{\sum x_i^2}\right]$,

$$= \frac{\sum x_i}{\sum x_i^2} = \frac{0}{\sum x_i^2} = 0, \text{ since } \sum x_i = 0$$

Therefore, equation 3.3 can be written as

$$\hat{\beta}_1 = \beta_1 \sum k_i X_i + \sum k_i U \quad \text{.....3.4}$$

Then, since k_i is constant, $\sum k_i X_i$ in equation 3.4 can be given as:

$$\sum k_i X_i = \sum X_i \left(\frac{x_i}{\sum x_i^2}\right)$$

$$= \frac{\sum X_i x_i}{\sum x_i^2}$$

$$= \frac{\sum \left(X_i - \overline{X}\right) X_i}{\sum x_i^2}$$

$$= \frac{\sum X_i^2 - \overline{X} \sum X_i}{\sum x_i^2}$$

$$= \frac{\sum X_i x_i}{\sum x^2{}_i} \quad \text{.....3.5}$$

And, $\sum x_i^2$ in the denominator of equation 3.5 can be expanded as:

$$\sum x_i^2 = \sum \left(X_i - \overline{X}\right)^2$$

$$= \sum (X_i^2 - 2\overline{X} X_i + \overline{X}^2)$$

$$= \sum X_i^2 - 2\overline{X} \sum X_i + \frac{n(\sum X_i)^2}{n^2}$$

$$= \sum X_i^2 - 2 \frac{\left(\sum X_i\right)^2}{n} + \frac{\left(\sum X_i\right)^2}{n}$$

$$= \sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}$$

$$= \frac{n \sum X_i^2 - \left(\sum X_i\right)^2}{n} \quad \text{.....3.6}$$

Thus, substituting equation 3.6 into equation 3.5, yields:

$$\sum k_i X_i = \frac{\sum X_i^2 - \overline{X}\sum X_i}{\frac{n\sum X_i^2 - \left(\sum X_i\right)^2}{n}}$$

$$= \frac{\sum X_i^2 - \left(\frac{\sum X_i}{n}\right)\left(\sum X_i\right)}{\frac{n\sum X_i^2 - \left(\sum X_i\right)^2}{n}}$$

$$= \left(\frac{n\sum X_i^2 - \left(\sum X_i\right)^2}{n}\right)\left(\frac{n}{n\sum X_i^2 - \left(\sum X_i\right)^2}\right)$$

$$= \underline{1} \text{.....3.7}$$

Hence, substituting equation 3.7 into equation 3.4, we get

$$\hat{\beta}_1 = \beta_1 (1) + \sum k_i U_i \text{ .....3.8}$$

Then, taking the expected value of equation 3.8 and noting that the values of X are fixed in repeated samples:

$$E\left(\hat{\beta}_1\right) = E(\beta_1) + E\left[\frac{\sum x_i u_i}{\sum x_i^2}\right]$$

$$= \beta_1 + \frac{\sum x_i E(u_i)}{\sum x_i^2}$$

$= \beta_1$, since $E(u_i) = 0$ by Assumption 2 of linear regression models

$$\Rightarrow E\left(\hat{\beta}_1\right) = \beta_1 \text{.....3.9}$$

Therefore, in general the mean of the ordinary least squares estimate $\hat{\beta}_1$ is equal to the true value of the population parameter β_1 .

3.1.2 The Variance of $\hat{\beta}_1$

We have shown in equation 3.2 that

$$\hat{\beta}_1 = \frac{\sum x_i Y_i}{\sum x_i^2} = \sum k_i Y_i \quad \text{.....3.10}$$

Recall that, by assumption, the k_i's are constant weights in hypothetical repeated sampling. Hence, taking the variance of equation 3.10 gives

$$Var\left(\hat{\beta}_1\right) = \sum k_i^2 \; Var\left(Y_i\right) \text{.....3.11}$$

But we have shown in chapter 2 that $Var(Y_i) = \sigma_u^2$. Therefore, equation 3.11 becomes,

$$Var\left(\hat{\beta}_1\right) = \sum k_i^2 . \sigma_u^2 \text{.....3.12}$$

$$= \sigma_u^2 \sum k_i^2 \text{ , since the variance of } U_i \text{ is homoscedastic by}$$

Assumption 5 of classical linear regression models.

Substituting $k_i = \frac{x_i}{\sum x_i^2}$ in equation 3.12 yields:

$$Var\left(\hat{\beta}_1\right) = \sigma_u^2 \sum\left(\frac{x_i^2}{\left(\sum x_i^2\right)^2}\right)$$

$$= \frac{\sigma_u^2}{\left(\sum x_i^2\right)^2} \sum x_i^2$$

$$\Rightarrow Var\left(\hat{\beta}_1\right) = \underline{\underline{\frac{\sigma_u^2}{\sum x_i^2}}} \text{ .....3.13}$$

3.1.3 The mean of $\hat{\beta}_0$

Recall that from equation 2.23 that

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

But, from equation 3.10

$$\hat{\beta}_1 = \sum k_i Y_i .$$

Hence:

$$\hat{\beta}_0 = \bar{Y} - \bar{X}\sum k_i Y_i$$

$$= \frac{\sum Y_i}{n} - \bar{X}\sum k_i Y_i$$

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

$$\hat{\beta}_0 = \sum\left(\frac{1}{n} - \bar{X} k_i\right) Y_i \quad \text{.....3.14}$$

Taking the expected value of equation 3.14 gives

$$E\left(\hat{\beta}_0\right) = \sum\left(\frac{1}{n} - \bar{X} k_i\right) .E\left(Y_i\right)$$ ..... $\bar{x}$, n *and* k_i are constants.

Now, recall that $E(Y_i) = \beta_0 + \beta_1 X_i$

$$\Rightarrow E\left(\hat{\beta}_0\right) = \sum\left(\frac{1}{n} - \bar{X} k_i\right)\left(\beta_0 + \beta_1 X_i\right)$$

$$= \sum\left[\frac{\beta_0}{n} + \frac{\beta_1 X_i}{n} - \beta_0 \bar{X} k_i - \beta_1 \bar{X} k_i X_i\right]$$

$$= \frac{n\,\beta_0}{n} + \frac{\beta_1 \sum X_i}{n} - \beta_0\,\overline{X}\sum k_i - \beta_1\,\overline{X}\sum k_i X_i$$

However, we have already shown in section 3.1.1 that $\sum k_i = 0$ and $\sum k_i X_i = 1$. Thus,

$$E\left(\hat{\beta}_0\right) = \beta_0 + \beta_1\,\overline{X} - \beta_1\,\overline{X}$$

$$= \beta_0$$

$$\Rightarrow \underline{\underline{E\left(\beta_0\right) = \beta_0}} \quad \text{.....3.15}$$

This implies that the mean of the ordinary least squares estimate $\hat{\beta}_0$ is equal to the true value of the population parameter β_0 .

3.1.4 The Variance of $\hat{\beta}_0$

Recall form equation 3.14 that

$$\hat{\beta}_0 = \sum\left(\frac{1}{n} - \overline{X}\,k_i\right) Y_i$$

Therefore, taking the variance of this equation will give you:

$$Var\left(\hat{\beta}_0\right) = \sum\left(\frac{1}{n} - \overline{X}\,k_i\right)^2 Var\left(Y_i\right), \text{ since } n, \overline{X} \text{ and } k_i \text{ are constants.}$$

$$= \sigma_u^2 \sum\left(\frac{1}{n^2} - \frac{2\,\bar{x}\,k_i}{n} + \bar{x}^2\,k_i^{\,2}\right), \text{ since } Var(Y_i) = \sigma_u^{\,2}$$

$$= \sigma_u^2 \left\{\frac{n}{n^2} - \frac{2\,\bar{x}\sum k_i}{n} + \bar{x}^2 \sum k_i^{\,2}\right\}$$

$$= \sigma_u^2 \left\{\frac{1}{n} + \frac{\bar{x}^2}{\sum x_i^2}\right\}, \text{ since } \sum k_i^{\,2} = \frac{1}{\sum x^2{}_i}.$$

$$= \sigma_u^2 \left(\frac{\sum x_i^2 + n\,\bar{x}^2}{n \sum x_i^2} \right)$$

Thus,

$$Var\left(\hat{\beta}_0\right) = \sigma_u^2 \left(\frac{\sum x_i^2 + n\,\bar{x}^2}{n \sum x_i^2} \right) \text{.....3.16}$$

3.1.5 The Variance of the Random Variable U

As we have shown above, the formula of the variances of the OLS estimates, $\hat{\beta}_0$ *and* $\hat{\beta}_1$ involve the variance of the random term U_i, i.e., σ_u^2. However, the true variance of U_i can not be computed since the values of U_i are not observable. Thus, a formula to estimate σ_u^2 should be developed. To do this, we use the device of repeated (hypothetical) sampling, through which we obtain all possible samples of size n; compute a regression line for each sample, and the, find the values of the residuals from each regression. Thus, the variance of the residuals defined as $e_i = Y_i - \hat{Y}_i$ * is defined as the expected value of the squared differences of $e_i's$ from their mean.

Symbolically,

$$Var\ (e_i) = E\left[e_i - E\ (e_i)\right]^2$$

Since, by definition $E\ (e) = 0$

$$Var\ (e_i) = E\left[e_i\right]^2 \text{ .....3.17}$$

---

*In this module e_i and $\hat{U}_i$ are interchangeably used to represent the estimated values of the stochastic term U_i

Now, the problem is to express this variance in terms of the sample observations of X and Y. To do this let's proceed as follows:

Recall that

$$e_i = Y_i - \hat{Y}_i \quad \text{.....3.18}$$

However, the difference between Y_i and its mean value $\overline{Y}$ is defined as:

$$y_i = Y_i - \overline{Y} \quad \text{.....3.19}$$

$$\Rightarrow Y_i = y_i + \overline{Y} \quad \text{.....3.19*}$$

Where $Y_i = \beta_0 + \beta_1 X_i + U_i$ and

$$\overline{Y} = \beta_0 + \beta_1 X_i + \overline{U}$$

Note: *While $E(U) = 0$ in repeated samples (that is in taking a very large no of sample we expect U_i to have a mean value of zero by assumption), in any particular single sample $\overline{U}$ is not necessarily zero*.

In addition, the difference between $\hat{Y}_i$ and the mean value $\overline{Y}$ is defined as:

$$\hat{y}_i = \hat{Y}_i - \overline{Y} \quad \text{.....3.20}$$

$$\Rightarrow \hat{Y}_i = \hat{y}_i + \overline{Y} \quad \text{.....3.20*}$$

Hence, substitution of equations 3.19* and 3.20* into equation 3.18 yields:

$$e_i = (y_i + \overline{Y}) - (\hat{y}_i + \overline{Y})$$

$$= y_i - \hat{y}_i \quad \text{.....3.21}$$

But we have shown in equation 3.19 that $y_i = Y_i - \overline{Y}$. Then, substituting the values of Y_i and $\overline{Y}$ defined under equation 3.18 in to equation 3.19, we get:

$$y_i = \beta_0 + \beta_1 X_i + U_i - \left[\beta_0 + \beta_1 \overline{X} + \overline{U}\right]$$

$$= \beta_0 + \beta_1 X_i + U_i - \beta_0 - \beta_1 \overline{X - U}$$

$$= \beta_1 \left(X_i - \overline{X}\right) + \left(U_i - \overline{U}\right)$$

$$= \beta_1 x_i + \left(U_i - \overline{U}\right), \text{ where } x_i = \left(X_i - \overline{X}\right)$$

$$\Rightarrow y_i = \beta_1 x_i + \left(U_i - \overline{U}\right) \text{ ……..3.21(a)}$$

Furthermore, from the definition under equation 3.20 we know that

$$\hat{y}_i = \hat{Y}_i - \overline{Y}$$

Thus,

$$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 X_i - (\hat{\beta}_0 + \hat{\beta}_1 \overline{X})$$

$$= \hat{\beta}_0 + \hat{\beta}_1 X_i - \hat{\beta}_0 - \hat{\beta}_1 \overline{X}$$

$$= \hat{\beta}_1 (X_i - \overline{X})$$

$$= \hat{\beta}_1 x_i$$

$$\Rightarrow \hat{y}_1 = \hat{\beta}_1 x_i \text{ ……..3.21(b)}$$

Therefore, substituting equations 3.21(a) and 3.21(b) into equation 3.21, we get:

$$e_i = y_i - \hat{y}_i = \beta_1 x_i + \left(U_i - \overline{U}\right) - \hat{\beta}_1 x_i$$

$$= \beta_1 x_i - \hat{\beta}_1 x_i + \left(U_i - \overline{U}\right)$$

$$= x_i (\beta_1 - \hat{\beta}_1) + \left(U_i - \overline{U}\right)$$

$$= \left(U_i - \overline{U}\right) + (\beta_1 - \hat{\beta}_1) x_i$$

$$= \left(U_i - \overline{U}\right) - (\hat{\beta}_1 - \beta_1) x_i$$

$$\Rightarrow e_i = \left(U_i - \overline{U}\right) - (\hat{\beta}_1 - \beta_1) x_i \text{ ……..3.22}$$

Note that the $e_i's$ in equation 3.22 are the regression residuals that define the deviations of actual observations from the estimated values. Then, taking the summation of the squares of the residuals in equation 3.22 over the n sample values yields:

$$\sum e_i^2 = \sum \left\{ \left(U_i - \overline{U}\right) - \left(\hat{\beta}_1 - \beta_1\right)x_i \right\}^2$$

$$= \sum \left\{ \left(U_i - \overline{U}\right)^2 + \left(\hat{\beta}_1 - \beta_1\right)^2 x_i^2 - 2\left(U_i - \overline{U}\right)\left(\hat{\beta}_1 - \beta_1\right)x_i \right\}$$

$$\Rightarrow \sum e_i^2 = \sum \left(U_i - \overline{U}\right)^2 + \left(\hat{\beta}_1 - \beta_1\right)^2 \sum x_i^2 - 2\left(\hat{\beta}_1 - \beta_1\right) \sum x_i \left(U_i - \overline{U}\right) \ldots\ldots 3.33$$

Thus, taking the expected values of equation 3.33, we get:

$$E\left(\sum e_i^2\right) = \underbrace{E\left[\sum \left(U_i - \overline{U}\right)^2\right]}_{A} + \underbrace{E\left[\left(\hat{\beta}_1 - \beta_1\right)^2 \sum x_i^2\right]}_{B} - \underbrace{2E\left[\left(\hat{\beta}_1 - \beta_1\right)\left(\sum x_i \left(U_i - \overline{U}\right)\right)\right]}_{C}$$

To avoid arithmetic complications, let us simplify the terms on the right hand side of the above equation one by one following the denotation given beneath individual terms in the equation.

$$\Rightarrow E(\sum e_i^2) = A + B - 2C \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.34$$

We know from the above denotation that $A = E\left[\sum \left(U_i - \overline{U}\right)^2\right]$

$$\Rightarrow A = E\left[\sum \left(U_i^2 + \overline{U}^2 - 2U_i\overline{U}\right)\right]$$

$$= E\left[\sum U_i^2 + n\overline{U}^2 - 2\overline{U}\sum U_i\right]$$

$$= E\left[\sum\left(U_i^2\right) + n\,\overline{U}\,\frac{\sum U_i}{n} - 2\,\overline{U}\,\sum U_i\right]$$

$$= E\left[\sum\left(U_i\right)^2 + \underbrace{\overline{U}\,\sum U_i - 2\,\overline{U}\,\sum U_i}\right]$$

$$= \left[\sum\left(U_i\right)^2 - \overline{U}\,\sum U_i\right]$$

$$= E\left\{\sum\left(U_i\right)^2 - \frac{\sum U_i}{n}\,.\,\sum U_i\right\}$$

$$= E\left\{\sum\,\left(U_i\right)^2 - \frac{\left(\sum U_i\right)^2}{n}\right\}$$

$$= E\left\{\sum\left(U_i\right)^2 - \frac{1}{n}\left(\sum U_i\right)^2\right\}$$

$$= \sum E\left(U_i\right)^2 - \frac{1}{n}\,E\left(\sum U_i\right)^2$$

Since $E\left(u_i\right)^2 = \sigma_u^{\,2}$, this equation can be expressed as

$$A = \sum\sigma_u^2 - \frac{1}{n}\,E\left[U_1 + U_2 + U_3 + ... + U_n\right]^2$$

$$= n\sigma_u^2 - \frac{1}{n}E\left[U_1^2 + U_2^2 + ... + U_n^2 + 2U_1U_2 + 2U_1U_3.... + 2U_1U_n + 2U_2U_1 + 2U_2U_{3+.....}\right]$$

$$= n\,\sigma_u^2 - \frac{1}{n}\,E\left\{\left(\sum U_i^2\right) + 2\,\sum_{i\neq j}U_i\,U_j\right\}$$

$$= n\,\sigma_u^2 - \frac{1}{n}\sum E(U_i)^2 + 2\sum_{i \neq j} E(U_i U_j)$$

We know from the assumptions of linear regression models that the covariance between different values of the error term, U_i is zero. Hence, the expected value of the cross product of the values of U_i is zero, i.e. $E(U_i U_j) = 0$ for $i \neq J$.

$$\Rightarrow A = n\,\sigma_u^2 - \frac{1}{n}\sum \sigma_u^2 + 0$$

$$= n\,\sigma_u^2 - \frac{1}{n}\, n\,.\,\sigma_u^2$$

$$= n\,\sigma_u^2 - \sigma_u^2$$

$$= \sigma_u^2 (n - 1)$$

$$\Rightarrow A = \sigma_u^2 (n - 1) \quad \text{.....3.35}$$

Now, let us come to the other term, B in equation 3.34.

Recall that we defined B as

$$B = E\left(\hat{\beta}_1 - \beta_1\right)^2 \sum x_i^2$$

$$= \sum x_i^2 \; E\left[\hat{\beta}_1 - \beta_1\right]^2 \quad \text{.....3.36}$$

However, we know that the variance of $\hat{\beta}_1$ can be written as:

$$Var\left(\hat{\beta}_1\right) = E\left[\hat{\beta}_1 - E\left(\hat{\beta}_1\right)\right]^2 \quad \text{.....3.37}$$

Recall from equation 3.9 that $E\left(\hat{\beta}_1\right) = \beta_1$. Thus, equation 3.36 can be rewritten as

$$Var\left(\hat{\beta}_1\right) = E\left[\hat{\beta}_1 - \beta_1\right]^2 \quad \text{.....3.37(a)}$$

But we have already shown in equation 3.13 that

$$Var\left(\hat{\beta}_1\right) = \sigma_u^2 \Big/ \sum x_i^2 \quad \text{.....3.37 (b)}$$

From the comparison of equations 3.37 (a) and (b), we see that

$$E(\hat{\beta}_1 - \beta_1) = \sigma_u^2 \Big/ \sum x_i^2 \quad \text{.....3.38}$$

Substituting equation 3.38 into equation 3.36, yields

$$B = \sigma_u^2 \Big/ \sum x_i^2 \left[\sum x_i^2\right]$$

$$= \underline{\underline{\sigma_u^2}} \quad \text{.....3.39}$$

Finally, expanding the last term, we get

$$C = E\left\{\left(\hat{\beta}_1 - \beta_1\right)\sum x_i \left(U_i - \overline{U}\right)\right\}$$

$$= E\left\{\left(\hat{\beta}_1 - \beta_1\right)\sum (x_i U_i - x_i \overline{U}\right\}$$

$$= E\left\{\left(\hat{\beta}_1 - \beta_1\right)\left[\sum (x_i U_i - \overline{U}\sum x_i\right]\right\}$$

However, we know that $\sum x_i = 0$.

Thus,

$$C = E\left\{\left(\hat{\beta}_1 - \beta_1\right)\left[\sum (x_i U_i\right]\right\} \quad \text{.....3.40}$$

But in equation 3.8 we have shown that

$$\hat{\beta}_1 = \beta_1 + \sum k_i U_i$$

$$\Rightarrow \hat{\beta}_1 - \beta_1 = \sum k_i U_i \quad \text{.....3.41}$$

Thus, substituting equation 3.41 into equation 3.40, we get

$$C = E\left\{ \left(\sum k_i U_i \right)\left(\sum x_i U_i \right) \right\}$$

$$= E\left\{ \left(\frac{\sum x_i U_i}{\sum x_i^2} \right) \sum x_i U_i \right\} \text{since } k_i = \frac{x_i}{\sum x_i^2}$$

$$= E\left\{ \frac{\left(\sum x_i U_i \right)^2}{\sum x_i^2} \right\}$$

$$= E\left\{ \frac{x_1^2 U_1^2 + x_2^2 U_2^2 + x_3^2 U_2^2 + ... + x_n^2 U_n^2 + 2 \sum x_i x_j U_i U_j}{\sum x_i^2} \right\}$$

$$= E\left\{ \frac{\sum x_i^2 U_i^2 + 2 \sum_{i \neq j} x_i x_j U_i U_j}{\sum x_i^2} \right\}$$

$$= \frac{\sum x_i^2 E\left(U_i\right)^2 + 2 \sum \{ x_i x_j E\left(U_i U_j\right) \}}{\sum x_i^2}$$

$$= \frac{\sum x_i^2 \sigma_u^2}{\sum x_i^2} \quad \text{since, } E(U_i U_j) = 0 \text{ for } i \neq J .$$

$$= \sigma_u^2$$

$$\Rightarrow C = \sigma_u^2 \text{ ……..3.42}$$

Then, substituting equations 3.42, 3.39 and 3.35, into 3.34, we get

$$E(\sum e_i^2) = \sigma_u^2 (n - 1) + \sigma_u^2 - 2\sigma_u^2$$

$$= \sigma_u^2 (n - 1) - \sigma_u^2$$

$$= \sigma_u^2 (n - 2)$$

$$\Rightarrow E(\sum e_i^2) = \sigma_u^2 (n - 2) \text{ ……..3.43}$$

Dividing both sides of equation 3.43 by $n-2$ gives

$$\sigma_u^2 = \frac{E(\sum e^2{}_i)}{n-2}$$

$$= E(\frac{\sum e^2{}_i}{n-2}) \quad \text{.....3.44}$$

Dear readers, from your knowledge of **Statistics for Economists (Econ 321)** you may recall that if an estimator, say $\hat{\theta}$, is an unbiased estimator of the true parameter, say θ, its expected value will be equal to the true parameter, θ (for the proof see section 3.2.2).

Hence, from equation 3.44 we can conclude that $\hat{\sigma}_u = \frac{\sum e^2}{n-2}$ is an unbiased estimator of the variance of U_i , σ_u^2.

3.1.6 Standard errors of least squares estimates

In chapter 2, we have noted that the least square estimates are functions of sample data and consequently the estimates change from sample to sample as the data are likely to change from sample to sample.

Hence, we have to develop a measure to verify the precision or reliability of the estimators. In statistics we know that the precision of an estimate is measured by its standard error. By definition, the standard error of an estimate is the standard deviation of its sampling distribution.

As the standard deviation of an estimator $\hat{\theta}$ *is* $\sqrt{Var(\hat{\theta})}$, the standard errors of the OLS estimates are given as:

$$Se\left(\hat{\beta}_0\right) = \sqrt{\text{var}\left(\hat{\beta}_0\right)} = \sqrt{\sigma_u^2 \left(\frac{\sum x_i^2 + n\bar{x}^2}{n\sum x_i^2}\right)}$$

$$= \sqrt{\frac{\sum X_i^2}{n\sum x_i^2}\sigma^2{}_u} \quad \text{.....3.45}$$

$$Se\left(\hat{\beta}_1\right) = \sqrt{\text{var}\left(\hat{\beta}_1\right)} = \sqrt{\sigma_u^2 \Big/ \sum x_i^2} \quad \text{.....3.46}$$

Where, $\hat{\sigma}^2{}_u = \frac{\sum e^2{}_i}{n-2}$

3.2 The Gauss – Markov Theorem

In this section we will show that the least squares estimates possess some ideal or optimum properties. These properties are established in the famous Gauss – Markov theorem. This theorem states that, given the assumptions of the classical regression models, the least squares estimators have minimum variance in the class of unbiased linear estimators; i.e., they are Best, Linear and Unbiased Estimators (BLUE).

To understand this theorem, we have to know the desirable properties of statistical estimators. These properties include

i) **Linearity:** The estimator should be a linear function of the sample observations.

ii) **Unbiasedness:** The average or expected value of an estimator should be equal to the true value.

iii) **Having minimum variance:** The estimator should have minimum variance in the class of linear and unbiased estimators.

Generally, estimators that have these properties are BLUE and it can be shown that the OLS estimators are BLUE, which can be seen from the following proof of Gauss – Markov theorem.

3.2.1 Linearity

The OLS estimators $\hat{\beta}_0$ *and* $\hat{\beta}_1$ are linear functions of the observed sample values of Y_i and X_i i.e, their formula involve the values of Y and X in their first power.

To prove this, recall that from equation 2.23 and 2. 24, we have

$$\hat{\beta}_1 = \frac{\sum x_i \, y_i}{\sum x_i^2} \quad and \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Then given that the X_i s always appear with the same values in the hypothetical repeated sampling process, it can easily be shown that the least square estimates depend on the values of Y only, i.e., $\hat{\beta}_0 = f(Y_i)$ *and* $\hat{\beta}_1 = f(Y_i)$

Proof

We know from equation 3.10 that

$$\hat{\beta}_1 = \sum \left(\frac{x_i}{\sum x_i^2} \right) Y_i = \sum k_i Y_i \quad , \quad where \quad k_i = \frac{x_i}{\sum x_i^2}$$

Since k_i's are assumed to be fixed/constants from sample to sample, they may be regarded as constant weights assigned to individual values of Y.

Hence,

$$\hat{\beta}_1 = \sum k_i \, Y_i = k_1 \, Y_1 + k_2 \, Y_2 + \ldots\ldots\ldots + k_n \, Y_n$$

$$\Rightarrow \hat{\beta}_1 = f(Y)$$

Thus, the estimate $\hat{\beta}_1$ is a linear function of the Y's, i.e., a linear combination of the values of the dependent variable.

By analogy we can show that $\hat{\beta}_0$ is linear in sample observations of Y. Under equation 3 .14

$$\hat{\beta}_0 = \sum \left(1/n - \overline{X}\, k_i\right) Y_i$$

We know that $\overline{X}$ *and* k_i are fixed constants from sample to sample. Thus $\hat{\beta}_0$ depends only on the values of Y, i.e., $\hat{\beta}_0$ is a linear function of sample values of Y.

3.2.2 Unbiasedness

In general an estimator, say $\hat{\theta}$, is an unbiased estimator of the true parameter, say θ, if the expected value of $\hat{\theta}$ is equal to θ. Since it has already been established in section 3.1.1 and 3.1.3 (under equations 3.9 and 3.15) that

$$E\left(\hat{\beta}_0\right) = \beta_0 \quad \text{and}$$

$$E\left(\hat{\beta}_1\right) = \beta_1$$

we can conclude that the OLS estimates $\hat{\beta}_0$ *and* $\hat{\beta}_1$ respectively are unbiased estimators of the true parameters β_0 and β_1.

The meaning of this property is that the estimates converge to the true value of the parameters as we increase the number of (hypothetical) samples taken from a certain parent population. In other words, if we take all possible samples (or a very large number of samples) of size n of observations on Y and X and compute the estimates $\hat{\beta}_0$ *and* $\hat{\beta}_1$ for each sample, we will obtain a large number of such estimates, whose mean (expected value) will be equal to the true parameters of the relationship.

Activity

Dear readers, which assumption(s) of linear regression model did we use to establish the unbiasedness properties of OLS estimates? ----------

---

3.2.3 Minimum Variance

The Gauss–Markov theorem state that the least squares estimates are best (have the smallest variance) as compared with any other linear unbiased estimators obtained from any other econometric methods. This property is the main reason for the popularity of the OLS method.

Here it is noteworthy that there might be other non–linear and biased estimators obtained from other econometric methods with smallest variance, but the comparison of the OLS estimates is restricted traditionally to the class of linear unbiased estimators. Now, let us show that the OLS estimators have minimum variance in the class of linear unbiased estimators.

Proof

i) Prove that variance of $\hat{\beta}_1$ is the lowest in the class of linear unbiased estimators of β_1

Recall that $Var\left(\hat{\beta}_1\right) = \sigma_u^2 \left(1 \Big/ \sum x_i^2 \right)$

Now, we want to prove that any other linear unbiased estimate of the true parameter, for example β_1^* , obtained from any other econometric method, has a bigger variance than the least squares estimate $\hat{\beta}_1$. This means we want to prove that

$$Var\ \left(\hat{\beta}_1\right)\ <\ Var\ \left(\beta_1^*\right).$$

Firstly: We know that β_1^* is by assumption a linear combination of the values of Y, which means it is the weighted sum of the sample values of Y where the weights being different from the weights $k_i \left[= x_i \Big/ \sum x_i^2 \right]$ of the least squares estimates.

Hence, suppose that

$$\beta_1^*\ =\ \sum c_i\, Y_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.48$$

Where $c_i\ =\ k_i\ +\ d_i$, k_i being the weights defined for the OLS estimates and d_i being an arbitrary set of weights similar (but not the same) to k_i.

Substituting $\beta_0\ +\ \beta_1\ X_i + U_i\ \ for\ Y_i$ in equation 3.48 we obtain

$$\beta_1^*\ =\ \sum c_i\ \left[\beta_0\ +\ \beta_1\ X_i\ +\ U_i\right]\ =\ \sum\ \left[\beta_0\ c_i\ +\ \beta_1 c_i\ X_i\ +\ c_i\ U_i\right]$$

$$=\ \beta_0 \sum c_i\ +\ \beta_i\ \sum c_i X_i\ +\ \sum c_i\ U_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.49$$

Secondly: The new estimate β_1^* is also assumed to be an unbiased estimator of the true β_1 , *i.e* $E\left(\beta_1^*\right)\ =\ \beta_1$.

Taking the expected values of the expression in equation 3.49, we get

$$E\left(\beta_1^*\right) = E\left[\beta_0 \sum c_i + \beta_1 \sum c_i X_i + \sum c_i U_i\right]$$

However, $E\left(\beta_1^*\right) = \beta_1$ if and only if

i) $\sum c_i = 0$

ii) $\sum c_i u_i = 0$

iii) $\sum c_i X_i = 1$

But the fact that $\sum c_i = 0$ *implies* $\sum d_i = 0$

This is due to the fact that $\sum c_i = \sum\left(k_i + d_i\right) = \sum k_i + \sum d_i$, which will be equal to zero only if $\sum d_i = 0$ $\sin ce$ $\sum k_i = 0$

Similarly, $\sum c_i x_i = 1$ *requires* $\sum d_i x_i$ to be equal to zero. This is because

$$\sum c_i X_i = \sum\left(k_i + d_i\right)X_i = \sum k_i X_i + \sum d_i X_i$$

$$= 1 + \sum d_i X_i$$

Hence, $\sum c_i X_i = 1$ if and only if $\sum k_i d_i = 0$

In summary

i) $\sum c_i = 0$ *ii*) $\sum d_i = 0$, *iii*) $\sum c_i X_i = 1$ *iv*) $\sum d_i X_i = 0$

Thirdly: The variance of the new estimator β_1^* is given as

$$Var\left(\beta_1^*\right) = Var\left(\hat{\beta}_1\right) + \sigma_u^2 \left(\sum d_i^2\right)$$

Proof

The procedure for the derivation of $Var\left(\beta_1^*\right)$ is the same as that used for deriving the variance of the least squares estimate $\hat{\beta}_1$.

Recall that in equation 3.2 we have shown that

$$\hat{\beta}_1 = \sum k_i Y_i$$

$$\Rightarrow Var\left(\hat{\beta}_1\right) = Var\left(\sum\left(k_i Y_i\right)\right) = \sum Var\left(k_i Y_i\right) = \sum\left[k_i^2 \; Var\left[Y_i\right]\right] = \sum k_i^2 \sigma_u^2 = \sigma_u^2 \sum k_i^2$$

By analogy we may establish the variance of β_1^* as follows

$$\beta_1^* = \sum c_i Y_i$$

$$\Rightarrow Var\left(\beta_1^*\right) = Var\left[\sum c_i Y_i\right]$$

$$= \sum Var\left[c_i Y_i\right] = \sum\left[c_i^2 \; Var\left[Y_i\right]\right] = \sigma_u^2 \sum c_i^2 \quad \text{……..3.50}$$

Where, c_i is constant weight independent of the values of Y.

Then,

$$\sum c_i^2 = \sum\left(k_i + d_i\right)^2$$

$$= \sum\left(k_I^2 + 2k_I \, d_I + d_I^2\right)$$

$$= \sum k_i^2 + 2\sum k_i \, d_i + \sum d_i^2$$

$$= \sum k_i^2 + \sum d_i^2 + 2\sum k_i \, d_i$$

But

$$\sum k_i \, d_i = \sum\left(\frac{x_i}{\sum x_i^2}\right) d_i$$

$$= \frac{\sum x_i d_i}{\sum x_i^2}$$

$$= \frac{\sum \left(X_i - \overline{X}\right) d_i}{\sum x_i^2}$$

$$= \frac{\sum d_i X_i - \overline{X} \sum d_i}{\sum x_i^2}$$

$$= \sum d_i X_i$$

$= 0$ Since, from the conditions set above $\sum d_i X_i = 0$ and $\sum d_i = 0$

$$\Rightarrow \sum c_i^2 = \sum k_i^2 + \sum d_i^2 \quad \text{.....3.51}$$

Substituting equation 3.51 into equation 3.50 yields

$$Var\left(\beta_1^*\right) = \sigma_u^2 \left[\sum k_i^2 + \sum d_i^2\right]$$

$$= \sigma_u^2 \sum k_i^2 + \sigma_u^2 \sum d_i^2 \quad \text{.....3.52}$$

But we have already shown that $\sigma_u^2 \sum k_i^2 = Var\left(\hat{\beta}_1\right)$

Thus, equation 3.52 becomes

$$Var\left(\beta_1^*\right) = Var\left(\hat{\beta}_1\right) + \sigma_u^2 \sum d_i^2 \quad \text{.....3.53}$$

Given that d_i's are defined as arbitrary constant weights and not all of them are zero at the same time, the second term is always positive.

$$\Rightarrow \sigma_u^2 \sum d_i^2 > 0$$

Therefore,

$$Var\ \left(\beta_1^*\right) > Var\ \left(\hat{\beta}_1\right)$$

Thus in a group of linear unbiased estimates of the true β_1, the least squares estimate has minimum variance. i.e. $\hat{\beta}_1$ is the best estimator of β_1

By following the same procedure, we can proof that the OLS estimator, $\hat{\beta}_0$ is the best estimator in the class of linear unbiased estimators of the true parameter β_0, the proof of which is shown below.

Firstly: We take a new estimator $\tilde{\beta}_0$ that we assume to be a linear function of Y, with weights being $c_i = k_i + d_i$, as defined above.

We have shown above that the OLS estimator $\hat{\beta}_0$ is given as:

$$\hat{\beta}_0 = \sum\left[\frac{1}{n} - \overline{X}\ k_i\right] Y_i$$

Hence, since $\tilde{\beta}_0$ is assumed to be linear by analogy it can be given as:

$$\hat{\beta}_0 = \sum\left[\frac{1}{n} - \overline{X}\ c_i\right] Y_i \quad \text{.....3.54}$$

Secondly: we want $\tilde{\beta}_0$ to be an unbiased estimator of the true β_0, i.e,

$$E\left(\tilde{\beta}_0\right) = \beta_0 .$$

Substituting for Y_i in equation 3.54 that $Y_i = \beta_0 + \beta_1 X_i + U_i$, we obtain

$$\tilde{\beta}_0 = \sum\left[1/n - \overline{X}\, c_i\right]\left[\beta_0 + \beta_1\ X_i + U_i\right]$$

$$= n\,\beta_0\left[1/n\right] - \beta_0\,\overline{X}\sum c_i + \beta_1\frac{\sum X_i}{n} - \beta_1\,\overline{X}\sum c_i X_i + \sum\left[\frac{1}{n} - \overline{X}X_i\right]U_i$$

$$= \beta_0 - \beta_0\,\overline{X}\sum c_i + \beta_1\left[\overline{X} - \overline{X}\sum c_i X_i\right] + \sum\left[\frac{1}{n} - \overline{X}X_i\right]U_i \ldots\ldots\ldots\ldots\ldots 3.55$$

Taking expected values of the terms in the both sides of equation 3.55, gives

$$E\left(\tilde{\beta}_0\right) = E\left\{\beta_0\left[1 - \overline{X}\sum c_i\right]\right\} + \beta_1\left[\overline{X} - \overline{X}\,E\left[\sum c_i X_i\right]\right] + E\left\{\left[\sum\left(\frac{1}{n} - \overline{X}\,X_i\right)\right]\left[U_i\right]\right\}$$

However, we assumed that $\tilde{\beta}_0$ is an unbiased estimator of β_0. Consequently, $E\left(\tilde{\beta}_0\right) = \beta_0$. But this holds in equation 3.56 if and only if the following conditions are satisfied:

i) $\sum c_i = 0$ $\quad ii)\ \sum c_i X_i = 1,$ $\quad iii)\ \sum c_i U_i = 0$

These conditions, following the definition $c_i = k_i + d_i$, imply that

i) $\sum d_i = 0$ $\quad ii)\ \sum d_i X_i = 0$

Activity

Dear readers, show that $\sum d_i = 0$ ***and*** $\sum d_i X_i = 0$

Thirdly, from equation 3.54, the variance of $\tilde{\beta}_0$ will be given as:

$$Var\left(\tilde{\beta}_0\right) = \sum Var\left[\left(\frac{1}{n} - \overline{X}\,c_i\right)Y_i\right]$$

$$= \sum\left(\frac{1}{n} - \overline{X}\ c_i\right)^2\ Var\ (Y_i)$$

$$= \sigma_u^2\ \sum\left[\frac{1}{n} - \overline{X}\ c_i\right]^2 \text{, since } Var\ (Y_i) = \sigma_u{}^2$$

$$= \sigma_u^2\ \sum\left[\frac{1}{n^2} + \overline{X}^2 c_i{}^2 - 2\frac{\overline{X}c_i}{n}\right]$$

$$= \sigma_u^2\left[n \cdot \frac{1}{n^2} + \overline{X}^2 \sum c_i^2 - \frac{2\overline{X}}{n}\sum c_i\right]$$

Given that $\sum c_i = 0\ and\ \sum c_i^2 = \sum k_i^2 + \sum d^2{}_i$, we have

$$Var\left(\tilde{\beta}_0\right) = \sigma_u^2\left[\frac{1}{n} + \overline{X}^2\left(\sum k_i^2 + \sum d_i^2\right)\right]$$

$$= \sigma_u^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum x_i^2}\right] + \sigma_u^2\ \overline{X}^2\ \sum d_i^2 \text{, since } \sum k_i{}^2 = \frac{1}{\sum x^2{}_i}$$

$$= Var\left(\hat{\beta}_0\right) + \sigma_u^2\ \overline{X}^2\ \sum d_i^2 \text{, since from}$$

equation 3.16 $Var\left(\hat{\beta}_0\right) = \sigma_u^2\left(\frac{\sum x_i^2 + n\,\bar{x}^2}{n\sum x_i^2}\right)$

But we know that $\sum d_i^2 > 0$, $since$ all $d_i's$ are not zero simultaneously

Therefore,

$$Var\left(\tilde{\beta}_0\right) > Var\left(\hat{\beta}_0\right)$$

Note: To establish the BLU properties of the OLS estimates we made use of the Assumptions 1 , 2 , 3, 5 and 6. However, it did not require the fulfillment of

Assumption 4 for $\beta's$ to have the BLU properties, i.e., they are BLU even if U is not normally distributed.

Activity

Show that $\sum c_i^2 = \sum k_i^2 + \sum d^2{}_i$ ***(hint: use*** $c_i = k_i + d_i$***)***

Exercise

1. Show that the OLS estimates of the parameters of the true relationship between two variables are best, linear and unbiased. Explain the assumption you used to proof the BLU properties of the OLS estimates.

2. The following results are obtained from a sample of 11 observations on the dependent variable (Y) and explanatory variable (X)

$\overline{X} = 520$

$\overline{Y} = 220$ $\qquad\qquad$ $\sum X_i Y_i = 1290$

$\sum X_i^2 = 3100$ $\qquad\qquad$ $\sum Y_i^2 = 539{,}500$

Find the variance of the OLS estimates of the coefficients of regression line of Y on X

3. Suppose in question #2 above, on rechecking the data it was found hat two pairs of observations were erroneously recorded as

Y	*X*		*Y*	*X*
90	120	instead of	80	110
140	220		150	210

i. Find the variance of the OLS estimates of the coefficients of regression line of Y on X using the correct data.

ii. Explain the effect of the data recording error on the variance of the estimates of the coefficients of the regression model in question #2.

4. Given the following two models:

Model I: $Y_i = \beta_0 + \beta_1 X_i + U_i$

Model II: $Y_i = \alpha_0 + \alpha_1 (X_i - \overline{X}) + U_i$

i. Find the variances of the OLS estimators of β_0 and α_0. Explain the differences, if any.

ii. Find the variances of the OLS estimators of β_1 and α_1. Explain the differences, if any.

iii. What is the advantage, if any, of model II over model I?

CHAPTER 4

STATISTICAL TESTS OF THE REGRESSION MODEL

Overview

In the previous chapters we discussed the OLS method of estimating the parameters of economic relationships, and dealt with the procedures of estimating the means and standard deviations of OLS estimates. Furthermore, we made attempts to look into the properties of OLS estimates.

In this chapter, however, emphasis will be given to statistical tests of least square estimates. Succinctly, chapter 4 deals with the statistical criteria, which is one of the three criteria discussed in chapter one used by economists to gauge the "goodness" of the parameter estimates. Specifically, the concern of this chapter will be to explain the two most commonly used statistical measures, such as: *the Coefficient of Determination (r^2) and the Standard Errors of the Parameter Estimates, $Se(\hat{\beta}_i)$*

At the end of this chapter, students will be able to:

- Define the coefficient of determination and drive its formula
- Differentiate between different types of significance tests and conduct individual significance tests using standard errors of the estimates, *t distribution and standard normal distribution.*
- Construct confidence intervals for the parameters of a model and use them to conduct tests of significance on the estimates of the parameters

4.1 The Coefficient of Determination (r^2): A Measure of "Goodness of fit"

After estimating the parameters and establishing the fact that the estimates possess appropriate properties, the next step is to measure the fitness of the regression line obtained from them to the sample observations of Y and X. In general, the fitness of the regression line is measured based on the dispersion of observations around the regression line. This is due to the fact that when we plot the data observed from the survey on XY-plane, we will see that all the actual observations wouldn't lie on the regression line. The actual observations will deviate from the predicted/estimated values and hence, there will be positive and/or negative values for the residual term U_i. Thus, we expect a good regression line to minimize these residuals around it. Generally, the closer the observations to the regression line, the better the goodness of fit of the line will be. But how to measure the goodness of fit of the regression line still remains to be the question that this part attempts to answer.

The common measure of the goodness of fit of the regression line is the square of the correlation coefficient, r^2. It shows the parentage of the total variation of the dependent variable, say Y, that can be explained by the independent (explanatory) variable, say X. In simple words, r^2 is a summary measure that tells how well the regression line fits the actual data.

To understand this meaning of r^2, let us consider the following graphical explanation in terms of Venn diagram, where circle Y represents the variation in the dependent variable Y and circle X represents the variation in the explanatory variable X, and that the overlap of the two circles (i.e. the shaded area) indicates the extent to which the variation in Y is explained by the variation in X.

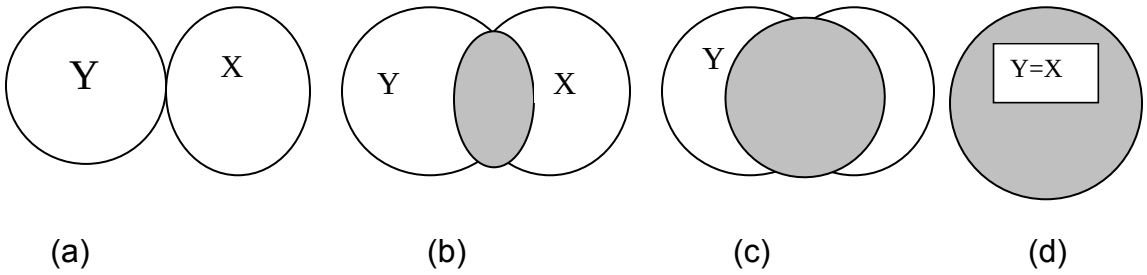


Figure 4.1 The Venn diagram view of r^2

It is evident from figure 4.1 that the greater the extent of the overlap between the two circles, the greater the variation in Y is explained by X . A numerical measure of this overlap can be considered as r^2. In the above figures as the overlap increases from "a" to "d", the r^2 value increases ipso facto. In figure "a" there is no overlap; which indicates that none of the variations in Y is explained by variation in X and hence r^2 is zero. On the other hand, in figure "d", the overlap is complete indicating that 100 percent of the variation in Y is explained by X and r^2 is 1. Hence, r^2 lies between 0 and 1.

To compute the value of r^2 and prove it to be the measure of goodness of fit, let us plot the observations on a rectangular co-ordinate system and then compute the means as:

$$\overline{X} = \sum_{i=1}^{n} X_i \Big/ n \qquad \textit{and} \qquad \overline{Y} = \sum_{i=1}^{n} Y_i \Big/ n$$

Then, draw perpendiculars through the points of these means as shown below.

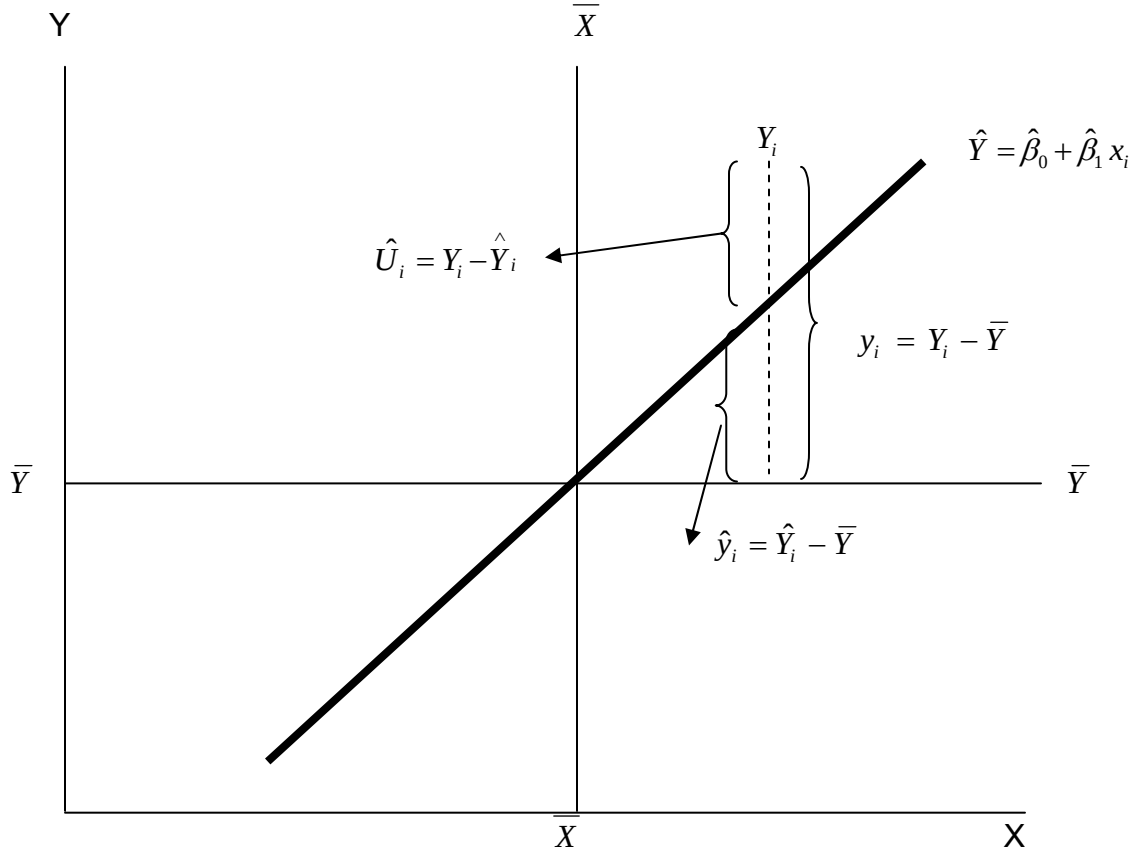


Figure 4.2 Partition of the Variation in Y_i

By fitting a regression line $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1$, we try to obtain the explanation of the variations of the dependent variable Y produced by the changes of the explanatory variable X. However, the fact that the actual observations deviate from the estimated line shows that the regression line explains only a part of the total variation of the dependent variable. In other words, a part of the variation defined as $\hat{U}_i = Y_i - \hat{Y}_i$ remains unexplained, where $\hat{U}_i$ is the estimate of U_i. (see Figure 4.2)

As it can be seen from figure 4.2, we may compute the total variation of the dependent variable by comparing each value of Y to the mean value $\overline{Y}$. Then,

adding the squares of all the resulting deviations we obtain the total sum of squares (TSS) as:

$$[\text{Total Variation in } Y\ ] \quad = \sum \left(Y_i - \bar{Y}\right)^2 = \sum y_i^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots 4.1$$

In the same way we can define the deviation of the regressed values, $\hat{Y}_i's$ from the mean value as $\hat{y}_i = \hat{Y}_i - \bar{Y}$. This is the part of the total variation of Y_i that is explained by the regression line. Thus, the sum of the squares of these deviations is the total explained variation of the dependent variable and is called the explained sum of squares (ESS), which is given as:

$$[\text{Explained Variation}] \quad = \sum_i \left(\hat{Y}_i - \bar{Y}\right)^2 = \sum_i \hat{y}_i^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots 4.2$$

The part of the variation of the dependent variable that is not explained by the regression line is given by the difference $\hat{U}_i = Y_i - \hat{Y}_i$ and is attributed to the existence of the disturbance variable, U. Thus, the sum of the squared residuals gives the total unexplained variation of the dependent variable Y around its mean; which may be called the residual sum of squares (RSS)

$$[\text{Unexplained Variation}] \quad = \sum_i \left(Y_i - \hat{Y}_i\right)^2 = \sum_i \hat{U}_i^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots 4.3$$

In summary, the variation in the observed Y can be decomposed as follows.

1) Deviation of the actual observations of Y_i from the regression line is given as

$$\hat{U}_i = Y_i - \hat{Y}_i \ldots\ldots\ldots\ldots\ldots\ldots 4.4$$

2) Deviation of the actual observations of Y_i from its mean value is obtained as:

$$y_i = Y_i - \bar{Y} \ldots\ldots\ldots\ldots\ldots\ldots 4.5$$

3) Deviation of the regressed value, $\hat{Y}_i$ from the mean value is given as:

$$\hat{y}_i = \hat{Y}_i - \overline{Y} \quad \text{.....4.6}$$

Then, by re arranging equation 4.5, we obtain

$$Y_i = y_i + \overline{Y} \quad \text{.....4.7}$$

Similarly from equation 4.6, we obtain

$$\hat{Y}_i = \hat{y}_i + \overline{Y} \quad \text{.....4.8}$$

Then, by substituting equation 4.7 and 4.8 in equation 4.4, we obtain

$$\hat{U}_i = (y_i + \overline{Y}) - (\hat{y}_i + \overline{Y})$$

$$= y_i + \overline{Y} - \hat{y}_i - \overline{Y}$$

$$= y_i - \hat{y}_i \quad \text{.....4.9}$$

By rearranging equation 4.9, we get

$$y_i = \hat{U}_i + \hat{y}_i \quad \text{..... 4.10}$$

Note from equation 4.10 that each deviation of the observed values of Y from its mean consists of two components: the explained variation and unexplained variation.

Recall from equation 4.1 that total variation is given as:

$$TSS = \sum y_i^2$$

Then, putting equation 4.10 into it, we obtain

$$\sum y_i^2 = \sum \left(\hat{y}_i + \hat{u}_i \right)^2$$

$$= \sum \left(\hat{y}_i^2 + 2\, \hat{y}_i\, \hat{u}_i + \hat{u}_i^2 \right)$$

$$= \sum \hat{y}_i^2 + 2 \sum \hat{y}_i\, \hat{u}_i + \sum \hat{u}_i^2 \quad \text{.....4.11}$$

But, $\sum \hat{y}_i\, \hat{u}_i = 0$. Why?

Proof

Recall from equation 4.6 that $\hat{y}_i = \hat{Y}_i - \overline{Y}$. Moreover, we know that

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

and

$$\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{X}$$

Thus, substituting these two equations into equation 4.6, we get

$$\hat{y}_i = \left(\hat{\beta}_0 + \hat{\beta}_1 X_i \right) - \left(\hat{\beta}_0 + \hat{\beta}_1 \overline{X} \right)$$

$$= \hat{\beta}_0 + \hat{\beta}_1 X_i - \hat{\beta}_0 - \hat{\beta}_1 \overline{X}$$

$$= \hat{\beta}_1 X_i - \hat{\beta}_1 \overline{X}$$

$$= \hat{\beta}_1 \left(X_i - \overline{X} \right)$$

$$= \hat{\beta}_1 x_i \quad \text{..... .....4.12}$$

Where, $x_i = X_i - \overline{X}$

Hence, by plugging equation 4.12 into equation 4.9, we obtain:

$$\hat{u}_i = y_i - \hat{\beta}_1 x_i \quad \text{.....4.13}$$

Thus, taking the sum of the product of equation 4.12 and equation 4.13, we get

$$\sum \hat{y}_i \hat{u}_i = \sum \left(\hat{\beta}_1 x_i \right) \left(y_i - \hat{\beta}_1 x_i \right)$$

$$= \sum \left[\hat{\beta}_1 x_i y_i - \hat{\beta}_1^2 x_i^2 \right]$$

$$= \hat{\beta}_1 \left[\sum x_i y_i - \hat{\beta}_1 \sum x_i^2 \right] \quad \text{.....4.14}$$

But we know that

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$$

Therefore, substituting the formula of $\hat{\beta}_1$ in equation 4.14, we get:

$$\sum \hat{y}_i \hat{u}_i = \hat{\beta}_1 \left[\sum x_i y_i - \left(\frac{\sum x_i y_i}{\sum x_i^2} \right) \sum x_i^2 \right]$$

$$= \hat{\beta}_1 \left[\sum x_i y_i - \sum x_i y_i \right]$$

$$= \hat{\beta}_1 [\ 0\ ]$$

$$= 0$$

Therefore, equation 4.11becomes

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum \hat{u}_i^2 \quad \text{.....4.15}$$

In words,

$$\begin{bmatrix} Total \\ Variation \end{bmatrix} = \begin{bmatrix} Explained \\ Variation \end{bmatrix} + \begin{bmatrix} Unexplained \\ Variation \end{bmatrix}$$

$$TSS = ESS + RSS$$

Hence, the explained variation as a percentage of total variation can be written as:

$$\Rightarrow \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{\sum \left(\hat{\beta}_1 x_i \right)^2}{\sum y_i^2}$$

$$= \frac{\hat{\beta}_1^2 \sum x_i^2}{\sum y_i^2} \text{ ..... Since } \hat{y}_i = \hat{\beta}_1 x_i$$

Given that $\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_2}$, this becomes

$$\frac{\sum \hat{y}_i^2}{\sum y_i^2} = \left(\frac{\sum x_i y_i}{\sum x_i^2} \right)^2 \frac{\sum x_i^2}{\sum y_i^2}$$

$$= \frac{\left(\sum x_i y_i \right)^2}{\left(\sum x_i^2 \right)^2} \frac{\sum x_i^2}{\sum y_i^2}$$

$$= \frac{\left(\sum x_i y_i \right)^2}{\sum x_i^2 \sum y_i^2}$$

$$= \left[\frac{\sum x_i\ y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}\right]^2 \quad \text{.....4.16}$$

But we know that the term $\frac{\left(\sum x_i\ y_i\right)}{\sqrt{\sum x_i^2\ \sum y_i^2}}$ is the formula of the correlation coefficient, r. Thus, equation 4.16 can be rewritten as

$$\frac{\sum \hat{y}_i^2}{\sum y_i^2} = r^2$$

Thus, $r^2 = \left[\frac{\sum x_i\ y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}\right]^2$

$$= \frac{\left(\sum x_i\ y_i\right)^2}{\sum x_i^2\ \sum y_i^2} \quad \text{.....4.17}$$

This show that r^2 determines the proportion of the variation in Y which is explained by variation in X and this is the reason why r^2 is some times called the coefficient of determination.

For example, if $r^2_{y\,,\,x} = 0.9$, it implies that the regression line gives a good fit to the observed data, since it explains 90 percent of the total variation of the Y values around their mean. The remaining 10 percent of the total variation in Y is unaccounted for by the regression line and is attributed to factors included in the disturbance variable U.

The properties of r^2:

i) It is a non-negative quantity

ii) It's limits are $0 \le r^2 \le 1$.

Activity

Dear students, why do you think r^2 does not assume negative values?

Leaving the justification of the first property as an exercise for readers, let us show that $0 \leq r^2 \leq 1$.

Proof

Recall that

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum u_i^2 \qquad \text{(see equation 4.15)}$$

Dividing through equation 4.15 by $\sum y_i^2$ we get

$$1 = \frac{\sum \hat{y}_i^2}{\sum y_i^2} + \frac{\sum u_i^2}{\sum y_i^2}$$

$$\Rightarrow 1 = r^2 + \frac{\sum \hat{u}_i^2}{\sum y_i^2}$$

$$\Rightarrow r^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2} \quad \text{.....4.18}$$

.. Note in equation 4.18 that $\frac{\sum \hat{u}_i^2}{\sum y_i^2}$ is the proportion of the unexplained variation of the $Y's$ around their mean, $\overline{Y}$. If all the observations lie on the regression line, $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, there will be no scatter of points. In other words the total

variation of Y is explained completely by the estimated regression line, and consequently there will be no unexplained variation, i.e., in equation 4.18

$$\frac{\sum \hat{u}_i^2}{\sum y_i^2} = 0 \text{ and hence } r^2 = 1.$$

On the other hand, if the regression line explains only part of the variation in Y, there will be some unexplained variation, i.e., in equation 4.18

$$\frac{\sum \hat{u}_i^2}{\sum y_i^2} > 0.$$

Therefore,

$$r^2 < 1.$$

Finally, if the regression line does not explain any part of the variation of Y, $\frac{\sum \hat{u}_i^2}{\sum y_i^2} = 1$ since $\sum y_i^2 = \sum \hat{u}_i^2$. Therefore, $r^2 = 0$, which shows that the limits of r^2 are $0 \le r^2 \le 1$.

To express the formula of r^2 in terms of the slope of the regression line $(\hat{\beta}_1)$, we proceed as follows.

Recall that in equation 4.17, we have already shown that

$$r^2 = \left(\sum x_i\, y_i\right)^2 \Big/ \sum x_i^2 \cdot \sum y_i^2$$

$$= \frac{\left(\sum x_i\, y_i\right)}{\sum x_i^2} \quad \frac{\left(\sum x_i\, y_i\right)}{\sum y_i^2}$$

But we have already shown that

$$\hat{\beta}_1 = \frac{\sum x_i\, y_i}{\sum x_i^2}$$

$$\Rightarrow r^2 = \hat{\beta}_1 \cdot \frac{\sum x_i y_i}{\sum y_i^2} \quad \text{.....4.19}$$

Multiplying the terms in the right hand side of equation 4.19 by $\frac{\sum x_i^2}{\sum x_i^2}$ (which in essence is the same as multiplying by 1), produces

$$r^2 = \hat{\beta}_1 \frac{\sum x_i y_i}{\sum y_i^2} \cdot \frac{\sum x_i^2}{\sum x_i^2}$$

By rearranging the terms in this equation, it can be rewritten as

$$\Rightarrow r^2 = \hat{\beta}_1 \frac{\sum x_i y_i}{\sum x_i^2} \cdot \frac{\sum x_i^2}{\sum y_i^2}$$

$$= \hat{\beta}_1^2 \frac{\sum x_i^2}{\sum y_i^2} \quad \text{.....4.20}$$

Furthermore, if we divide the numerator and denominator in equation 4.20 by sample size n (for small sample size by $n-1$), we obtain

$$r^2 = \hat{\beta}_1^2 \frac{\frac{\sum x_i^2}{n}}{\frac{\sum y_i^2}{n}}$$

$$\Rightarrow r^2 = \hat{\beta}_1^2 \left(S_x^2 / S_y^2 \right) \quad \text{.....4.21}$$

Where S_x^2 *and* S_y^2 are sample variances of X and Y, respectively.

Finally, using these definitions of r^2, we can express RSS and TSS as follows:

We have already shown that

$$r^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{ESS}{TSS}$$

$$\Rightarrow ESS = (TSS)\, r^2$$

$$= r^2 \sum y_i^2 \quad \text{.....4.22}$$

However, we know that

$$\text{TSS} = \text{ESS} + \text{RSS}$$

$$\Rightarrow \text{RSS} = \text{TSS} - \text{ESS}$$

$$= TSS - \left(r^2\, TSS\right) \text{ , since } ESS = \left(TSS\right) r^2$$

$$= \text{TSS } (1 - \text{r}^2)$$

$$\Rightarrow \text{RSS} = \sum y_i^2 \left(1 - r^2\right) \quad \text{.....4.23}$$

Thus, combining equation 4.22 and 4.23, we get

$$TSS = r^2 \sum y_i^2 + \sum y_i^2 \left(1 - r^2\right) \quad \text{.....4.24}$$

4.2 Hypothesis Testing

As you may remember from your **"Statistics for Economists (Econ 321)"**, the problem of statistical hypothesis testing deals with testing whether a given observation or finding is compatible with some pre stated hypothesis. The word "compatible", here, means that the finding is "sufficiently" close to the hypothesized value so that we do not reject the stated hypothesis. In statistics, the stated hypothesis is known as the null hypothesis, denoted by H_0, and it is tested against an alternative hypothesis, H_1.

For example, if a theory leads us to believe that the slope coefficient is unity

(i.e., $\beta_1 = 1$), the interest in hypothesis testing is to show how close or consistent is the observed (or estimated) $\hat{\beta}_1$ with the stated hypothesis. Thus, this hypothesis can be written as

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Generally, the theory of hypothesis testing is concerned with developing rules or procedures for deciding whether to reject or not to reject the null hypothesis. There are two mutually complementary approaches for devising such rules: Test of significance approach and Confidence interval approach

Both these approaches assume that the variable (or the estimator) under consideration has some probability distribution and that hypothesis testing involves making assertions about the value(s) of the parameter(s) of the assumed distribution.

4.2.1 The Test of Significance Approach

This approach is a procedure by which sample results are used to verify the validity of the null hypothesis. In this case the rule to accept or reject the null hypothesis is developed on the basis of test statistic obtained from the data at hand. To develop the decision rule for the rejection or acceptance of H_0, we can use different test statistics based on the assumptions about the distribution of the population of the variable under consideration. In this part we are mainly interested in OLS estimates and test of significance is conducted to see their statistical significance.

In general, the least square estimates (for instance $\hat{\beta}_0$ *and* $\hat{\beta}_1$) are obtained from a sample of observations on Y and X. Since sampling errors are inevitable in all estimates, it is necessary to apply tests of significance in order to measure the

size of the error and determine the degree of confidence in the validity of the estimates.

There are several tests under the test of significance approach to accomplish this purpose: *the standard error test, the Z-test and the Student's t- test.*

4.2.2.1 The standard error test of the least square estimates

This test uses the standard error of the estimates as a test statistic to decide on the rejection or acceptance of the null hypothesis. In other words, it helps us to decide whether the estimates $\hat{\beta}_0 \ and \ \hat{\beta}_1$ are significantly different from zero; i.e., whether the sample from which they have been estimated has come from a population whose true parameters are zero $(\beta_0 = 0 \ \ and/or \ \ \beta_1 = 0)$.

Formally, we test the null hypothesis $H_0 : \beta_i = 0$ against the alternative hypothesis $H_1 : \beta_i \neq 0$

To conduct the standard error test, we pass through the following three steps.

Step 1. *Obtain the standard errors of the estimates*

We have already shown in chapter 3 under equations 3.45 and 3.46 that the standard errors of OLS estimates for β_0 and β_1 are given as:

$$Se\left(\hat{\beta}_0\right) = \sqrt{Var\left(\hat{\beta}_0\right)} = \sqrt{\frac{\hat{\sigma}_u^2 \sum X_i^2}{n \sum x^2}} = \sqrt{\frac{\left(\sum \hat{u}_i^2\right) \sum X_i^2}{(n-2) \sum x_i^2}}$$

$$Se\left(\hat{\beta}_1\right) = \sqrt{Var\left(\hat{\beta}_1\right)} = \sqrt{\hat{\sigma}_u^2 \Big/ \sum x_i^2} = \sqrt{\frac{\sum \hat{u}_i^2}{(n-2) \sum x_i^2}}$$

Step 2. *Compare the standard deviations (errors) obtained in step 1, with the numerical values of* $\hat{\beta}_0$ *and* $\hat{\beta}_1$

Dear readers, you may remember that in chapter 2 we have already developed the formula to obtain numerical values of $\hat{\beta}_0$ *and* $\hat{\beta}_1$*. For the purpose of convenience, their formulae are given below (see equations 2.23 and 2.24)*

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

And

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$$

Step 3. *Make decision*

Decision Rule

If the standard error of an estimate is smaller than half its numerical value (i.e., $Se(\hat{\beta}_i) < \left(\hat{\beta}_i / 2\right)$, then the estimate is statistically significant. This means that we reject the null hypothesis $(H_0 : \beta_i = 0)$ if $Se(\hat{\beta}_i) < \left(\hat{\beta}_i / 2\right)$, and this is equivalent to accepting that the true population parameter β_i is different from zero.

Note: In this test to arrive at the conclusion regarding the significance or non-significance of the OLS estimate, $\hat{\beta}$, we have implicitly used a two-tail test at 5 % level of significance.

Generally, the acceptance or rejection of the null hypothesis has a definite economic meaning. Namely, the acceptance of the null hypothesis $(H_0 : \beta_i = 0)$ 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. This is because the conducted test provided evidence that changes in X leave Y unaffected, which may mean that there is no relationship between X and Y .

In general, we may state the statistical significance of the estimates in one of the following ways:

1. The estimates are significantly different from zero.
2. The estimates are statistically significant
3. We reject the null hypothesis

Remark: *To facilitate the comparison of the standard errors of the estimates to their numerical value, it is convenient to print the standard errors in parentheses under the parameter estimates to which they refer.*

4.2.2.2 The Z – Test of the Least – Square Estimates

This test basically assumes that the parent population of the variable has standard normal distribution. In other words, the Z -test is applicable only if:

a) The distribution of the variable is normal and
b) The population variance of the variable is known or
c) The population variance is unknown; provided that the sample with which we work is sufficiently large (i.e., $n > 30$).

To decide on the rejection or acceptance of the null hypothesis, this test compares the value of Z obtained from the sample estimates with the critical values of Z at the given significance level.

Decision Rule

If the empirical value of Z falls in the critical region, we reject the null hypothesis because the probability of observing the empirical Z (if our hypothesis were true)

is very small. In other words, it is improbable that such Z would be observed if H_0 were true.

The Z – test may be outlined as follows

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

Under the usual assumptions regarding the values of U, (namely U_i is normally distributed with mean 0 and constant variance, σ_u^2, i.e., $U \sim N(0, \sigma_u^2)$) the least squares estimates $\hat{\beta}_0$ *and* $\hat{\beta}_1$ have the following normal distribution:

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma\left(\hat{\beta}_0\right)\right) \text{, where } \sigma(\hat{\beta}_0) = \sqrt{\sigma_u^2 \frac{\sum X_i^2}{n\sum x_i^2}}$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma\left(\hat{\beta}_1\right)\right) \text{, where } \sigma(\hat{\beta}_1) = \sqrt{\sigma_u^2 \frac{1}{\sum x_i^2}}$$

The normal distributions above, then, can be standardized, i.e, they can be transformed in to the units of the standard normal variable Z, which has zero mean and unit variance, $Z \cong N(0,1)$ through the transformation formula given as follows.

$$Z_i = \frac{\textit{The value of the variable} - \textit{the expected value of the variable}}{\textit{Standard deviation of the variable}} \quad \ldots\ldots\ldots\ldots\ldots\ldots 4.25$$

Since the variables of interest here are the OLS estimates of the parameters β_0 and β_1, the formula of Z in equation 4.25 becomes

$$Z = \frac{\hat{\beta}_0 - \beta_0}{\sigma(\hat{\beta}_0)} = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sigma_u^2 \sum x_i^2 / n \sum x_i^2}} \quad , \quad for \quad \hat{\beta}_0$$

And 4.26

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma_u^2 / \sum x_i^2}} \quad , \quad for \quad \hat{\beta}_1$$

After calculating the values of Z from the equations in 4.26, we need to choose the level of significance. The 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.*

In applied econometric work it has become customary to perform a two-tail test. The choice of a two tail test implies no a priori knowledge regarding the sign of the coefficient whose significance is being tested.

Our final step is to compare the calculated value of Z (obtained from equation 4.26) with its critical value for a given significance level (such as1%, 5% or 10%) and determine the critical/rejection region.

For Example, for 5 % significance level the rejection region for the null hypothesis is given as either $Z < -1.96$ or $Z > 1.96$ as shown below.

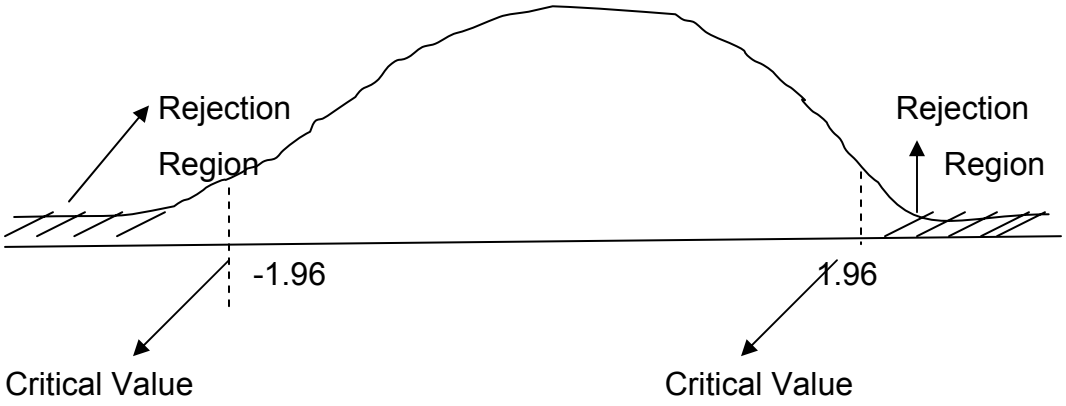


Figure 4.3 Test of Significance using Z test

In applied econometrics it has become customary to test the hypothesis that the true population parameter is zero, which is stated as

$$\Rightarrow H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

Therefore, the empirical value of Z is computed from equation 4.26 as:

$$\Rightarrow Z = \frac{\hat{\beta}_i - \beta_i}{\sigma(\hat{\beta}_i)} = \frac{\hat{\beta}_i - 0}{\sigma(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{\sigma(\hat{\beta}_i)} \quad \text{.....4.26*}$$

Thus, in the case of the test for the null hypothesis $H_0 : \beta_i = 0$ the procedure of Z test reduces to the simple step of dividing the estimated value of the parameter by its standard deviation, and comparing the resulting Z value with the theoretical/critical values of Z.

Given that for 5 % level of significance the critical value of Z is 1.96, which is approximately equal to 2, the decision rule about accepting or rejecting the null hypothesis suggests that we reject the null hypothesis if

$$Z_{calculated} > 1.96 \Rightarrow Z > 2$$

But in equation 4.26*, we have seen that

$$\Rightarrow Z = \frac{\hat{\beta}_i - \beta}{\sigma(\hat{\beta}_i)} = \frac{\hat{\beta}_i - 0}{\sigma(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{\sigma(\hat{\beta}_i)}$$

Since the decision rule states that the null hypothesis is rejected if Z is greater than 2, the above equation becomes,

$$\Rightarrow Z > 2$$

$$\Rightarrow \frac{\hat{\beta}_i}{\sigma(\hat{\beta}_i)} > 2$$

This implies that Z would be greater than 2 if and only if $\hat{\beta}_i$ is greater than twice the value of the denominator, i.e.,

$$\hat{\beta}_i > 2\sigma\left(\hat{\beta}_i\right)$$

$$\Rightarrow \sigma(\hat{\beta}_i) < \frac{1}{2}\hat{\beta}_i \text{ .....4.27}$$

This is the rule of thumb under the standard error test. Thus, the standard error test and Z – test are identical; they are two ways of saying the same thing.

Thus, the statements

1. We reject the null hypothesis when $Z > 2$ and
2. We reject the null hypothesis when $\sigma(\hat{\beta}_i) < \frac{1}{2}\hat{\beta}_i$ are two different ways of saying the same thing.

4.2.2.3 The Student's t test

The t test is one of the procedures to test whether the estimates of the parameters are statistically significant or not. Generally, this test is used when the following conditions are satisfied.

1. If the population of the variable is normal
2. If the population variance is unknown and
3. If the sample size is small ($n < 30$)

To conduct this test we use the t transformation formula. Through this formula we transform the values of any variable X in to t units in similar way to the Z transformation formula. However, the t value depends in addition on the number of degrees of freedom (df) and it includes the variance estimates S_x^2 instead of the true variance.

The t – transformation formula is given as:

$$t = \frac{X_i - \mu}{S_x} \text{ , with n – 1 degrees of freedom}$$

$$\text{Where } S_x = \sqrt{\frac{\sum\left(X_i - \overline{X}\right)^2}{n-1}}$$

By analogy, the t statistic for OLS estimates ($\hat{\beta}_i$) is obtained from the following formula.

$$t = \frac{\hat{\beta}_i - \beta_i}{\sqrt{Var\left(\hat{\beta}_i\right)}} \quad \text{.....4.28}$$

Where $n-K$ is degrees of freedom, n is the sample size and K is the number of explanatory variables.

Note: The t – distribution is always symmetric, with mean equal to zero and variance $(n\ 1)(n\ 3)$ which approaches unity when n is large.

To perform the t test, we follow the following steps

Step 1*. Define the null and alternative hypothesis*

Step 2*. Choose the desired level of significance (α)*

Step 3*. Define the number of degrees of freedom*

Step 4*. Obtain the calculated value of t from equation 4.28*

Step 5*. Obtain the critical values (define the critical region)*

Step 6*. Make decision.*

In econometrics the customary form of the null hypothesis is

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

In this case the t statistic in equation 4.28 reduces to:

$$t = \hat{\beta}_i \Big/ Se\left(\hat{\beta}_i\right) \quad \text{.....4.29}$$

Equation 4.29 states that the sample (or calculated) value of t is obtained by dividing the estimate $\hat{\beta}_i$ by its standard deviation. Then to make decision, we compare the value of t to its critical value obtained for $n-K$ degrees of freedom and specified level of significance.

Decision Rule

Reject the null hypothesis if calculated t value falls in the critical region or rejection region, i.e., reject H_0 *if* $|t\ cal| > t_{\alpha/2}$, where $t_{\alpha/2}$ is the critical value of t obtained from t-tables.

In other words, $-t_{\alpha/2} \leq t_{cal} \leq t_{\alpha/2}$ is called acceptance region.

The t test can be performed in an approximate way by simple inspection. From t table we see that the value of t changes very slowly where the degrees of freedom $n-K$ are greater than 8. For example, $t_{\alpha/2}$ value for significance level of 5% takes values between 2.30 (when $n-K=8$) and 1.96 (when $n-K=\infty$).

It is clear from this that the change from 2.30 to 1.96 is obviously very slow. Consequently, we can ignore the degrees of freedom when $n-K>8$ and say that the critical value of t is 2. Thus, the two tail t test at 5 % significance level reduces to the following rule.

Finally, if the observed t value is greater than 2 (or less than –2), we reject the null hypothesis.

But, given that $t_{cal} = \hat{\beta}_i / S(\hat{\beta}_i)$, the sample value of *t* would be greater than 2 if the estimate is greater than twice its standard deviation.

$$\Rightarrow t_{cal} > 2 \text{ , } if \ \ \hat{\beta}_i > 2\, S\left(\hat{\beta}_i\right)$$

Or

$$S\left(\hat{\beta}_i\right) < 1/2\ \hat{\beta}_i$$

Thus, we see that the decision rules:

a) We reject H_0 if $t_{cal} > t_{\frac{\alpha}{2}}$ and

b) We reject H_0 if $S(\hat{\beta}_i) < \hat{\beta}/2$ are essentially the same.

Note: This is an approximation to the formal t test discussed above and it is valid only for $n - K > 8$

4.2.2 Confidence Interval Approach to Hypotheses Testing

4.2.2.1 Confidence Intervals for β_0 and β_1

In the preceding discussion we tried to develop a decision rule to reject the null hypothesis. Now, however, the question is about the implication of the decision. What does the rejection of the null hypothesis imply?

For example, rejection of the H_0 does not mean that our OLS estimate is the correct estimate of the true population parameter β_i, but it simply means that our estimate has come from a sample drawn from a population whose parameter β_i is different from zero. Therefore, in order to define how close the true parameter lines to the estimate, we must construct confidence intervals for the true parameter.

In other words we must establish limiting values around the estimate within which the true parameter is expected to lie with a certain “degree of confidence”, and this implies that with a given probability the population parameter will be with in the defined confidence interval.

To establish confidence interval for the parameters, we choose a probability in advance and refer to it as the confidence level (or confidence coefficient). It is customary in econometrics to choose the 95 % confidence level. This means that in repeated sampling the confidence limits, computed from the samples, would

include the true population parameter in 95 percent of the cases. In the other 5 percent of the cases the population parameter will fall outside the confidence limits. In other words, it means that given the confidence coefficient 95 % in the long run in 95 out of 100 cases intervals like the one we constructed following the above procedure will contain the true population parameter.

4.2.2.2 Confidence interval from standard normal distribution

It is shown in previous chapters that following the normal distribution assumption of U_i, the OLS estimators $\hat{\beta}_0$ *and* $\hat{\beta}_1$ are normally distributed. Thus the standard normal variable Z can be obtained as:

$$Z : \frac{\hat{\beta}_i - \beta_i}{\sigma\left(\hat{\beta}_i\right)} \quad \text{.....4.30}$$

Then, we choose a confidence coefficient, say 95% and obtain the critical value of Z. Finally, find the probability that the value of Z lies between $-Z_{\alpha/2}$ *and* $Z_{\alpha/2}$. For example, for a confidence coefficient of 95% the critical value of Z is 1.96 and the probability is 0.95, as shown below.

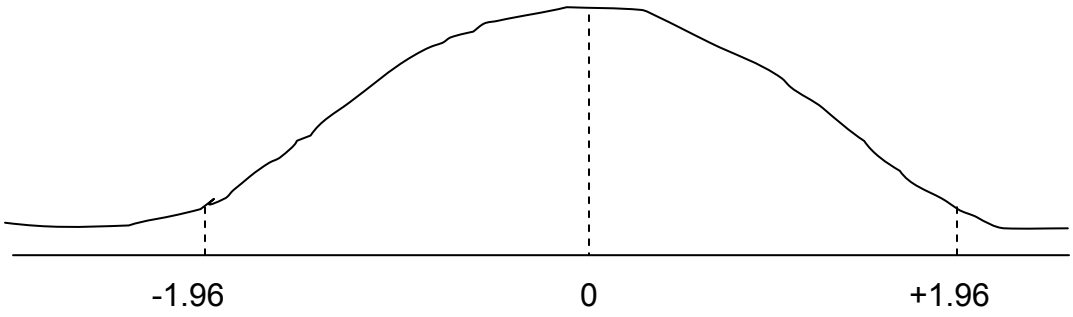


Figure 4.4 Critical Values of *Z* for Confidence Level of 90%

Thus, the probability with which Z lies between -1.96 and 1.96 can be written as:

$$\mathrm{P}\,(-1.96 < \mathrm{Z}\,1.96) = 0.95 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 4.31$$

Substituting equation 4.30 into equation 4.31 for Z, we obtain

$$P\left(-1.96 < \frac{\left(\hat{\beta}_i - \beta_i\right)}{\sigma\left(\hat{\beta}_i\right)} < 1.96\right) = 0.95$$

Multiplying all the terms in the bracket by $\sigma\left(\hat{\beta}_i\right)$, we get

$$P\left\lfloor -1.96\,\sigma\left(\hat{\beta}_i\right) < \hat{\beta}_i - \beta_i < 1.96\,\sigma\left(\hat{\beta}_i\right)\right\rfloor = 0.95$$

Subtracting $\hat{\beta}_i$ from all the terms in the bracket, we get

$$P\left\lfloor -1.96\,\sigma\left(\hat{\beta}_i\right) - \hat{\beta}_i < -\beta_i < 1.96\,\sigma\left(\hat{\beta}_i\right) - \hat{\beta}_i\right\rfloor = 0.95 \ldots\ldots\ldots\ldots 4.32$$

Finally, multiplying all the terms in the bracket by -1 yields,

$$\Rightarrow P\left\lfloor \hat{\beta}_i - 1.96\,\sigma\left(\hat{\beta}_i\right) < \beta_i < \hat{\beta}_i + 1.96\,\sigma\left(\hat{\beta}_i\right)\right\rfloor = 0.95 \ldots\ldots\ldots\ldots\ldots\ldots\ldots 4.33$$

From equation 4.33, the lower limit is

$$\hat{\beta}_i - 1.96\,\sigma\left(\hat{\beta}_i\right) \textit{ and}$$

The upper limit is

$$\hat{\beta}_i + 1.96\,\sigma\left(\hat{\beta}_i\right)$$

The meaning of the confidence interval in equation 4.33 is that the unknown population parameter, β_i, will lie within the defined limits 95 times out of 100 cases. In general the confidence interval for β_i is given as:

$$\hat{\beta}_i \pm Z_{\alpha/2}\ \sigma(\hat{\beta}_i)$$

4.2.2.3 Confidence interval from the student's t distribution

We have discussed above how to use normal distribution (and hence Z statistic) to make probability statements about β_i provided that the true population variance σ_u^2 is known. But in practice it is rarely known and is determined by the unbiased estimator $\hat{\sigma}_u^2$, and makes the use of Z – statistic less likely. In this case we use t – statistic to develop the confidence intervals of the true parameters using OLS estimates.

The procedure of constructing a confidence interval with the t – distribution is similar to the one outlined above, with the exception that in the latter case we take into account the degrees of freedom (df).

As we have already discussed, the t – statistic for $\hat{\beta}_i$ is given as

$$t = \frac{\hat{\beta}_i - \beta_i}{Se(\hat{\beta}_i)}$$

Then, to construct confidence interval for a parameter, we follow the following steps.

Step 1. Choose a confidence level, say 95 %

Step 2. Find the t critical value with $n-K$ degrees of freedom and use its implication for the probability of t lying between $-t_{\alpha/2}$ *and* $t_{\alpha/2}$. This probability for two tail t test is given as

$$P\left(-t_{0.025} \leq t < t_{0.025}\right) = 0.95$$

$$\Rightarrow P\left(-t_{0.25} < \frac{\hat{\beta}_i - \beta_i}{Se\left(\hat{\beta}_i\right)} < t_{0.025}\right) = 0.95$$

Following the procedures we followed in the case of Z statistics, we obtain the following.

$$P\left(\hat{\beta}_i - t_{0.025}\ Se\left(\hat{\beta}_i\right) < \beta_i < \hat{\beta}_i + t_{0.025} Se\left(\hat{\beta}_i\right)\right) = 0.95 \ldots\ldots\ldots\ldots\ldots\ldots 4.34$$

Thus, the lower limit of β_i is

$$\hat{\beta}_i - t_{0.025}\ Se\left(\hat{\beta}_i\right) \text{ and}$$

the upper limit is

$$\hat{\beta}_i + t_{0.025}\ Se\left(\hat{\beta}_i\right).$$

In compact form equation 3.34 can be written as

$$\hat{\beta}_i \pm t_{0.25}\ Se\left(\hat{\beta}_i\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 4.35$$

Expending this to any confidence level $\left[100\ (1-\alpha)\ \%\right]$, the confidence interval in equation 4.35 is written as

$$\hat{\beta}_i \pm t_{\alpha/2}\ Se\left(\hat{\beta}_i\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 4.36$$

Where α is the significance level.

Therefore, the confidence interval for β_0, at α significance level is

$\hat{\beta}_0 \pm t_{\alpha/2} Se(\hat{\beta}_0)$ *and that for* β_1 is $\hat{\beta}_1 \pm t_{\alpha/2} Se(\hat{\beta}_1)$

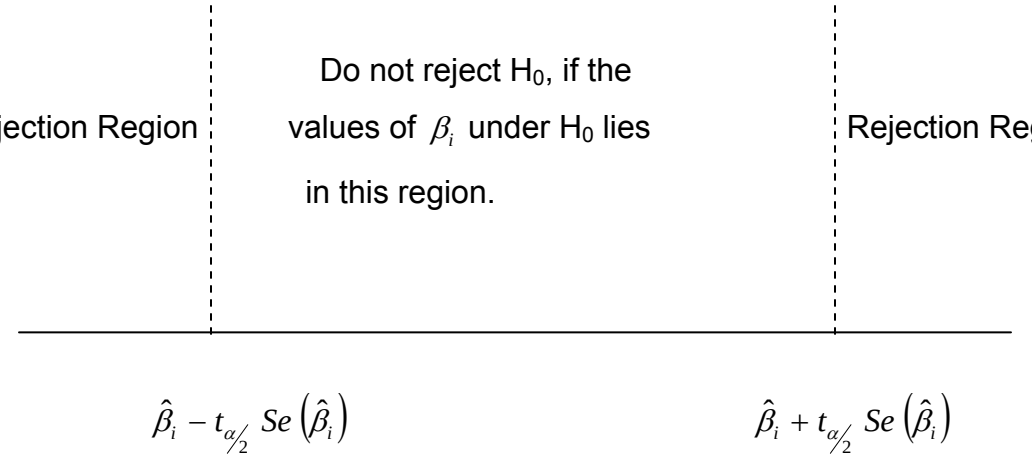
Finally, we can use these confidence intervals to test hypotheses about the parameters (β_i)

To use these confidence intervals for hypothesis testing purpose, we use the following decision rule.

Decision Rule

After constricting a $100(1-\alpha)\%$ confidence interval for β_i reject the null hypothesis, if β_i under the null hypothesis falls out side the confidence interval.

In the figure given below, we have two rejection regions: the area below the lower limit and the area above the upper limit.



Note that in statistics, when we reject the null hypothesis, we say that our finding is statistically significant.

Exercise

1. Suppose that a graduating class student from the Department of Economics in Arbaminch University has collected annual data on exchange rate and relative prices in U.S. and Ethiopia from 1990 to 2004. Finally, he obtained the following regression results, where *Y* = exchange rate of the Ethiopian birr to the U.S. dollar (Birr/$) and *X* = ratio of the U.S. consumer price index to the Ethiopian consumer price index; that is, *X* represents the relative prices in the two countries:

$$\hat{Y}_t = 6.682 - 4.318\, X_t$$

Se = (1.22) (1.333) $r^2 = 0.528$

i. Interpret his regression result. How would you interpret r^2?
ii. Does the negative value of X_t make economic sense? What is the underlying economic theory?
iii. Suppose he were to redefine *X* as the ratio of Ethiopian CPI to the U.S. CPI. Would that change the sign of *X*? Why?

1. Given that that the researcher in question #1 used the following model

$$Y_t = \beta_0 + \beta_1 X_t + U_t$$

i. Test the hypothesis that the coefficient of X_t is statistically insignificant
ii. Construct 90% confidence interval for β_1 , and use the your confidence interval to test the hypothesis that $\beta_1 = 0$

2. A store manager selling TV sets observes the following sales on ten different days, where Y is the number of TV sets sold and X is the number of sales representatives.

Y	3	6	10	5	10	12	5	10	10	8
X	1	1	1	2	2	2	3	3	3	2

i. Find the variances of the OLS estimates of the coefficients in the regression of Y on X

ii. Calculate r^2 and interpret its value

iii. Construct 90% confidence interval for the coefficient of X

iv. Test the hypothesis that the coefficient of X is not statistically significant

3. Assume that two graduating class students from Economics Department at Arba Minch University have conducted independent researches to analyze the relationship between agricultural output (Y) and labor input (L) in Gamo Goffa Zone. The two researchers used data on Y and L for the time period of 1960–1990 and obtained the following results.

Student I:

$$\hat{Y}_t = \underset{(23.55)}{-338} + \underset{(0.16)}{3.05}\, L_t \qquad r^2 = 0.958$$

Student II

$$\ln \hat{Y}_t = \underset{(0.266)}{-5.48} \underset{(0.012)}{-2.08} \ln L_t \qquad r^2 = 0.985$$

Given that the numbers in the parentheses are t ratios

i. Construct 99% confidence interval for the coefficient of L for the two models, and use the intervals to verify whether the coefficients of L are statistically significant or not.

ii. Test the hypothesis that the coefficients of L are not significant at 1% significance level

4. Show that the regression of Y on X produces the same results as the regression of X on Y if $r^2 = 1$

CHAPTER FIVE

MULTIPLE LINEAR REGRESSION MODELS

Overview

In the previous chapters, we have discussed simple linear regression models, where we had only one explanatory variable that explains the change in the dependent variable. Since these one explanatory variable models are inadequate to explain the real world economic relationships, in this chapter we will turn our attention to multiple linear regression models where the dependent variable is influenced by more than one explanatory variable. For simplicity purpose, we start our analysis of multiple regression models with three variables model or two explanatory variables model, and in subsequent sections of the chapter we will deal with models involving more than three variables.

At the end of this chapter students will be able to:

- Know multiple regression models and differentiate between multiple regression models and simple regression models
- Obtain OLS estimates for the coefficients of multiple regression models
- Estimate the coefficient of determination, R^2, for multiple regression models
- Differentiate between simple correlation coefficients, partial correlation coefficients and multiple correlation coefficients
- Estimate the means and variances of OLS estimates of the coefficients of multiple regression models
- Conduct significance tests of individual estimates and test the overall significance of multiple regression models
- Interpret the estimates of the coefficients of multiple regression model
- Apply matrix algebra to obtain the OLS estimates of the coefficients of multiple regression models, and to obtain the means and variances of the estimates and the coefficient of determination in multiple regression models.

5.1 The Three Variables Models

These models are the simplest possible multiple regression models and they have only two explanatory variables. To illustrate these models, let us start with the simplest form of the theory of demand.

Economic theory postulates that quantity demanded (say Y) of a given commodity depends on its price (X_1) and consumers' income (X_2) . This shows that

$$Y = f(X_1, X_2) \text{.....5.1}$$

Given that economic theory does not specify the mathematical form of the demand function, we start our investigation by assuming that the relationship between Y, and X_1 and X_2 is linear. Therefore, the mathematical model of equation 5.1 will be:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \text{ .....5.2}$$

Equation 5.2 shows that the relationship between Y, and the explanatory variables, X_1 and X_2, is exact in the sense that the variations in the quantity demanded of Y are fully explained by changes in price and income (X_1 and X_2).

If this form were true, any observation on Y, X_1 and X_2 would determine a point which would lie on a plane. However, if we gather observations on these variables and plot them on a diagram, we will observe that not all of them lie on a plane: some will lie on it, but others will lie above or below it. This scatter is due to the reasons we discussed in chapter 2. Hence, we need to take the influence of such factors into account by introducing a random variable U to equation 5.2. Then, it becomes stochastic and yields an econometric model given as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i \text{ ..... 5.3}$$

On a priori grounds, we would expect the coefficient β_1 to have a negative sign, given **the law of demand.** Furthermore, since for normal commodities the quantity demanded changes in the same direction as income, β_2 is expected to be positive.

To complete the specification of our model in equation 5.3, we need some assumptions about the random variable U. The assumptions used here are the same as those used in the simple linear regression models.

5.2 Estimation of the Parameters of Multiple Linear Regression Model

Having specified our model as in equation 5.3, now it is time to use sample observations on Y, X_1 and X_2, and obtain estimates of the true parameters β_0 , β_1 *and* β_2 . In other words, in this part we will learn how to use the sample regression function (SRF) to estimate the population regression function (PRF).

Noting that equation 5.3 is known as population regression function, SRF is given as:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{u}_i \quad \text{.....5.4}$$

where $\hat{u}_1$ is the sample estimate of U_i.

By rearranging equation 5.4, $\hat{u}_i$ is given as

$$\hat{u}_i = Y_i - \hat{Y}_i \quad \text{..... 5.5}$$

Where, $\hat{Y}_i = \hat{\beta}_0 + \beta_1 X_{1i} + \hat{\beta}_2 X_{2i}$

Then, as we have already discussed in chapter 2, OLS estimates of the PRF given in equation 5.3 would be obtained by minimizing the sum of squared residuals, $\hat{u}_i$. In other words, the OLS estimates would be obtained by choosing the values of the unknown parameters that minimize the deviations of the estimated values from the actual observations.

Symbolically, the OLS estimates are obtained from the following minimization problem.

$$Min\sum_{i=1}^{n}\hat{u}_i^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

$$= \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}\right)^2 \text{ ……..……5.6}$$

As you may remember from **Calculus for Economists**, for $\sum\hat{u}_i^2$ to attain its minimum the following three conditions, called First-Order-Conditions must be fulfilled.

i) $\frac{\partial\sum\hat{u}_i^2}{\partial\hat{\beta}_0} = 0$ ……..5.7

ii) $\frac{\partial\sum\hat{u}_i^2}{\partial\hat{\beta}_1} = 0$ ……..5.8

iii) $\frac{\partial\sum\hat{u}_i^2}{\partial\hat{\beta}_2} = 0$ ……5.9

Performing the partial differentiations to equations 5.7 to 5.9, we get the following system of three normal equations in three unknowns, $\hat{\beta}_0$, $\hat{\beta}_1$ *and* $\hat{\beta}_2$

$$\sum Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum X_{1i} + \hat{\beta}_2 \sum X_{2i} \text{ ……..... 5.10}$$

$$\sum X_{1i} X_i = \hat{\beta}_0 \sum X_{1i} + \hat{\beta}_1 \sum X_{i\,i}^2 + \hat{\beta}_2 \sum X_{1i} X_{2i} \text{ ……5.11}$$

$$\sum X_{2i} Y_i = \hat{\beta}_0 \sum X_{2i} + \hat{\beta}_1 \sum X_{1i} X_{2i} + \hat{\beta}_2 \sum X_{2\,i}^2 \text{ …. 5.12}$$

Equations 5.10, 5.11 and 5.12 are called the OLS normal equations of a model with three variables.

Dear readers if you face any difficulty in performing partial differentiations to equations 5.7 to 5.9 to obtain the above normal equations, please refer to the procedures we followed in chapter 2 to get normal equations of simple linear regression models.

Then, by solving simultaneously (or by using *Cramer's Rule*) the three normal equations defined in equations 5.10, 5.11 and 5.12, we get the formula for OLS estimates $\hat{\beta}_0$, $\hat{\beta}_1$ *and* $\hat{\beta}_2$, of the parameters β_0 , β_1 *and* β_2 . These procedures produce the formulae of the estimates in terms of actual observations of the dependent and explanatory variables. However, it is customary to express these formulae in terms of the deviations of sample observations of the variables from their mean values. Hence, in this module, hoping that interested students can drive the formula of the OLS estimates in terms of actual observations using the procedures we used in chapter 2, the formulae of the estimates here are developed in deviation forms.

To express the formula of the OLS estimates in deviation form (deviation of sample observations of the variables from their mean), we proceed as follows.

Now, recall from equation 5.5 that

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

However, we have seen in chapter 4 that

$$Y_i = y_i + \overline{Y} \quad \text{.....5.13}$$

and

$$\hat{Y}_i = \hat{y}_i + \overline{Y} \quad \text{.....5.14}$$

$$\Rightarrow \hat{u}_i = y_i + \overline{Y} - \left(\hat{y}_i + \overline{Y}\right)$$

$$= y_i + \overline{Y} - \hat{y}_i - \overline{Y}$$

$$\Rightarrow \hat{u}_i = y_i - \hat{y}_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 5.15$$

From equation 5.14, we get

$$\hat{y}_i = \hat{Y}_i - \overline{Y}. \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 5.16$$

By dividing trough equation 5.10 by n, we obtain

$$\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{X}_1, + \hat{\beta}_2 \overline{X}_2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 5.17$$

Substituting equations 5.17 into equation 5.16, we get

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} - \hat{\beta}_0 - \hat{\beta}_1 \overline{X}_1 - \hat{\beta}_2 \overline{X}_2$$

$$= \hat{\beta}_1 \left(X_{1i} - \overline{X}_1\right) + \hat{\beta}_2 \left(X_{2i} - \overline{X}_2\right)$$

$$= \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 5.18$$

Where, $x_{1i} = X_{1i} - \overline{X}_1$ and $x_{2i} = X_{2i} - \overline{X}_2$

Then, by substituting equations 5.18 in to equation 5.15, we obtain

$$\sum \hat{u}_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum \left(y_i - \left[\hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}\right]\right)^2$$

$$= \sum \left(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}\right)^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots 5.19$$

As we discussed in chapter 2, the principle of OLS suggests that to obtain OLS estimates, we have to minimize the sum of the squares of the deviations in equation 5.19, which requires the following conditions to hold.

i) $\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_1} = 0$ *and*

ii) $\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_2} = 0$

Then, performing the above partial differentiations will yield the OLS normal equations of a three variable model in deviations form, and these equations are given as

$$\sum x_{1i}\, y_i = \beta_1 \sum x_{1\,i}^2 + \hat{\beta}_2 \sum x_{1\,i}\, x_{2\,i} \quad \ldots\ldots\ldots\ldots\ldots\ldots \text{5.20}$$

$$\sum x_{2\,i}\, y_i = \hat{\beta}_1 \sum x_{1\,i}\, x_{2\,i} + \hat{\beta}_2 \sum x_{2\,i}^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots \text{5.21}$$

Finally, by using **Cramer's Rule**, the formula of $\hat{\beta}_0$, $\hat{\beta}_1$ *and* $\hat{\beta}_2$ in which the variables are expressed in terms of deviations from their means, will be given as

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\, \overline{X}_1 - \hat{\beta}_2\, \overline{X}_2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 5.22$$

$$\hat{\beta}_1 = \frac{\left(\sum x_{1i}\, y_i\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{2i}\, y_i\right)\left(\sum x_{1i}\, x_{2i}\right)}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i}\, x_{2i}\right)^2} \ldots\ldots\ldots\ldots\ldots\ldots 5.23$$

$$\hat{\beta}_2 = \frac{\left(\sum x_{2i}\, y_i\right)\left(x_{1i}^2\right) - \left(\sum x_{1i}\, y_i\right)\left(\sum x_{1i}\, x_{2i}\right)}{\left(\sum x_{1\,i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i}\, x_{2i}\right)^2} \ldots\ldots\ldots\ldots\ldots\ldots 5.24$$

Remark

From equations 5.20 and 5.21 we can establish the following important formulae.

1. $\sum \hat{u}_i\, x_{1\,i} = 0$

2). $\sum \hat{u}_i \, x_{2i} \;\; = \; 0$

Proof

Recall from equation 5.15 that

$$\hat{u}_i \; = \; y_i \; - \; \hat{y}_i$$

$$\Rightarrow y_i \; = \; \hat{y}_i \; + \; \hat{u}_i$$

But from equation 5.18, we know that $\hat{y}_i = \hat{\beta}_1 \, x_{1i} \; + \; \hat{\beta}_2 \, x_{2i}$

$$\Rightarrow y_i \; = \hat{\beta}_1 \; x_{1i} \; + \hat{\beta}_2 \;\; x_{2i} \; + \hat{u}_i \text{……………………………………….5.25}$$

Multiplying both sides of equation 5.25 by x_{1i} and then taking the sum of the results over the number of sample observations, we get

$$\sum y_i \; x_{1i} \;\; = \; \hat{\beta}_1 \sum x_{1i}^2 \; + \; \hat{\beta}_2 \sum x_{1i} \;\; x_{2i} \; + \; \sum \hat{u}_i \;\; x_{1i} \text{ …………………...5.26}$$

But normal equation 5.20 shows that

$$\sum y_i \; x_{1i} \;\; = \hat{\beta}_1 \sum x^2{}_{1i} \;\; + \; \hat{\beta}_2 \sum x_{1i} \;\; x_{2i} \; + \sum \hat{u}_i^2 \;\; x_{1i}$$

Thus, for this to be true, in equation 5.26

$$\sum \hat{u}_i^2 x_{1i} \;\; = \; 0 \text{ ……………………………………………………….5.27}$$

Similarly, from normal equation 5.21, we can show that

$$\sum \hat{u}_i^2 x_{2i} \;\; = \; 0 \text{ ………………………………………………………….5.28}$$

5.3 The Multiple Coefficient of Determination, R^2

In the simple linear regression model, we have seen that r^2 measures the goodness of fit of the regression equation; that is, it gives the percentage of the total variation in the dependent variable Y explained by the explanatory variable X. This definition of r^2 can be easily extended to multiple linear regression models. Thus, if we want to know the proportion of the variation in Y explained by explanatory variables jointly in multiple linear regression models, we use the multiple linear regression models country part of r^2 ;i.e., R^2

To drive R^2 in the three explanatory variables model, recall that by definition

$$R^2 = \frac{ESS}{TSS}$$

$$\Rightarrow R^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2}$$

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

As we have shown in chapter 4 equation 4.18, R^2 (in terms of regression residuals) can be defined as

$$R^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2}$$

$$= \frac{\sum y_i^2 - \sum \hat{u}_i^2}{\sum y_i^2} \quad \text{.....5.29}$$

Now, let us express $\sum \hat{u}_i^2$ in equation 5.29 in its expanded form by using elementary algebra as follows.

$$\Rightarrow \sum \hat{u}_i^2 = \sum (\hat{u}_i)(\hat{u}_i) \quad \text{.....5.30}$$

Then, substituting equation 5.15 into 5.30, we get

$$\sum \hat{u}_i^2 = \sum \hat{u}_i \left(y_i - \hat{y}_i \right) \quad \text{.....5.31}$$

Again, substituting equation 5.18 in place of $\hat{y}_i$ into 5.31, yields

$$\sum \hat{u}_i^2 = \sum \hat{u}_i \left[y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} \right]$$

$$= \sum \hat{u}_i y_i - \hat{\beta}_1 \sum \hat{u}_i x_{1i} - \hat{\beta}_2 \sum \hat{u}_i x_{2i} \quad \text{.....5.32}$$

Since from equations 5.27 and 5.28, $\sum \hat{u}_i \; x_{1i} = 0$ and $\sum \hat{u}_i \; x_{2i} = 0$, then equation 5.32 becomes

$$\sum \hat{u}_i^2 = \sum \hat{u}_i \; y_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{5.33}$$

Now, let us substitute equation 5.15 into 5.33 in place $\hat{u}_i$

$$\Rightarrow \sum \hat{u}_i^2 = \sum \left(y_i - \hat{y}_i\right) \; y_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{5.34}$$

By plugging the value of $\hat{y}_i$ from equation 5.18 into equation 5.34, we get

$$\sum \hat{u}_i^2 = \sum \left(y_i - \hat{\beta}_1 \; x_{1\,i} - \hat{\beta}_2 \; x_{2\,i}\right) y_i$$

$$= \sum y_i^2 - \hat{\beta}_1 \sum x_{1\,i} \; y_i - \hat{\beta}_2 \sum x_{2i} \, y_i \quad \ldots\ldots\ldots\ldots\ldots\ldots \text{5.35}$$

Further, substituting equation 5.35 in the formula of R^2 in equation 5.29, we obtain:

$$R^2 = \frac{\sum y_i^2 - \left(\sum y_i^2 - \hat{\beta}_1 \sum x_{1\,i} \; y_i - \hat{\beta}_2 \sum x_{2i} \, y_i\right)}{\sum y_i^2}$$

$$= \frac{\sum y_i^2 - \sum y_i^2 + \hat{\beta}_1 \sum y_i \; x_{1i} + \hat{\beta}_2 \sum y_i \; x_{2i}}{\sum y_i^2}$$

$$\Rightarrow \quad R^2 = \frac{\hat{\beta}_1 \sum y_i \; x_{1i} + \hat{\beta}_2 \sum y_i \; x_{2i}}{\sum y_i^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{5.36}$$

Note: Equation 5.36 is the formula of R^2 for the two explanatory (or three) variables models.

As we discussed in chapter 4, R^2 measures the proportion of the variation in the dependent variable explained by explanatory variables jointly and its value

ranges from 0 to 1 (i.e., $0 \leq R^2 \leq 1$), where $R^2 = 1$ implies that the fitted regression line explains 100% of the variations in the dependent variable, while $R^2 = 0$ implies that the model does not explain any of the variations in the dependent variable.

Note: The higher the value of R^2, the greater would be the percentage of the variation of Y explained by the regression plane, i.e., the better the "goodness of fit" of the regression plane to the sample observations.

The general formula for R^2 can be developed by inspecting the formula of R^2 for two-variable and three-variable models.

1. **A model with one explanatory variable (some times known as two-variables model).**

The model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + U_i$$

The formula of R^2 (see equation 4.17a)

$$R^2_{Y,x_1} = \frac{\hat{\beta}_1 \sum y_i x_{1i}}{\sum y_i^2} \quad \text{……..} \text{……} \text{5.37}$$

2. **A model with two explanatory variables (some times called tree-variables model)**

The model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i$$

The formula of R^2 (see equation 5.36)

$$R^2_{Y,x1,x2} = \frac{\hat{\beta}_1 \sum y_i x_{1i} + \hat{\beta}_2 \sum y_i x_{2i}}{\sum y_i^2} \quad \text{.....5.36a}$$

Dear readers, what is the difference between the formula of R^2 in the above two models? What pattern have you inspected from equations 5.37 and 5.36a?

Equation 5.37 shows that the formula of R^2 involves an additional term $\hat{\beta}_2 \sum y_i x_{2i}$ in the numerator than that in equation 5.37. This implies that R^2 for a model with one additional explanatory variable, X_{2i}, has one extra term formed from the product of $\hat{\beta}_2$ and $\sum y_i x_{2i}$.

Generally, by inspecting the formula of R^2 from the two models given above, we see that for each additional explanatory variable, the formula of R^2 includes an additional term in the numerator, where the additional term is formed from the product of the estimate of the parameter corresponding to the new variable, and the sum of the product of the deviations of the new variable and the dependent variable.

Consequently, for K-variables model, the formula of R^2 becomes:

$$R^2 = \frac{\hat{\beta}_1 \sum y_i x_{1i} + \hat{\beta}_2 y_i x_{2i} + ... + \hat{\beta}_K \sum y_i x_{Ki}}{\sum y_i^2} \text{ ……5.38}$$

The important property of R^2 is that it is a non-decreasing function of the number of explanatory variables present in the model, i.e., as the number the number of explanatory variables increases, the value of R^2 almost invariably increases and never decreases.

To illustrate this, let us start from the origin of the formula of R^2.

Recall from chapter 4 equation 4.18 that

$$R^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2} \text{ ……..5.39}$$

Note from equation 5.39 that while $\sum y_i^2$ in the formula of R^2 is independent of the number of explanatory variables (because it is simply $\sum \left(Y_i - \overline{Y}\right)^2$), $\sum \hat{u}_i^2$ depends on the number of regressors present in the model. Intuitively, it is clear that as the number of explanatory variables increases, the residual term, $\sum \hat{u}_i^2$, is likely to decrease (at least it will not increase); hence R^2 will increase.

This suggests that in comparing two regression models with the same dependent variable but differing number of explanatory variables, one should be very wary of choosing the model with the highest R^2.

Generally, to compare two R^2 terms, one must take into account the number of X-variables present in the model and adjust R^2 for degrees of freedom. This can

be readily done if we consider an alternative coefficient of determination: adjusted- R^2 (symbolically $\overline{R^2}$), which is obtained as follows:

$$\overline{R}^2 = 1 - \frac{\sum \hat{u}_i^2 \Big/ (n-k)}{\sum y_i^2 \Big/ n-1} \quad \text{.....5.40}$$

$$\Rightarrow \overline{R}^2 = 1 - \frac{\hat{\sigma}_u^2}{\sigma_y^2} \quad , \quad because \quad \hat{\sigma}_u^2 = \frac{\sum \hat{u}_i^2}{n-k} \quad and \quad \hat{\sigma}_y^2 = \frac{\sum y_i^2}{n-1}$$

Equation 5.40 can also be written as

$$\overline{R}^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2} \left(\frac{n-1}{n-k}\right) \text{.....5.41}$$

Note: The term adjusted means that $\overline{R}^2$ is obtained by adjusting R^2 for the degrees of freedom (df) associated with the sums of squares $\left(\sum \hat{u}_i^2 \ and \ \sum y_i^2\right)$ entering equation 5.39.

To see the relationship between $\overline{R}^2$ and simple R^2, consider from equation 5.39 that

$$R^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2}$$

$$\Rightarrow \frac{\sum \hat{u}_i^2}{\sum y_i^2} = -\left(R^2 - 1\right)$$

$$\Rightarrow \frac{\sum \hat{u}_i^2}{\sum y_i^2} = 1 - R^2 \quad \text{..... ...5.42}$$

Substituting equation 5.42 into equation 5.41, we get:

$$\overline{R}^2 = 1 - \left[\left(1 - R^2\right)\left(\frac{n-1}{n-k}\right)\right] \text{.....5.43}$$

Note:

1) *For* $K > 1$, $\overline{R}^2 < R^2$

2) $\overline{R}^2$ can be negative.

However, in case it turns out to be negative in any application, its value is taken as zero.

Dear readers, which R^2 should one use in practice then? Why?

Generally, there is no consensus on the answer to this question. But the advice is to treat $\overline{R}^2$ as just another summery statistic besides the simple R^2.

Some times researchers play the game of maximizing $\overline{R}^2$. In other words, they choose the model that gives the highest $\overline{R}^2$. But this may be dangerous because in regression analysis our objective is not to obtain a high $\overline{R}^2$ per se, but rather to obtain dependable estimates and draw statistical inference about them. Thus, emphasis should be given to this objective and not to maximization of $\overline{R}^2$.

5.4 Simple, Partial and Multiple Correlation Coefficients

The correlation coefficient in general measures the degree of linear association between two variables. For the three variable models we can compute three correlation coefficients: r_{y,x_1}, r_{y,x_2} *and* r_{x_1,x_2}. These correlation coefficients are called gross or simple correlation coefficients or coefficients of zero order.

But the question, then, is: *Does the simple correlation coefficient, say* r_{y,x_1} *, measure the "true" degree of (linear) association between Y and* X_1 *, when a second explanatory variable* X_2 *is associated with both Y and* X_1 *?*

To answer this question, suppose the true regression model is:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + u_i \quad \text{.....5.44}$$

But, suppose for some reason we omit X_2 from the model and regress Y on X_1 as follows:

$$Y_i = \beta_0 + \beta_1 X_{1i} + v_i \quad \text{.....5.45}$$

Then, we can put the above question in other words as: *Will* β_1 *in equation 5.45 be equal to the true coefficient* b_1 *in equation 5.44?*

The answer is no. In general, r_{y,x_1} is not likely to reflect the true degree of association between Y and X_1 in the presence of X_2. It is likely to give false impression of the nature of association between Y and X_1. Therefore, what we need is a correlation coefficient that is independent of the influence, if any, of X_2 on X_1 and Y. Such a correlation coefficient is known as partial correlation coefficient (or first order correlation coefficient). For example, $r_{y,x_1.x_2}$ is the partial correlation coefficient between Y and X_1, holding X_2 constant.

Partial correlation coefficients can be derived from simple correlation coefficients as follows.

$$r_{y,x_1.x_2} = \frac{r_{y,x_1} - \left(r_{y,x_2}\right)\left(r_{x_1,x_2}\right)}{\sqrt{\left(1 - r^2{}_{y,x_2}\right)\left(1 - r^2{}_{x_1,x_2}\right)}} \quad \text{.....5.46}$$

$$r_{y,x_2.x_1} = \frac{r_{y,x_2} - \left(r_{y,x_1}\right)\left(r_{x_1,x_2}\right)}{\sqrt{\left(1 - r^2{}_{y,\,x_1}\right)\left(1 - r^2{}_{x_1\,,x_2}\right)}} \quad \text{..... .....5.47}$$

$$r_{x_1\,,\,x_2.y} = \frac{r_{x_1\,,\,x_2} - \left(r_{x_1\,,\,y}\right)\left(r_{x_2\,,\,y}\right)}{\sqrt{\left(1 - r^2{}_{y\,,\,x_1}\right)\left(1 - r^2{}_{y\,,\,x_2}\right)}} \quad \text{.....5.48}$$

The correlation coefficients given in equations 5.46 to 5.48 are also called first order correlation coefficients.

Note that by order of the correlation coefficients we mean the number of secondary variables or the number of variables held constant when we are analyzing the effect of one variable on some other.

In one explanatory variable model, the simple r has a straight forward meaning. It measures the degree of (linear) association (and not causation) between the dependent variable and explanatory variable. But once we go beyond the one explanatory model, we need to pay due attention to the interpretation of the simple correlation coefficients.

From the correlation coefficients in equations 5.46 to 5.48, we can observe the following properties of partial correlation coefficients.

i) Even if $r_{y,x_1} = 0$, $r_{y,x_1\,.\,x_2}$ will not be zero unless r_{y,x_2} *or* $r_{x_1\,,\,x_2}$ *or* both are zero.

ii) The terms $r_{y,x_1.x_2}$ *and* r_{y,x_1} need not have the same sign.

iii) In single explanatory variable models, we have seen that the value of r^2 lies between 0 and 1. Similarly, the values of the squared partial correlation coefficients lie between 0 and 1.

iv) The fact that $r_{y,x_1} = r_{x_1,\,x_2} = 0$, does not mean that $r_{y,x_2} = 0$

v) If $r_{y,x_1} = 0$, *and* r_{y,x_2} *and* $r_{x_1,\,x_2}$ are non-zero, then

a) $r_{y,x_1\,.x_2}$ will be negative, if they are of the same sign.

b) $r_{y,x_1\,.\,x_2}$ will be positive, if they are of opposite sign.

Example

Let Y denote crop yield, X_1 represent rain fall and X_2 represent temperature. Suppose further that there is no association between Y and X_1 (i.e. $r_{y,x_1} = 0$); the associations between Y and X_2, and X_1 and X_2 are positive and negative respectively.

Then as equation 5.59 shows, $r_{y,x_1\,.x_2}$ will be positive, that means, holding temperature constant, there will be positive association between crop yield and rainfall. This seemingly paradoxical result, however, is not surprising. Since temperature, X_2, affects both Y and X_1, in order to find out the net relationship between rain fall and crop yield, we need to remove the effect of the "nuisance" variable, X_2. This example shows how one might be misled by the simple coefficients of correlation.

Note: $r_{y\,,\,x_1\,.\,x_2}$ is called the coefficient of partial determination and is interpreted as the proportion of the variation in Y not explained by X_2 but has been explained by the inclusion of X_1 into the model.

The three or more variables model analogue of r is R (coefficient of multiple correlations) and it is a measure of the degree of association between Y and all the explanatory variables jointly. In practice, however, R is of little importance.

5.5 The Mean and Variance of the Parameter Estimates

The means of the estimates of the parameters of the three variable models are derived in the same way as those in the two variable models. By analogy to the procedures used to proof that the OLS estimates in the two variable models, the OLS estimates, $\hat{\beta}_0$, $\hat{\beta}_1$ *and* $\hat{\beta}_2$, of the three variables models, are unbiased estimates of the true parameters, i.e.,

$$E\left(\hat{\beta}_0\right) = \beta_0 \ ,$$

$$E\left(\hat{\beta}_1\right) = \beta_1 \ \ ,and$$

$$E\left(\hat{\beta}_2\right) = \beta_2$$

Similarly, the variances of the OLS estimates in three variable models are obtained using the same procedures to those used in the case of two variable models, and their formulae obtained from these procedures are given as follows.

1. $$Var\left(\hat{\beta}_0\right) = \hat{\sigma}_u^2 \left[\frac{1}{n} + \frac{\overline{X_1}^2 \sum x_{2i}^2 + \overline{X_2}^2 \sum x_{1i}^2 - 2\overline{X}_1 \overline{X}_2 \sum x_{1i} x_{2i}}{\sum x_{1i}^2 \sum x_{2i}^2 - \left(\sum x_{1i} x_{2i}\right)^2}\right] \quad \text{……..5.49}$$

Or

$$Var\left(\hat{\beta}_0\right) = \frac{\sigma_u^2}{n} + \overline{X}_1^2 \; Var\left(\hat{\beta}_1\right) + \overline{X}_2^2 Var\left(\hat{\beta}_2\right) + 2\,\overline{X}_1\,\overline{X}_2\, Cov\left(\hat{\beta}_1\,,\,\hat{\beta}_2\right)$$

Where

$$Cov\left(\hat{\beta}_1\,,\,\hat{\beta}_2\right) = \frac{-r_{12}\sigma_u^2}{\left(1 - r_{12}^2\right)\sqrt{\sum x_{1i}^2 \sum x_{2i}^2}}\,,$$

$$\hat{\sigma}_u^2 = \sum \hat{u}_i^2 \Big/ n - k$$

and

$$r_{12} = \frac{\sum x_{1i}^2 \sum x_{2i}^2}{\sqrt{\sum x_{1i}^2 \sum x_{2i}^2}}$$

2. $$Var\left(\hat{\beta}_1\right) = \frac{\sigma_u^2\left[\sum x_{2i}^2\right]}{\sum x_{1i}^2 \sum x_{2i}^2 - \left(\sum x_{1i}\, x_{2i}\right)^2}$$ ……..5.50

Or

$$Var\left(\hat{\beta}_1\right) = \frac{\sigma_u^2}{\sum x_{1i}^2\left(1 - r_{12}^2\right)}$$

3. $$Var\left(\hat{\beta}_2\right) = \frac{\sigma_u^2\left[\sum x_{1i}^2\right]}{\sum x_{1i}^2 \sum x_{2i}^2 - \left(\sum x_{1i}\, x_{2i}\right)^2}$$ ……5.51

Or

$$Var\left(\hat{\beta}_2\right) = \frac{\sigma_u^2}{\sum x_{2i}^2\left(1 - r_{12}^2\right)}$$

Note that equations 5.49 to 5.51 give the formulae of the variances of the OLS estimates for a model with three variables: a model with one dependent variable and two explanatory variables. The question, then, is how to find the variances of the estimates of the models that have more than three variables, say $K-$variables models.

The generalization of the formulae of the variances of the estimates is facilitated by using the concept of determinants as follows.

1. A model with two variables

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i$$

Then, $Var\left(\hat{\beta}_1\right) = \sigma_u^2 \Big/ \sum x_{1i}^2$

2. A model with tree variables

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

Then,

a) $Var\left(\hat{\beta}_1\right) = \dfrac{\hat{\sigma}_u^2 \sum x_{2i}^2}{\sum x_{1i}^2 \sum x_{2i}^2 - \left(\sum x_{1i} x_{2i}\right)^2}$

b) $Var\left(\hat{\beta}_2\right) = \dfrac{\hat{\sigma}_u^2 \sum x_{1i}^2}{\sum x_{1i}^2 \sum x_{2i}^2 - \left(\sum x_{1i} x_{2i}\right)^2}$

These expressions of the variances may be written in the form of determinants normal equations as follows:

The normal equations of the model with three variables in deviation form are (see equations 5.20 and 5.21 above):

$$\sum x_{1i} y_i = \hat{\beta}_1 \sum x_{1i}^2 + \hat{\beta}_2 \sum x_{1i} x_{2i} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots 5.20^*$$

$$\sum x_{2i} y_i = \hat{\beta}_1 \sum x_{1i} x_2 + \hat{\beta}_2 \sum x_{2i}^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots 5.21^*$$

Let A be the matrix of the known terms appearing in the right hand side of equations 5.20* and 5.21*.

Then,

$$A = \begin{bmatrix} \sum x_{1i}^{\ 2} & \sum x_{1i}\ x_{2i} \\ \sum x_{1i}\ x_{2i} & \sum x_{2\ i}^{2} \end{bmatrix}$$

Following this, the variance of the estimates, $\left[Var\ \left(\hat{\beta}_i\right)\right]$ can be obtained from the following rule.

The variance of each parameter estimate, say $\hat{\beta}_i$, is the product of σ_u^2 and the ratio of the minor determinant associated with this estimate divided by the (complete) determinant, where the complete determinant is obtained from matrix A defined above.

Using this rule, $Var\ \left(\hat{\beta}_1\right)$ is defined as:

$$Var\ \left(\hat{\beta}_1\right) = \sigma_u^2\ \frac{M_{11}\begin{vmatrix} \sum x_{1i}^{\ 2} & \sum x_{1i}\ x_{2i} \\ \sum x_{1i}\ x_{2i} & \sum x_{2\ i}^{2} \end{vmatrix}}{\begin{vmatrix} \sum x_{1i}^{\ 2} & \sum x_{1i}\ x_{2i} \\ \sum x_{1i}\ x_{2i} & \sum x_{2\ i}^{2} \end{vmatrix}}$$

$$\Rightarrow Var\ \left(\hat{\beta}_1\right)\ =\ \sigma_u^2\ \frac{\sum x_{2\ i}^{2}}{\sum x_{1\ i}^{2}\ \sum x_{2i}^{2} - \left(\sum x_{1i}\ x_{2i}\right)^2}$$

Similarly,

$$Var\left(\hat{\beta}_2\right) = \sigma_u^2 \frac{M_{11}\begin{vmatrix} \sum x_{1i}^{\ 2} & \sum x_{1i}\ x_{2i} \\ \sum x_{1i}\ x_{2i} & \sum x_{2\ i}^2 \end{vmatrix}}{\begin{vmatrix} \sum x_{1i}^{\ 2} & \sum x_{1i}\ x_{2i} \\ \sum x_{1i}\ x_{2i} & \sum x_{2\ i}^2 \end{vmatrix}}$$

$$\Rightarrow Var\left(\hat{\beta}_2\right) = \sigma_u^2 \frac{\sum x_{1\ i}^2}{\sum x_{1\ i}^2 \sum x_{2i}^2 - \left(\sum x_{1i}\ x_{2i}\right)^2}$$

Therefore, inspections of the above procedures tell us that the variance of the estimates of a model including can be computed from the ratio of two determinants:

i) The determinant appearing in the numerator is the minor formed after striking out the row and column of the terms corresponding to the coefficient whose variance is being computed.

ii) The determinant appearing in the denominator is the complete determinant of the known terms appearing on the right – hand side of the normal equations.

5.6 Interpretation of Regression Coefficients

The interpretation of the regression coefficients depends on the type of the model used in the regression analysis.

1. Linear model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 5.52$$

Where,

Y = consumption of coffee in pound per day.

X_{1i} = the price of coffee per pound

X_{2i} = price of tea per pound

Therefore, $\hat{\beta}_1$ in equation 5.52 implies that as the price of coffee increases by one unit coffee consumption, on average, decreases by $\hat{\beta}_2$ pounds per day; here we are expecting the sign of $\hat{\beta}_2$ to be negative since tea and coffee are presumed to be substitute commodities.

2. Log-linear or double-log model

$$\ln Y_i = \beta_0 + \beta_1 \ln X_{1i} + \beta_2 \ln X_{2i} + u_i \text{ .....5.53}$$

Where,

Y = consumption of coffee in pound per day.

X_{1i} = the price of coffee per pound

X_{2i} = price of tea per pound

In this case the estimates of the coefficients give a direct estimate of elasticity. For example, if $\hat{\beta}_1$ = -0.2530, it implies that if the price of coffee per pound goes up by 1 percent, on average, coffee consumption goes down by about 0.25 percent, per day.

Note that a change in $\ln Y$ gives a relative or proportional change in Y, where as a change in Y gives an absolute change.

In simple regression models like,

$Y_i = \beta_0 + \beta_1 X_i + u_i$

we are concerned with measuring the effect of an explanatory variable, X , on the dependent variable, Y , This effect is measured by $\hat{\beta}_1$,

$$where \quad \hat{\beta}_1 = \frac{Cov\ (y\ ,\ x)}{Var\ (x)} = \frac{\sum x_i\ y_i}{\sum x_i^2}$$

On the other hand, in the multiple regression equations with two explanatory variables X_1 and X_2 we can talk of the joint effect of X_1 and X_2, and the partial effect of X_1 or X_2 on Y.

To illustrate this assume the following model

$$Y_i = \beta_0 + \beta_1\ X_{1i} + \beta_2\ X_{2i} + u_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 5.54$$

In this case the partial effect of X_1 is measured by $\hat{\beta}_1$ and the partial effect of X_2 is measured by $\hat{\beta}_2$. Here, by partial effect we mean the effect of one variable holding the other variables constant or after eliminating the effect of the other variable. Thus, $\hat{\beta}_1$ can be interpreted as a measure of the effect of X_1 on Y, after eliminating the effect of X_2 on X_1. This interpretation suggests that we can drive the estimator $\hat{\beta}_1$ by estimating two separate simple regressions. To do this we usually follow the following steps.

Step 1: Estimate a regression of X_1 on X_2 as follows and obtain $\hat{b}_i$ *and* e_i,

where e_i is the estimate of w_i in equation 5.55.

The model

$$X_{1i} = b_0 + b_1\ X_{2i} + w_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 5.55$$

Then, by using the analogy of equation 5.15, the estimate of w_i in equation 5.55, can be given as:

$$e_i = x_{1i} - \hat{x}_{1i} \quad \text{.....5.56}$$

Where, as usual $\hat{x}_{1i} = \hat{X}_{1i} - \overline{X}_1$

By following the procedures we have followed to drive equation 4.12, we get

$$\hat{x}_{1i} = \hat{b}_1 \; x_{2i} \quad \text{.....5.57}$$

Substituting equation 5.57 into 5.56, yields

$$e_i = x_{1i} - \hat{b}_1 x_{2i} \quad \text{.....5.58}$$

Where,

$$\hat{b}_1 = \frac{\sum x_{1i} \; x_{2i}}{\sum x_{1i}^{\;2}}$$

In equation 5.55 we know that e_i is the part of X_1 that is left after removing the effect of X_2 on X_1. In other words, it represents the part of the variation in X_1 left unexplained by X_2.

Step 2:- Regress Y_i *on* e_i (obtained from step 1) as given in equation 5.59 below, and then observe the relationship between α_1 and $\hat{\beta}_1$ (the estimate obtained from multiple regression equation 5.54.

The regression Y_i *on* e_i can be modeled as:

$$Y_i = \alpha_0 + \alpha_1 e_i + v_i \quad \text{.....5.59}$$

Where, v_i is the stochastic term that satisfies the usual assumptions of the error term.

Then, the OLS estimate of α_1 is given as

$$\hat{\alpha}_1 = \frac{Cov\,(y_i\,,\,e_i)}{Var\,(e_i)} \quad \text{.....5.60}$$

To make the procedure of expanding equation 5.60, let us solve the terms in the numerator and denominator in turn.

1. Using the formula of variance,

$$Var\left(e_i\right) = \frac{\sum e_i^2}{n} \quad \text{.....5.61}$$

Substituting e_i from equation 5.58 into equation 5.61, gives

$$\sum e_i^2 = \sum\left(x_{1i} - b_1\hat{x}_{2i}\right)^2$$

$$= \sum\left(x_{1\,i}{}^2 + \hat{b}_1^2\ x_{2\,i}^2 - 2\ \hat{b}_1\ x_{1\,i}\,x_{2\,i}\right)$$

$$= \sum x_{1\,i}{}^2 + b_1^2\ \sum x_{2\,i}{}^2 - 2\ \hat{b}_1 \sum x_{1\,i}\,x_{2\,i} \quad \text{.....5.62}$$

Plugging in $\frac{\sum x_{1i}\ x_{2i}}{\sum x_{2\ i}^2}$ for $\hat{b}_1$ in equation 5.62

$$\sum e_i{}^2 = \sum x_{1i}{}^2 + \left(\frac{\sum x_{1i}\ x_{2\,i}}{\sum x_{2\ i}^2}\right)^2 \sum x_{2\ i}^2 - 2\left(\frac{\sum x_{1i}\ x_{2i}}{\sum x_{2\ i}^2}\right)\sum x_{1i}\ x_2$$

$$= \sum x_{1\,i}{}^2 + \frac{\left(\sum x_{1i}\ x_{2\,i}\right)^2}{\sum x_{2\ i}^2} - 2\ \frac{\left(\sum x_{1\,i}\ x_{2\,i}\right)^2}{\sum x_{2\ i}^2}$$

$$= \sum x_{1\ i}{}^2 - \frac{\left(\sum x_{1i}\ x_{2\,i}\right)^2}{\sum x_{2\,i}{}^2}$$

$$\Rightarrow \sum e_i{}^2 = \frac{\sum x_{1i}{}^2 \sum x_{2\ i}^2 - \left(\sum x_{1\,i}\ x_{2\,i}\right)^2}{\sum x_{2\ i}{}^2} \quad \text{.....5.63}$$

Thus, by substituting equation 5.63 into 5.61

$$Var\ (e_i) = \frac{\sum x_{1i}^{\ 2}\ \sum x_{2i}^{\ 2} - \left(\sum x_{1i}\ x_{2i}\right)^2 \Big/ \sum x_{2i}^{\ 2}}{n} \quad \text{……5.64}$$

2. Using the formula of covariance between variables

$$Cov\ (y_i\ ,\ e_i) = \frac{\sum y_i\ e_i}{n} \quad \text{……5.65}$$

By substituting e_i from equation 5.58 into equation 5.65, we get

$$Cov\ (y_i\ ,\ e_i) = \frac{\sum y_i\left(x_{1i} - \hat{b}_1\ x_{2i}\right)}{n}$$

$$= \frac{\sum\left(y_i\ x_{1i} - \hat{b}_1\ y_i\ x_{2i}\right)}{n}$$

$$= \frac{\sum y_i\ x_{1i} - \hat{b}_1 \sum y_i x_{2i}}{n} \quad \text{……..5.66}$$

Since $\hat{b}_1 = \frac{\sum x_{1i}\ x_{2i}}{\sum x_{2\ i}^2}$, equation 5.66 becomes

$$Cov\ (y_i\ ,\ e_i) = \frac{\sum y_i\ x_{1i} - \left(\frac{\sum x_{1i}\ x_{2i}}{\sum x_{2\ i}^2}\right) \sum y_i\ x_{2i}}{n}$$

$$= \frac{\sum y_i\ x_{1i}\ \sum x_{2i}^{\ 2} - \frac{\left(\sum x_{1i}\ x_{2i}\right)\left(\sum y_i\ x_{2i}\right)}{\sum x_{2\ i}^2}}{n} \quad \text{……5.67}$$

Now, substituting equations 5.64 and 5.67 into equation 5.60, we get

$$\hat{\alpha}_1 = \frac{Cov\ (y_i\ ,\ e_i)}{Var\ (e_i)} = \frac{\left(\frac{\sum y_i\ x_{1i}\ \sum x_{2i}^{\ 2} - \frac{\left(\sum x_{1i}\ x_{2i}\right)\left(\sum y_i\ x_{2i}\right)}{\sum x_{2\ i}^{2}}}{n} \right)}{\left(\frac{\sum x_{1i}^{\ 2}\ \sum x_{2i}^{\ 2} - \left(\sum x_{1i}\ x_{2i}\right)^2 \Big/ \sum x_{2i}^{\ 2}}{n} \right)}$$

$$\Rightarrow \hat{\alpha}_1 = \frac{\sum y_i\ x_{1i}\ \sum x_{2\ i}^{2} - \sum y_i\ x_{2i}\ \sum x_{1i}\ x_{2i}}{\sum x_{1\ i}^{\ 2}\ \sum x_{2\ i}^{\ 2} - \left(\sum x_{1i} x_{2i}\right)^2} = \hat{\beta}_1 \text{.....5.68}$$

The implication of equation 5.68 is that the pure effect of X_1 on Y after eliminating the effect of X_2 both on Y and X_1 can be obtained by simply estimating the multiple regression equation by OLS and hence, $\hat{\beta}_1$ gives the pure effect of X_1 on Y.

5.7 Statistical Inference in Multiple Regression Model

In this part we simply extend what we discussed in chapter four on interval estimation and hypothesis testing. Although in many ways the concepts on statistical significance developed there can be applied straight forward to the estimates obtained from multiple regression model, a few additional features are unique to such models.

As we discussed in chapter four, following the usual assumption $U_i \sim N\left(0\ ,\ \sigma_u^2\right)$, it can be shown that the estimators $\hat{\beta}_0$, $\hat{\beta}_1$ *and* $\hat{\beta}_2$ are

normally distributed with means equal to β_0 , β_1 and β_2 and variances given in equations 5.49, 5.50 and 5.51, respectively.

Then, if we standardize the distribution of the OLS estimates, $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$, it follows $t-$distribution with $n-k$ degrees of freedom. This is due to the fact that σ_u^2 is unknown; hence, we cannot use $Z-$distribution for statistical inference in multiple regression models. In other words, the $t-$distribution will be used to establish confidence intervals as well as to test hypothesis about the true population partial regression coefficients.

The $t-$values used in statistical inference are defined as follows

$$t_{calculated} = \frac{\hat{\beta}_i - \beta_i}{Se\left(\hat{\beta}_i\right)} \quad \text{.....5.69}$$

5.7.1 Statistical significance of individual coefficients in multiple regression

To test the statistical significance of individual coefficients of multiple regression coefficients, we use the same procedures we used for the estimates of simple regression models.

To conduct the $t-$test, we follow the following steps.

Step 1. Define the null and alternative hypothesis

Step 2. Choose the desired level of significance (α)

Step 3. Define the number of degrees of freedom

Step 4. Obtain the calculated value of t from equation 5.69

Step 5. Obtain the critical values (define the critical region)

Step 6. Make decision.

In multiple regression models, the customary form of hypothesis for statistical significance of individual coefficients is

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

Then, equation 5.69 becomes

$$t_{calculated} = \frac{\hat{\beta}_i}{Se\left(\hat{\beta}_i\right)} \text{.....5.70}$$

Decision Rule

Reject H_0, if the $t-$value obtained lies in the rejection region or less than its critical value obtained from t tables for the given significance level and degree of freedom ($n-K$).

5.7.2 Testing the overall significance of the sample regression function

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

In this section, however, we are concerned with a joint hypothesis that the slope coefficients, say $\beta 1$ and $\beta 2$ in three variables model (see equation 5.54), are jointly or simultaneously equal to zero.

The hypothesis in this case is defined as:

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

H_1 : Not all slope coefficients are simultaneously zero.

A test of such a hypothesis is called a test of the **overall significance** of the observed or estimated regression line, that is, whether *Y* is linearly related to both *X*1 and *X*2

As you might have noted, in testing the individual significance of an observed partial regression coefficient, we assumed implicitly that each test of significance was based on a different (i.e., independent) sample. Thus, in testing the significance of $\hat{\beta}_1$ under the hypothesis that $\beta 1 = 0$, it was assumed tacitly that the testing was based on a different sample from the one used in testing the significance of $\hat{\beta}_2$ under the null hypothesis that $\beta 3 = 0$. But if we use the same sample data to test the joint hypothesis defined above, we shall be violating the assumption underlying the test procedure.

Therefore, we cannot use the usual *t* test to test the joint hypothesis that the true partial slope coefficients are zero simultaneously. However, this joint hypothesis can be tested by the **analysis of variance** (ANOVA), which can be given as follows.

Table 5.1. ANOVA Table for Three variables Regression Model

Source of Variation	SS	Df	MSS
Explained Variation (ESS)	$\hat{\beta}_1 \sum y_i \ x_{1i} + \hat{\beta}_2 \sum y_i \ x_{2i}$	2	$\frac{\hat{\beta}_1 \sum y_i x_{1i} + \hat{\beta}_2 y_i x_{2i}}{2}$
Unexplained Variation (RSS)	$\sum \hat{u}_i^2$	$n-3$	$\hat{\sigma}_u^2 = \sum \hat{u}_i^2 / n-3$
Total Variation (TSS)	$\sum y^2{}_i = \hat{\beta}_1 \sum y_i \ x_{1i} + \hat{\beta}_2 \sum y_i \ x_{2i} + \sum \hat{u}_i^2$	$n-1$	

Using the usual assumption of normal distribution for U_i and the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$, the variable F defined (from the above table) as follows has F distribution with 2 and $n-3$ degrees of freedom.

$$F = \frac{ESS/df}{RSS/df}$$

$$= \frac{ESS/3-1}{RSS/n-3}$$

$$= \frac{\hat{\beta}_1 \sum y_i x_{1i} + \hat{\beta}_2 y_i x_{2i} / 2}{\sum \hat{u}_i^2 / n-3}$$

$$\Rightarrow F = \left(\frac{n-3}{2}\right)\left(\frac{R^2}{1-R^2}\right) \text{.....5.71}$$

Note that equation 5.71 yields the value of F for three variables regression models. By analogy, the value of F for $K-$variables regression models is defined as

$$F = \left(\frac{n-K}{K-1}\right)\left(\frac{R^2}{1-R^2}\right) \text{.....5.72}$$

Decision Rule:

If the *F* value computed from equation 5.71 exceeds the critical *F* value from the *F* table at the α percent level of significance, we reject H_0; otherwise we do not reject it. Alternatively, if the *p* value of the observed *F* is sufficiently low, we can reject H_0.

5.8 Matrix Approach to Multiple Linear Regression Models

In fitting a multiple regression model, it is much more convenient to express the mathematical operations using matrix notation. To illustrate, suppose that there are $K-$regrssor variables, n sample observations and that the model relating the regressors to the dependent variable is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + U_i \text{ ..... 5.73}$$

Where, $i = 1,2,...,n$

Note that equation 5.73 is a shorthand expression for the following set of n simultaneous equation

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \beta_3 X_{31} + ...+ \beta_{1c} X_1 + u_1$$

$$Y_2 = \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \beta_3 X_3 + ... + \beta_k X_{k2} + u_2$$

; ; ;

; ;

$$Y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \beta_3 X_{3n} + ... + \beta_k X_{kn} + u_n$$

This system of n equations can be expressed in matrix notation as:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix}_{(n\ x1)} = \begin{bmatrix} 1 & X_{12} & X_{21} & . & . & . & X_{k1} \\ 1 & X_{12} & X_{22} & . & . & . & X_{k2} \\ 1 & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 1 & X_{1n} & X_{2n} & . & . & . & X_{kn} \end{bmatrix}_{[n\ x(k+1)]} \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_{1c} \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ . \\ . \\ . \\ u_n \end{bmatrix}$$

$$\Rightarrow Y \quad = \quad X \quad \beta \quad U$$

$$\Rightarrow Y = X\beta + U, \text{ where}$$

$Y = \text{n x 1}$ column vector of observations on the dependent variable Y

$X = \text{n x (k+1)}$ matrix giving n observations on K variables, where the first column of 1's represents the intercept term.

$\beta = (\text{k+1})\text{ x 1}$ column vector of the unknown parameters, $\beta_0, \beta_1, \beta_2 ....., \beta_k$

$U = \text{n x 1}$ column vector of n disturbance u_i

As usual to make the model specification stage complete and to obtain the OLS estimates of the coefficients of the model, we need to make reasonable assumptions about the distribution of the stochastic term u_i . The counterpart of the usual assumptions of classical linear regression model for matrix approach are:

Assumption 1. $E(U) = 0$, where U and 0 in this case are n x 1 column vectors; 0 being a null vector

Assumption 2. $E(U\,U') = \sigma_u^2 I$, where I is an n x n identity matrix

Assumption 3, X , which the $\mathrm{n\ x\ (k+1)}$ matrix, is non stochastic, i.e., it consists of a set of fixed values.

Assumption 4. The rank of X is $\mathrm{P\,(X) = k+1}$, where $\mathrm{k+1}$ is the number of columns in X and $\mathrm{k+1}$ is less than the number of observations n. This means that the columns of the X matrix are linearly dependent; i.e., there is no exact linear relation ship among the X variables.

Assumption 5. The U vector has a multivariate normal distribution.

$$\Rightarrow U \sim N\left(0\ ,\ \sigma_u^2 I\right)$$

Dear students, what relationships do you observe between these assumptions and the assumptions we made in chapter 2 for simple regression models?

For example, assumption 2 above is a compact form expressing the two assumptions of classical linear regression model in scalar notation, namely the assumption of homoscedasticity of $U_i's$ (constant σ_u^2) and the assumption of non-auto correlated $U's$ (*i.e* $E\left(u_i\ u_j\right) = 0$) . To understand how it goes, we use the multiplication operation for matrix notations.

Using the multiplication rule of matrix notations,

$$U\,U' = \begin{bmatrix} U_1 \\ U_2 \\ . \\ . \\ . \\ U_n \end{bmatrix} \begin{bmatrix} U_1 , U_2 \ldots . U_n \end{bmatrix} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 5.74$$

Where, U' is the transpose of U .

The multiplication of the vectors (the column vector and row vector) in equation 5.74, yields

$$\Rightarrow\ U\,U' = \begin{bmatrix} (u_1^2) & (u_1 u_2) & . & . & . & . & . & (u_1 u_n) \\ (u_2 u_1) & (u_2^2) & . & . & . & . & . & (u_2 u_n) \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ (u_n u_1) & (u_n u_2) & . & . & . & . & . & (u_n^2) \end{bmatrix} \ldots\ldots\ldots\ldots\ldots\ldots 5.75$$

Taking the expected value of equations 5.75 gives,

$$E(U\,U') = E \begin{bmatrix} (u_1^2) & (u_1 u_2) & . & . & . & . & . & (u_1 u_n) \\ (u_2 u_1) & (u_2^2) & . & . & . & . & . & (u_2 u_n) \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ (u_n u_1) & (u_n u_2) & . & . & . & . & . & (u_n^2) \end{bmatrix}$$

$$= \begin{bmatrix} E(u_1^2) & E(u_1 u_2) & . & . & . & . & . & E(u_1 u_n) \\ E(u_2 u_1) & E(u_2^2) & . & . & . & . & . & E(u_2 u_n) \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ E(u_n u_1) & E(u_n u_2) & . & . & . & . & . & (u_n^2) \end{bmatrix} \ldots\ldots\ldots\ldots 5.76$$

Since $U's$ are homoscedastic and their values are non-autocorrelated (i.e. $E\left(U_i^2\right) = \sigma_u^2 \quad and \quad E\left(u_i\, u_j\right) = 0$), equation 5.76 becomes

$$E\left(U\,U'\right) = \begin{bmatrix} \sigma_u^2 & 0 & . & . & 0 \\ 0 & \sigma_u^2 & .. & . & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & 0 & . & . & \sigma_u^2 \end{bmatrix}$$

$$= \sigma_u^2 \begin{bmatrix} 1 & 0 & . & . & . & 0 \\ 0 & 1 & . & . & . & 0 \\ . & . & . & . & . & . \\ 0 & 0 & . & . & . & 1 \end{bmatrix}$$

$$\Rightarrow E\left(U\,U'\right) = \sigma_u^2\ I \quad \text{……..5.77}$$

Where, I is an $\mathrm{n\ x\ n}$ identity matrix and is called the variance-covariance matrix of the disturbance term U. In this matrix, the elements in the main diagonal that runs from the upper left corner to the lower right corner give the variances and the elements off the main diagonal give the covariance.

So far we have specified our model through matrix approach and made plausible assumptions. Having done these tasks, now we are at the estimation stage.

To illustrate this stage in matrix approach, the population regression function (PRF) is given as

$$PRF: Y = X\beta + U \quad \text{…… ……5.78}$$

The main interest here is to obtain OLS estimates of β from sample observations on Y and X; i.e., based on SRF which can be given in scalar representation as:

$$SRF: Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_k X_{ki} + \hat{U}_i \quad \ldots\ldots\ldots 5.79$$

The matrix representation of equation 5.79 is

$$\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix}_{(n\,x1)} = \begin{bmatrix} 1 & X_{12} & X_{21} & . & . & . & X_{k1} \\ 1 & X_{12} & X_{22} & . & . & . & X_{k2} \\ 1 & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 1 & X_{1n} & X_{2n} & . & . & . & X_{kn} \end{bmatrix}_{[n\,x(k+1)]} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_2 \\ . \\ . \\ . \\ \hat{\beta}_k \end{bmatrix}_{(k+1,1)} + \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ . \\ . \\ . \\ \hat{u}_n \end{bmatrix}_{(n\,x1)}$$

This can be written in a compact form as

$$Y = X\hat{\beta} + \hat{U} \quad \ldots\ldots\ldots 5.80$$

In chapter 2, we discussed that according to the principle of OLS, the estimates are obtained by minimizing the sum of the squares of the residuals (RSS).

$$\Rightarrow \min \sum \hat{U}_i^2 \quad \ldots\ldots\ldots 5.81$$

In matrix notation, equation 5.81 is given as

$$\min \hat{U}.\hat{U} \quad \ldots\ldots\ldots 5.82$$

To obtain the OLS estimates from equation 5.82, recall from equation 5.80 that

$$Y = X\hat{\beta} + \hat{U}$$

$$\Rightarrow \hat{U} = Y - X\hat{\beta} \quad \ldots\ldots\ldots 5.83$$

$$\Rightarrow \hat{U}' = \left(Y - X\hat{\beta}\right)' \quad \ldots\ldots\ldots 5.84$$

From equations 5.83 and 5.84, we obtain

$$\hat{U}'\hat{U} = \left(Y - X\hat{\beta}\right)'\left(Y - X\hat{\beta}\right)$$

Using the properties of the transpose of matrix notation, equation 5.85 becomes

$$\hat{U}'\hat{U} = \left(Y' - \hat{\beta}'X'\right)\left(Y - X\hat{\beta}\right), \text{ since } \left(X\hat{\beta}\right)' = \hat{\beta}'X'$$

$$= Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots 5.85$$

Note in equation 5.85 that $Y'X\hat{\beta}$ is a scalar. This is because

i) Since Y is an (n x 1) column vector, its transpose Y' will be (1 x n) row vector

ii) Since X is an [n x (k + 1)] matrix, the product of Y' and X (i.e. $Y'X$) will be a row vector of order [1 x (k + 1)].

iii) Since $\hat{\beta}$ is a [(k + 1) x 1] column vector, the product $Y'X\hat{\beta}$ will be a scalar.

Thus, $\hat{\beta}'X'Y$, which is the transpose of $Y'X\hat{\beta}$, will also be a scalar.

$\Rightarrow \hat{\beta}'X'Y = Y'X\hat{\beta}$ (Since the transpose of a scalar is the scalar itself)

Consequently, equation 5.85 can be written as

$$\hat{U}'\hat{U} = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta} \quad \ldots\ldots\ldots\ldots\ldots\ldots 5.86$$

$$\hat{U}'\hat{U} = Y'Y - 2\left(Y'X'\hat{\beta}\right)' + \hat{\beta}'X'X\hat{\beta}$$

$$\Rightarrow \hat{U}'\hat{U} = Y'Y - 2\hat{\beta}'\left(Y'X\right)' + \hat{\beta}'X'X\hat{\beta} \quad \ldots\ldots\ldots\ldots\ldots\ldots 5.87$$

Now, differentiating equation 5.87 partially with respect to $\hat{\beta}$ and setting the result to zero, we obtain

$$\frac{\partial \hat{U}'\hat{U}}{\partial \hat{\beta}} = -2\,X'Y + 2\,X'X\,\hat{\beta} = 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots 5.88$$

Note that to get equation 5.88, we used the rule of differentiation of matrix notations, namely

1) $$\partial\left(\frac{\hat{\beta}'A'}{\partial \hat{\beta}}\right) = A'$$

2) $$\frac{\partial \hat{\beta}'A\hat{\beta}}{\partial \hat{\beta}} = 2A\hat{\beta}$$

Where, A is ([k + 1]x[k + 1]) matrix and $\hat{\beta}$ is a ([k + 1]x1) column vector.

Equation 5.88, can be rewritten as

$$-2\,X'Y = -2\,X'X\,\hat{\beta} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 5.89$$

Now, if the inverse of $X'X$ (i.e. $(X'X)^{-1}$) exists, multiplying both sides of equation 5.89 by $(X'X)^{-1}$, we get

$$\Rightarrow (X'X)^{-1}X'Y = (X'X)^{-1}(X'X)\hat{\beta} \ldots\ldots\ldots\ldots\ldots\ldots\ldots 5.90$$

But since $(X'X)^{-1}(X'X) = I$, we get

$$(X'X)^{-1}\,X'Y = I\,\hat{\beta} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 5.91$$

$$\Rightarrow \hat{\beta} = (X'X)^{-1} X'Y \quad \text{.....5.92}$$

Since $I_{(k+1\ X\ K+1)}\ \hat{\beta}_{(k+1\ X\ 1)} = \hat{\beta}$

Note that equation 5.92 gives the OLS estimates of the coefficients in matrix notions. In this equation

$$X'X = \begin{bmatrix} n & \sum X_{1i} & \cdots & \sum X_{ki} \\ \sum X_{2i}X_{1i} & \sum X_{2i}^2 & \cdots & \sum X_{2i}X_{ki} \\ & & \cdots & \\ \sum X_{ki}X_{1i} & \sum X_{ki}X_{21i} & & \sum X_{k_i^2} \end{bmatrix} \text{ and}$$

$$X'Y = \begin{bmatrix} \sum Y_i \\ \sum X_{1i}Y_i \\ \cdots \\ .. \\ \sum X_{ki}Y_i \end{bmatrix}$$

Furthermore, the matrix notations can be used to obtain the variance of the OLS estimates and the covariance between any two such estimates.

By definition the Variance-covariance matrix of $\hat{\beta}$ is given as:

$$Var\,Cov(\hat{\beta}) = E\left\{ \left[\hat{\beta} - E(\hat{\beta}) \right] \left[\hat{\beta} - E(\hat{\beta}) \right]' \right\} \text{..... 5.93}$$

But we know from previous discussions that

$$\hat{\beta} = (X'X)^{-1} X'Y \quad \text{.....5.92*}$$

and

$$Y = X\beta + U \quad \text{.....5.78*}$$

Now substituting $X\beta + U$ from equation 5.78* for Y in equation 5.92*, we obtain

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

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

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

$$\Rightarrow \hat{\beta} - \beta = (X'X)^{-1} X'U \quad \text{..... 5.94}$$

We have already shown in chapter three that the OLS estimates of β are unbiased and hence $E(\hat{\beta}) = \beta$. Therefore, equation 5.93 becomes

$$Var\,Cov(\hat{\beta}) = E\left\{(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right\} \quad \text{..... 5.95}$$

By substituting 5.94 into 5.95, we get

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

Note that $\left((X'X)^{-1} X'U\right)' = U'X(X'X)^{-1}$. Hence, this equation becomes

$$VarCov(\hat{\beta}) = E\left\{(X'X)^{-1} X'UU'X(X'X)^{-1}\right\} \quad \text{.....5.96}$$

Since by assumption 3 of matrix notation the X‘s are non-stochastic, the expected value of the terms in the right hand side of equation 5.96, will be

$$Var\,Cov\ \left(\hat{\beta}\right) = \left(X'X\right)^{-1} X'E\left(UU'\right) X\left(X'X\right)^{-1}$$

Since $E\left(UU'\right) = \sigma^2{}_u\, I$

$$Var\,Cov\ \left(\hat{\beta}\right) = \left(X'X\right)^{-1} X'\sigma_u^2\ I\,X\ \left(X'X\right)^{-1}$$

$$\Rightarrow Var\,Cov\ \left(\hat{\beta}\right) = \underline{\underline{\sigma_u^2\ \left(X'X\right)^{-1}}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ 5.97$$

Note that equation 5.97 gives the variance-covariance matrix of $\hat{\beta}$; where the main diagonal that runs form the upper left corner to the button right corner gives the variance and the off diagonal elements gives covariance between the estimates.

As we can see from equation 5.97, the variance-covariance matrix σ_u^2, which is unknown and needs to be estimated from sample observations. In the previous chapters (chapter 2 and 4), we have shown that the unbiased estimator of σ_u^2 for simple linear regression model and for two explanatory variables model, respectively, is given as:

$$\hat{\sigma}_u^2 = \frac{\sum \hat{U}_i^2}{n-2}$$

and

$$\hat{\sigma}_u^2 = \frac{\sum \hat{U}_i^2}{n-3}$$

Similarly, the general formula for k -explanatory variables model is given as:

$$\hat{\sigma}_u^2 = \frac{\sum \hat{U}_i^2}{n-k}$$

In matrix notation, the estimator of σ_u^2 is given as:

$$\hat{\sigma}_u^2 = \frac{\hat{U}'\hat{U}}{n-k} \quad \text{.....5.98}$$

Though the value of $\hat{U}'\hat{U}$ can be obtained form the estimated residuals, they can be estimated from observed values of Y and X. To do this, recall that

$$\sum \hat{U}_i^2 = RSS = TSS - ESS$$

But

$$TSS = \sum y_i^2$$

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

$$= \sum Y_i^2 - 2\bar{Y}\sum Y_i + \bar{Y}^2$$

$$= \sum Y_i^2 - n\bar{Y}^2 \quad \text{.....5.99}$$

In matrix notation, $\sum Y_i^2 = Y'Y$. Hence, equation 5.99 becomes

$$TSS = Y'Y - n\bar{Y}^2 \quad \text{..... 5.100}$$

To obtain the formula of *ESS* in matrix notation let us consider the pattern of the formula of ESS in models with different number of explanatory variables.

i) In a model with one explanatory variable

$$ESS = \hat{\beta}_1 \sum y_i x_{1i}$$

ii) In a model with two explanatory variables

$$ESS = \hat{\beta}_1 \sum y_i x_{1i} + \hat{\beta}_2 \sum y_i x_{2i}$$

iii) In a model with k – explanatory variables

$$ESS = \hat{\beta}_1 \sum y_i x_{1i} + \hat{\beta}_2 \sum y_i x_{2i} + ... + \hat{\beta}_k \sum y_i x_{ki}$$

Following these patterns, *ESS* in matrix notation is given as:

$$ESS = \hat{\beta}'X'Y - n\overline{Y}^2 \quad \text{………………………… 5.101}$$

Therefore, from equations 5.100 and 5.101, we define *RSS* as:

$$RSS = \hat{U}'\hat{U}$$

$$= TSS - ESS$$

$$= Y'Y - \hat{\beta}'X'Y \quad \text{………………………… 5.102}$$

Finally, we can drive formula for the coefficient of determination, R^2 using matrix notation. To do this recall that

$$R^2 = \frac{ESS}{TSS} \quad \text{………………………………………..5.103}$$

Plugging in equations 5.100 and 5.101, into equation 5.103, we obtain

$$R^2 = \frac{\hat{\beta}'X'Y - n\overline{Y}^2}{Y'Y - n\overline{Y}^2} \quad \text{…………………………………… 5.104.}$$

Exercise

1. Suppose that the following results were obtained from observations on 100 employees of XYZ Company.

$\sum X_{1i} = 123$

$\sum X_{1i}Y_i = 1290$

$\sum Y_i = 460$

$\sum Y_i^2 = 539{,}500$

$\sum X_{2i} = 96$

$\sum Y_i^2 = 3924$

$\sum X_{1i}^2 = 232$

$\sum X_{2i}^2 = 167$

$\sum X_{2i}Y_i = 615$

$\sum X_{2i}X_{1i} = 125$

$\sum X_{1i}Y_i = 870$

Given that the PRF is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

Where, Y, X_1 and X_2 denote hourly wage, education level in years of schooling and years of work experience, and u represents the error term that satisfies the usual assumptions of classical linear regression models.

Then, based on the above information,

i. Compute the OLS estimates of the coefficient of the model
ii. Interpret the coefficients of the model
iii. Find the standard errors of the OLS estimates. Are all the estimates plausible from the point of view of economic theory?

2. Using the information given in equation #1 above, form its matrix representation and
    i. Compute the OLS estimates of the coefficient of the model
    ii. Calculate the coefficient of multiple determination
    iii. Calculate adjusted R^2
    iv. Test the hypothesis that $H_0 : \beta_1 = 0$

CHAPTER SIX

VIOLATIONS OF BASIC ASSUMPTIONS OF LINEAR REGRESSION MODELS

Overview

In previous chapters we dealt with linear regression models, which are based on some basic assumption for their applicability. In this chapter, however, we will try to relax some of these assumptions and see the effect of these relaxations on OLS estimates, and inferences and predictions made using the OLS estimates. Specifically, in this chapter we will try to relax three of the basic assumptions, namely the assumptions of non auto-correlated error terms, homoscedastic variance of the error terms and no perfect multicollinearity among the explanatory variables.

In the first part of this chapter we will analyze the causes, consequences and detection mechanisms for autocorrelation of the error term. In the subsequent sections of the chapter, we will treat issues related to heteroscedasticity of the variance of the error term and multicollinearity among the explanatory variables.

At the end of this chapter students will be able to:

- Understand the causes and consequences of auto-correlated error terms; and differentiate between different types of autocorrelation schemes
- Use different mechanisms to detect whether the error terms are auto-correlated; and take remedial measures to cure the problem of auto-correlation, if any.
- Understand the causes and consequences of the heteroscedasticity of the variance of the error term

- Detect whether the variance of the error term is heteroscedastic or homoscedastic; and if it is heteroscedastic, take remedial measures to cure the consequential problem.
- Understand the causes and consequences of perfect and near to perfect multicollinearity among the explanatory variables
- Identify whether there is no perfect or near to perfect multicollinearity among the explanatory variables

6.1 Autocorrelation

6.1.1 Introduction

One of the assumptions of classical regression models is that the successive values of the error term, U are temporally independent. In other words, this assumption means that the values of the error term in any one period are independent from its values in any previous periods. This implies that the covariance between any two values of the error term U, such as $u_i \; and \; u_j$, is equal to zero.

If the value of U in any particular period is correlated with its own preceding value (or values), then there exists autocorrelation or serial correlation between the values of the error term.

Generally, autocorrelation refers to relationships between the successive values of the same variable; not between the values of different variables. In this section we are particularly interested in the autocorrelation of the values of U.

6.1.2 Sources of autocorrelation

In practical exercise, autocorrelation between the values of U may arise from different sources. The possible sources of autocorrelation for the values of the error term are discussed as follows.

1. **Omission of explanatory variables**

It is customary that most economic variables tend to be auto-correlated. If an auto-correlated variable has been excluded from the set of explanatory variables, obviously its influence will be reflected in the random variable U. This case may be called "**quasi – autocorrelation**", since it is due to the auto-correlated pattern of omitted explanatory variables; and not due to the behavioral pattern of the values of the true U.

2. Misspecification of the mathematical form of the model

If we have adopted a mathematical form that differs from the true form of the relationship, the values of U may show serial correlations.

3. Interpolations in statistical observations

Most of the published time series data involve some interpolation and "smoothing" processes. In this process, the average of the values disturbance term over successive time periods will be taken. Consequently, the successive values of U are interrelated and exhibit autocorrelation pattern.

4. Misspecification of the true random term, U

It may well be expected in many cases that the successive values of the true U are corrected. Thus even the purely random factors (such as wars, drought, storms, strikes etc) impose influences that are spread over more than one period of time. These factors result in serially dependent values of the disturbance term U, so that if we assume $E(u_i\ u_j) = 0$, we will really misspecify the true pattern of the values of U. This case of autocorrelation may be called "**true autocorrelation**" because its root lies in the U term itself.

This discussion of sources of autocorrelation makes it obvious that the assumption of temporal independence of the values of U can be easily violated in practice. Hence, a logical question that follows is: **"What form will the autocorrelation among the values of the error term take?"**

Most standard econometric text books deal with the simple case of linear relationship between any two successive values of U, which is known as a **first-order autoregressive scheme**. In this scheme the value of U in any particular period depends on its own value in the preceding period alone. This scheme is also called **first-order Markov process.**

On the other hand, if the values of U in a certain period depend on its value in the two previous periods, then the form of autocorrelation is called **second-order autoregressive scheme**, and so on. Nonetheless, in most applied research it is assumed that if autocorrelation exists, it takes the form of the simple first-order autoregressive scheme. Consequently, in this module we will give emphasis to first-order autoregressive scheme.

6.1.3 The first-order autoregressive scheme

In this section we will limit our analysis of the autocorrelation problem to the simple first -order autoregressive scheme, since most classical text books refer to this model as the most frequently assumed model of autocorrelation in applied econometric research. Hence, we will try to see its distribution. Concisely, we will try to establish the mean, variance and covariance of U when its values are correlated with the simple Markov process.

The first-order Markov scheme implies that

$$u_t = f\left(u_{t-1}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 6.1.1$$

In this process, the relationship between $u_t \; and \; u_{t-1}$ is assumed to be linear and is given as:

$$u_t = \rho u_{t-1} + \in_t \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 6.1.2$$

Where

ρ is the autocorrelation coefficient between $u_t \; and \; u_{t-1}$ and

$\in_t$ is a random variable satisfying all the usual assumptions of the error term, such as

$$E\left(\in_t\right) = 0$$

$$E\left(\in_t \in_{t-1}\right) = 0 \quad \text{and}$$

$$E\left(\in_t^2\right) = \sigma_\in^2 \text{, which is constant.}$$

Dear students, why do you think we use ρ as a coefficient of u_{t-1} in equation 6.1.2? --
----

To answer this question, let us write equation 6.1.2 in general form as follows.

$$u_t = a u_{t-1} + \in_t \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ 6.1.3$$

Note that equation 6.1.2 represents a linear relationship between u_t *and* u_{t-1}, where a is the slope coefficient and the constant intercept is equal to zero. Furthermore, equation 6.1.2 is a simple linear regression model with suppressed constant term. Then, if we apply OLS to equation 6.1.3, we obtain

$$\hat{a} = \frac{\sum_{t=2}^{n} u_t u_{t-1}}{\sum_{t=2}^{n} u_{t-1}^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ 6.1.4$$

On the other hand, we know that the autocorrelation coefficient, ρ, between u_t *and* u_{t-1} is given as:

$$\rho_{u_t, u_{t-1}} = \frac{\sum u_t u_{t-1}}{\sqrt{\sum u_t^2 \sum u_{t-1}^2}} \ldots\ldots\ldots\ldots\ldots\ldots\ 6.1.5$$

Equation 6.1.5 is a special form of correlation coefficients between any two variables (X and Y), which are defined as

$$r_{x,y} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$

Note that for large sample size (i.e. for large t),

$$\sum u_t^2 \cong \sum u_{t-1}^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 6.1.6$$

Then, equation 6.1.5 can be rewritten as:

$$\rho_{u_t,\, u_{t-1}} = \frac{\sum u_t \; u_{t-1}}{\sqrt{\left(\sum u_{t-1}^2\right)^2}}$$

$$= \frac{\sum u_t \; u_{t-1}}{\sum u_{t-1}^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 6.1.7$$

Now, we can see that equations 6.1.4 and 6.1.7 are similar, and this is the reason why in most text books, the simple first-order autoregressive scheme is given as:

$$u_t = \rho u_{t-1} + \in_t$$

Then, the complete form of the first-order Markov process (i.e. the pattern of autocorrelation for all the values of U) can be developed as follows.

We know that

$$u_t = f\left(u_{t-1}\right) = \rho\, u_{t-1} + \in_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 6.1.8$$

Then,

$$u_{t-1} = f\left(u_{t-2}\right) = \rho u_{t-2} \in_{t-1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 6.1.9$$

$$u_{t-2} = f\left(u_{t-3}\right) = \rho u_{t-3} + \in_{t-2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 6.1.10$$

$$u_{t-3} = f(u_{t-4}) = \rho u_{t-4} + \in_{t-3} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 6.1.11$$

$$u_{t-r} = f\left(u_t - (r+1)\right) = \rho\, u_t - (r+1) + \in_{t-r} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 6.1.12$$

Thus, in order to define the error term in any particular period t for the first-order Markov process, we work as follows. We start from the autocorrelation relationship in period t, which is given as

$$u_t = \rho u_{t-1} + \in_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 6.1.8^*$$

Substituting equation 6.1.9 into equation 6.1.8* yields

$$u_t = \rho \left[\rho\, u_{t-2} + \in_{t-1}\right] + \in_t \text{ , since } \quad u_{t-1} = \rho u_{t-2} + \in_{t-1}$$

$$= \rho^2\, u_{t-2} + \rho \in_{t-1} + \in_t \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{ 6.1.13}$$

Substituting equation 6.1.10 into equation 6.1.13, we get

$$u_t = \rho^2 \left[\rho u_{t-3} + \in_{t-2}\right] + \rho \in_{t-1} + \in t$$

$$u_t = \rho^3\, u_{t-3} + \rho^2 \in_{t-2} + \rho \in_{t-2} + \in_t \ \ldots\ldots\ldots\ldots\ldots \text{ 6.1.14}$$

Plugging in equation 6.1.11 into equation 6.1.14, we get

$$u_t = \rho_3 \left(\rho u_{t-4} + \in_{t-3}\right) + \rho^2 \in_{t-2} + \rho_{\in t-1} + \in_t$$

$$= \rho^4\, u_{t-4} + \rho^3 \in_{t-3} + \rho^2 \in_{t-2} + \rho \in_{t-1} + \in_t \ldots\ldots\ldots 6.1.15$$

If we continue with the substitution process for r periods lag (where r is large), we find that

$$u_t = \rho^r\, u_{t-r} \in_t + \rho \in_{t-1} + \rho^2 \in_{t-2} + \rho^3 \in_{t-3} + \ldots \qquad 6.1.16$$

As it is common to any correlation coefficient, $|\rho| < 1$. Then, as the power of ρ increases to infinity, the term with the lagged u_t, (i.e. $\left[\rho^r u_{t-r}\right]$), tends to zero. In other words, as $r \to \infty$, $\rho^r \to 0$.

Thus, u_t in equation 6.1.16 can be written as

$$u_t = \in_t + \rho \in_{t-1} + \rho^2 \in_{t-2} + \rho^3 \in_{t-3} + \ldots$$

$$= \sum_{r=0}^{\infty} \rho^r \in_{t-r} \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{ 6.1.17}$$

Thus, equation 6.1.17 gives the value of the error term when it is autocorrelated with a first-order autoregressive scheme.

Finally, we will establish the mean, variance and covariance of u_t; auto correlated with a first-order autoregressive scheme.

1. **Mean of auto-correlated** u_t

The mean of u_t is its expected value. Hence, to find the mean of u_t, we take the expected value of both sides of equation 6.1.17.

$$\Rightarrow E(u_t) = E\left[\sum_{r=0}^{\infty} \rho^r \in_{t-r}\right]$$

$$= \sum_{r=0}^{\infty} \rho^r E(\in_{t-r}) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 6.1.18$$

But by assumption, $E(\in_{t-r}) = 0$. Therefore, equation 6.1.18 becomes

$$E(u_t) = 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 6.1.19$$

It is evident from equation 6.1.19 that the mean of an auto-correlated stochastic term, u_t, is zero if the autocorrelation is simple first-order autoregressive scheme.

2. **Variance of auto-correlated** u_t

From the definition of variance, we know that

$$Var(u_t) = E[u_t - E(u_t)]^2$$

$$= E[u_t]^2 \text{ , since from equation 6.1.19, } E(u_t) = 0$$

$$\Rightarrow Var(u_t) = E(u_t^2) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 6.1.20$$

Substituting the value of u_t from equation 6.1.17, we get

$$Var(u_t) = E\left[\sum_{r=0}^{\infty} \rho^r \in_{t-r}\right]^2$$

$$= E\left\{\left(\in_t + \rho \in_{t-1} + \rho^2 \in_{t-2} + ...\right)^2\right\}$$

$$= E\left\{\in_t^2 + \rho^2 \in_{t-1}^2 + \rho^4 \in_{t-2}^2 + ... + 2(cross\ .products)\right\}$$

$$= E\left\{\sum_{r=0}^{\infty} \left(\rho^r\right)^2 \in_{t-r}^2 + 2(cross\ \ product)\right\}$$

$$= E\left[\sum_{r=0}^{\infty} \left(\rho^r\right)^2 \in_{t-r}^2\right] + 2E\underbrace{(cross\ \ product)}\ldots\ldots\ldots\ldots 6.1.21$$

However, in equation 6.1.2 we assumed that $\in_t$ is purely stochastic and its values are non auto correlated, i.e.,

$E\left(\in_t \in_{t-1}\right) = 0 \Rightarrow E(cross\ \ product) = 0$ in equation 6.1.21

Then,. Equation 6.1.21 becomes

$$Var(u_t) = E\left[\sum_{r=0}^{\infty} \left(\rho^r\right)^2 \in_{t-r}^2\right]$$

$$= \sum_{r=0}^{\infty} \left(\rho^r\right)^2 E\left(\in_{t-r}^2\right)$$

$$= \sigma_{\in}^2 \sum_{r=0}^{\infty} \left(\rho^r\right)^2$$

$$= \sigma_{\in}^2\left[1 + \rho^2 + \rho^4 + \rho^6 + \rho^8 + ...\right]\ldots\ldots\ldots\ldots\ldots 6.1.22$$

Note that the expression in the bracket is the sum of a geometric progression of infinite terms, whose first term is 1 and the common ratio is ρ^2. Following the summation rule of a geometric progression, the sum of the terms in the bracket

converges to $\frac{1}{1-\rho^2}$ since $|\rho| < 1$. Generally, the summation rule of a geometric progression states that the sum of n terms of a geometric progression with the first term of "a" and the common ratio of "λ" is given by the formula.

$$S = \frac{a\left(1-\lambda^n\right)}{1-\lambda}$$

For infinite series (i.e., $(n \to \infty)$) with $|\lambda| < 1$, "S" reduces to $S = \frac{1}{1-\lambda}$

Therefore, using this rule equation 6.1.22 becomes

$$Var\,(u_t) = \sigma_{\in}^2\left(\frac{1}{1-\rho^2}\right)\text{……..……6.1.23}$$

Equation 6.1.23 shows that the variance of an error term is constant even though it is autocorrelated, provided that it is autocorrelated with first order autoregressive scheme.

3. The covariance of auto-correlated u_t, $Cov\,(u_t\ , u_{t-1})$

Recall from equation 6.1.17 that

$$u_t = \sum_{r=0}^{\infty}\rho^r \in_{t-r}$$

$$= \in_t + \rho \in_{t-1} + \rho^2 \in_{t-2} + \ldots \qquad 6.1.24$$

Thus, the value of u_t in one period lag is

$$u_{t-1} = \in_{t-1} + \rho \in_{t-2} + \rho^2 \in_{t-3} + \rho^3 \in_{t-4} + \ldots \qquad 6.1.25$$

By definition, $Cov\,(u_t\ , u_{t-1})$ is given as

$$Cov\,(u_t\ , u_{t-1}) = E\left\{\left[\,u_t - E(u_t)\,\right]\left[\,u_{t-1} - E\left(u_{t-1}\right)\,\right]\right\}\text{……..6.1.26}$$

Since from equation 6.1.19, $E(u_t) = E(u_{t-1}) = 0$ equation 6.1.26 becomes

$$Cov\left(u_t\ ,\ u_{t-1}\right) = E\left\{u_t\, u_{t-1}\right\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 6.1.27$$

Substituting equations 6.1.24 and 6.1.25 into equation 6.1.27, yields

$$Cov\left(u_t\ ,u_{t-1}\right) = E\left\{\left[\in_t\ +\ \rho\in_{t-1}\ +\ \rho^2\ \in_{t-2}\ +\ ..\right]\left[\in_{t-1}\ +\ \rho\in_{t-2}\ +\ \rho^2\in_{t-3}\ +\ ...\right]\right\}$$

$$= E\left\{\left[\in_t + \rho\left(\in_{t-1} + \rho\in_{t-2} + ...\right)\right]\left[\in_{t-1}\ +\ \rho\in_{t-2}\ +\ \rho^2\in_{t-3}\ +\ ....\right]\right\}$$

$$= \underbrace{E\left\{\left(\in_t\right)\left(\in_{t-1}\ \rho\in_{t-2}\ +\rho^2\ \in_{t-3}+...\right)\right\}}_{E(cross\ product)=0} + E\left\{\left[\left(\underbrace{\in_{t-1}+\rho\in_{t-2}\ +...}_{identical}\right)\right]\left[\underbrace{\in_{t-1}+\rho\in_{t-2}\ +\rho_{t-2}^2}_{identical}\right]\right\}$$

$$= \rho E\left\{\left[\in_{t-1}\ +\ \rho\in_{t-2}\ +\ \rho^2\ \in_{t-3}\ +\ \rho^3\in_{t-4}\ +\ ....\right]^2\right\}$$

$$= \rho E\left\{\in_{t-1}^2\ +\ \rho^2\ \in_{t-2}^2\ +\rho^4\in_{t-3}^2\ +\ \rho^6\in_{t-4}+...\ +\ (cross\ \ product)\right\}$$

$$= \rho\left\{E\left(\in_{t-1}^2\right)+\ \rho^2E\left(\in_{t-2}^2\right)\rho^4E\left(\in_{t-3}^2\right)+\rho^6\ E\left(\in_{t-4}\right)+\ ...+\underbrace{E\left(cross\ \Pr oduct\right)}_{=0\ by\ assoation}\right\}$$

$$= \rho\left\{\sigma_\in^2\ +\ \rho^2\sigma_\in^2\ \ +\ \rho^4\sigma_\in^2\ +\ \rho^6\sigma_\in^2\ \ +..\right\}$$

$$= \rho\ \sigma_\in^2\ \underbrace{\left(1+\ \rho^2\ +\ \rho^4\ +\ \rho^6\ +....\right)}_{geometric\ \ progression}$$

$$\Rightarrow Cov\left(u_t\ ,\ u_{t-1}\right) = \ \rho\ \sigma_\in^2\left(\frac{1}{1-\ \rho^2}\right)\ \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ 6.1.28$$

But we know from equation 6.1.23 that

$$Var\left(u_t\right) = \sigma_u^2\ =\ \sigma_\in^2\left(\frac{1}{1-\ \rho^2}\right).$$

Substituting this in equation 6.1.28, we get

$$Cov\left(u_t\ ,\ u_{t-1}\right) = \underline{\underline{\rho\ \sigma_u^2}} \quad \text{.....6.1.29}$$

Following the same procedure to the procedures used above, we find that

$$Cov\left(u_t\ ,\ u_{t-2}\right) = \rho^2\sigma_u^2$$

$$Cov\left(u_t\ ,\ u_{t-3}\right) = \rho^3\ \sigma_u^2$$

. . .
. . .
. . . .

$$Cov\left(u_t\ ,\ u_{t-r}\right) = \rho^r\ \sigma_u^2 \text{ (for } r \neq t\text{)}.$$

From this pattern we can conclude that as the lag, say r, increases (i.e., as the lag gets far apart from the current period) the (spill over) effect of u_{t-r} *on* u_t declines. This is due to the fact that $|\rho| < 1$, and hence ρ^r declines and approaches zero as r approaches infinite

Summary

We conclude that when there is autocorrelation of first-order autoregressive scheme, the auto-correlated disturbance term has the following characteristics.

1) $u_t = \sum_{r=0}^{\infty} \rho^r \in_{t-r}$

2) $E\left(u_t\right) = 0$

3) $Var\left(u_t\right) = \sigma_u^2 = \sigma_{\in}^2\left(\frac{1}{1-\rho^2}\right)$

4) $Cov\left(u_t\ u_{t-s}\right) = \rho^s\ \sigma_u^2 \neq 0$

6.1.4 Consequences of Autocorrelation

1) Even when the residuals are serially correlated, the OLS estimates of the parameters are statistically unbiased

2) With the autocorrelated values of the disturbance term, the variances of the OLS estimates are likely to be larger than those obtained from other econometric methods. This implies that autocorrelation causes the actual (true) variance of $\hat{\beta}_i$ to be large, while the estimates of these variances obtained under the assumption of $E(u_i \ u_j) = 0$ are smaller. Thus, we can say that the variance of the $\hat{\beta}_i$ given by the simple OLS formula will under estimate the true variance of $\hat{\beta}_i$.

3) The variance of the random term may be seriously underestimated if the $u's$ are autocorrelated. This can be another source of the underestimation of the true variance of the parameter estimate, $\hat{\beta}_i$.

4) If the values of u are autocorrelated, the predictions based on ordinary least squares (OLS) will be inefficient, in the sense that they will have larger variance as compared with predictions based on estimates obtained from other econometric techniques.

6.1.5 Tests for Autocorrelation

1) Graphical method

Graphically, some clues about the possible existence of autocorrelation among the values of the error term can be obtained. The graphs used for this purpose can be:

i) **The graphs obtained by plotting the regression residuals against their own lagged values**: In these plotted diagrams, if most of the points fall in quadrants I and III, then it shows that there is positive autocorrelation. On the other hand, if most of the points fall in quadrants II and IV, then it gives the clue that there is negative autocorrelation among the values of the error term. If there is no noticeable pattern, then we can expect that there is no autocorrelation.

ii) **The graphs obtained by plotting the regression residuals against time**: If the values of the regression residuals in successive periods show a regular pattern, we conclude that the error term is autocorrelated. Specifically, if the successive values of the regression residuals change sign frequently, then there will be negative autocorrelation. In contrary, the regression residuals do not change their sign frequently so that several positive values of them are followed by several negative values, then it can be a signal for positive autocorrelation. If they show no pattern, then it can be concluded that there is no autocorrelation.

Generally, the graphical method gives rough idea about the existence and pattern of autocorrelation, which makes the use of other methods compulsory. Thus, in this module we deal with the most common method of detecting autocorrelation: **the Durbin – Watson Test**

2) The Durbin – Watson Test

Durbin and Watson have suggested a test for the possible existence of autocorrelation among the values of the stochastic/error term, which came to be known as **the Durbin – Watson Test.** This test is applicable to small samples as

well as large samples. However, the test is appropriate only for the first order autoregressive scheme, i.e., $u_t = \rho u_{t-1} + \in_t$.

The procedures used in this test may be outlined as follows.
In the first step, state the null hypothesis as shown below

$$H_0 : \rho = 0$$

Against,

$$H_1 : \rho \neq 0$$

In this hypothesis, if H_0 is true it implies that the $u's$ are not autocorrelated with first-order autoregressive scheme. On the other hand, if H_0 is rejected it implies that the $u's$ are autocorrelated with first order autoregressive scheme.

In the Durbin – Watson test**,** to test the null hypothesis that there is no autocorrelation we use the Durbin-Watson "d" statistic, defined as:

$$d = \frac{\sum_{t=2}^{n} \left(\hat{u}_t - \hat{u}_{t-1}\right)^2}{\sum_{t=1}^{n} \hat{u}_t^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{6.1.30}$$

From this, it can be shown that the values of d lie between 0 and 4, and that when $d = 2$, then $\rho = 0$. Thus, in essence testing the $H_0 : \rho = 0$ is equivalent to testing $H_0 : d = 2$.

Expanding the d – statistics in 6.1.30, we get

$$d = \frac{\sum_{t=2}^{n} \left(e_t - e_{t-1}\right)^2}{\sum_{t=1}^{n} e_t^2} \qquad \text{where } e_t = \hat{u}_t \text{ and } e_{t-1} = \hat{u}_{t-1}$$

$$= \frac{\sum_{t=2}^{n} \left(e_t^2 - 2\, e_t e_{t-1} + e_{t-1}^2\right)}{\sum_{t=1}^{n} e_t^2}$$

$$= \frac{\sum_{t=2}^{n} e_t^2 + \sum_{t=2}^{n} e_{t-1}^2 - 2\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}$$

Note that for large samples

$$\sum_{t=2}^{n} \cong \sum_{t=2}^{n} e_{t-1}^2 \cong \sum_{t=1}^{n} e_t^2$$

$$\Rightarrow d \cong \frac{\sum_{t=2}^{n} e_{t-2} + \sum_{t=2}^{n} e_{t-1}^2}{\sum_{t=2}^{n} e_{t-1}^2} - 2\,\frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_{t-1}^2}$$

$$\Rightarrow d \cong 2\left(\frac{\sum_{t=2}^{n} e_{t-1}^2}{\sum_{t=2}^{n} e_{t-1}^2} - \frac{\sum_{t=2}^{n} e_{t-1}}{\sum_{t=2}^{n} e_{t-1}^2}\right)$$

$$\Rightarrow \; d \;\cong\; 2\left(1-\frac{\sum e_t\, e_{t-1}}{\sum e_{t-1}^2}\right)$$

But we know that

$$\hat{\rho} \;=\; \frac{\sum e_t\, e_{t-1}}{\sum e_{t-1}^2}$$

Therefore,

$$\underline{\underline{d \;\cong\; 2\;(1-\hat{\rho})}} \qquad , \quad where\ |\rho| < 1.$$

From this expression we can see that:

1) If there is no autocorrelation $\hat{\rho} = 0$ *and by implication* $d = 2$. This means that if from the sample data we find $d \cong 2$, we accept that there is no autocorrelation in the function.

2) If $\hat{\rho} = +1 \Rightarrow d = 0,$ which implies that there is perfect positive autocorrelation.

3) If $\hat{\rho} = -1$, $d = 4,$ This implies that there is perfect negative autocorrelation

4) If $2 < d < 4$ then it implies that there is some degree of negative autocorrelation which gets stronger as the values of d get higher.

It should be clear that in Durbin-Watson test the null hypothesis of zero autocorrelation, $\rho = 0$ is carried out indirectly, by testing the equivalent hypothesis $d = 0$. Therefore, after formulating the hypotheses, we use the sample residuals ($e_t's$) and compute the empirical value of the Durbin-Watson

d statistic. Then, finally, we compare the empirical d with the theoretical values d that define the critical region of the test, as show in figure 6.1 below.

The problem with this test is that the exact distribution of d is not known. However, Durbin and Watson have established upper limits of d, (d_u) and (d_L), which are appropriate to test the hypothesis of zero first – order autocorrelation against the alternative hypothesis of positive first – order autocorrelation.

Durbin and Watson have tabulated these upper and lower values at the 5 per cent and1 percent level of significance. The tables assume that the $u's$ are normal, homoscedastic and not autocorrelated, and that the $X's$ are truly exogenous.

The test compares the empirical value of d with d_L and d_u in the Durbin-Watson tables and with their transformed values, $(4-d_L)$ and $(4-d_u)$. The comparison using d_L and d_u investigates the possibility of positive autocorrelation; while the comparison with $(4-d_L)$ and $(4-d_u)$ investigates the possibility of negative autocorrelation.

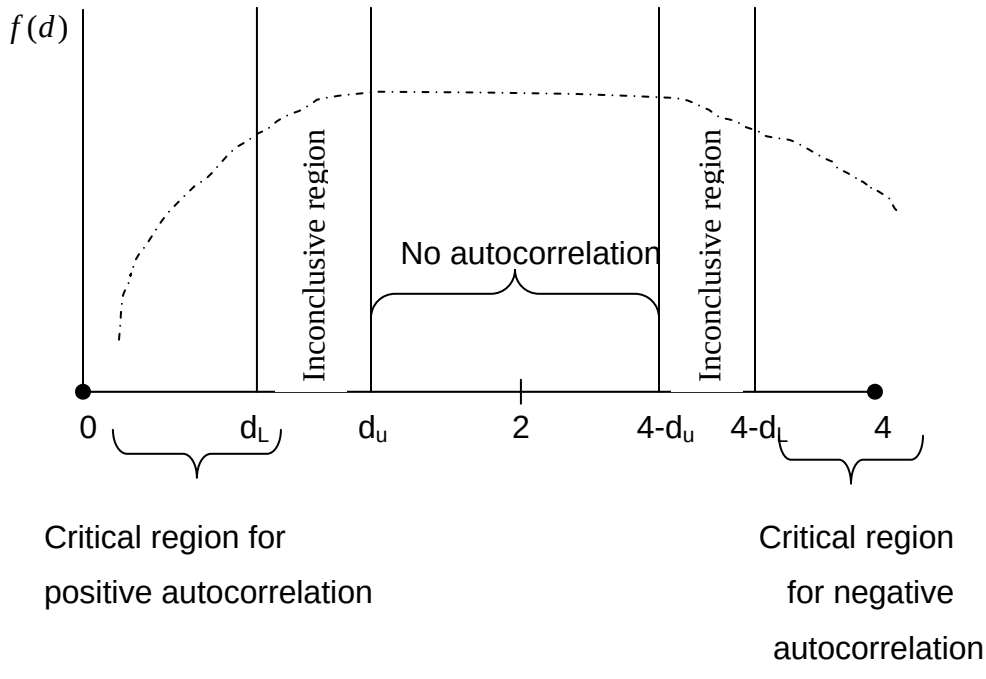


Figure. 6.1 The Distribution of d Statistic

Decision Rule

1) If $d > d_L$, we reject the H_0 hypothesis of no autocorrelation and accept that there is positive autocorrelation of first order.

2) If $d > 4 - d_L$, we reject the H_0 hypothesis of no autocorrelation and accept that there is negative autocorrelation of the first order.

3) If $d_u < d < (4 - d_u)$ we accept the H_0 hypothesis of no autocorrelation.

4) If $d_L < d < d_u$ *or* $4 - d_u < d < 4 - d_L$ the test is inconclusive

Shortcomings of the Durbin – Watson test

1) The d statistic is not an appropriate measure of autocorrelation if among the explanatory variables there are lagged values of an endogenous variable.

2) The test is not appropriate to test higher order serial correlation or for other forms of autocorrelation (e.g., non-linear autocorrelation).

3) Existence of inconclusive region

Remark:

Various writers, including Durbin himself, have suggested alternative tests for serial correlation, which are more accurate and more powerful than the Durbin – Watson test. However these tests are invariably more complicated and costly in computation. Given the short comings of the alternative tests several econometricians have followed the practice of applying the Durbin – Watson test in the following amended form:

Reject $H_0 : \rho = 0 \ \ if \ \ d < d_u \ \ \ or \ \ d > 4 - d_u$ i.e., include the inconclusive region of Durbin-Watson test into the rejection region and accept $H_0 : if \ \ d_u < d < 4 - d_u$

6.1.6 Solutions for Autocorrelation

Before taking remedial measures we should make sure that there is true autocorrelation, i.e., it is not due to omissions of variables and misspecification of the model. For true serial correlations, one of the solutions is the transformation of the original data so as to produce a model whose random variable satisfies the assumptions of classical regression models.

To illustrate, assume the usual first order autoregressive scheme

$$\Rightarrow u_t = \rho u_{t-1} + \in_t$$

Then, the appropriate transformation is to subtract from the original observations of each period the product of $\hat{\rho}$ times the value of the variables in the previous period.

Let the original model be

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + ... + \beta_k X_{kt} + u_t \quad \ldots\ldots\ldots 6.1.31$$

Where,

$u_t = \rho\, u_{t-1} + \in_t$ and $\in_t$ satisfies all the usual assumptions of random term.

From the original model in equation 6.1.31, the relationship for the period $t-1$ is

$$Y_{t-1} = \beta_0 + \beta_1 X_{1t-1} + \beta_2 X_{(t-1)} + ... + \beta_k X_{k(t-1)} + u_{t-1} \ldots\ldots\ldots 6.1.32$$

Multiplying equations 6.1.32 by ρ, we get

$$\rho Y_{t-1} = \rho b_0 + \rho b_1\, x_{1\,(t-1)} + \rho b_2\, x_{2\,(t-1)} + ... + \rho b_k\, x_{k(t-1)} + \rho u_{t-1} \ldots\ldots\ldots 6.1.33$$

Subtracting equation 6.1.33 from equation 6.1.31, we get:

$$Y_t - \rho Y_{t-1} = \beta_0 - \rho\beta_0 + \beta_1 \left(X_{1t} - \rho X_{1t-1}\right) + \beta_2 \left(X_{2t} - \rho X_{2t-1}\right) + ... + \beta_k \left(X_{kt} - \rho X_{k(t-1)}\right) + u_t - \rho u_{t-1}$$

$$\Rightarrow Y^* = \alpha + \beta_1\, X_1^* + \beta_2\, X_1^* + ... + \beta_k\, X_k^* + v_t \ldots\ldots\ldots\ldots\ldots 6.1.34$$

Where

$$Y^* = Y_t - \rho Y_{t-1} \ , \quad \alpha = \beta_0 (1-\rho) \ , \ X_1^* = X_{1t} - \rho X_{1t-1} \quad \text{and} \ v_t = u_t - \rho u_{t-1}$$

Note that with the above transformation we are able to retain only $n-1$ observations in our analysis since we loose one observation in the process. To avoid this loss, we use the following transformation of the first observation of the variables:

$$Y_1^* = Y_1 \sqrt{1-\rho^2}$$

$$X_{j1}^* = X_{j1} \sqrt{1-\rho^2} \quad , \quad j = 1 \ , \ 2 \ , \ .... \ k$$

In most applied econometric research, when autocorrelation is suspected, the investigator makes some reasonable assumptions (guesses) about the value of the autoregressive coefficient (ρ) . However, the usual case is to assume that $\rho = 1$. Under this assumption the appropriate transformation is to take the first differences of the original data and apply ordinary least squares to the transformed model as given below.

$$Y_t - Y_{t-1} = \beta_1 (X_{1t} - X_{1t-1}) + \beta_2 (X_{2t} - X_{2t-1}) + ... + \beta_k (X_{kt} - X_{kt-1}) + v_t$$

Where $v_t = u_t - u_{t-1}$

6.2 Heteroscedasticity: What happens if the error variance is not constant?

6.2.1 Introduction

The third assumption of linear regression models about the random error term, u is that its probability distribution remains the same over all observations of X. This means that the variance of each u is the same for all values of the explanatory variable.

$$\Rightarrow Var(u_i) = E\left\{(u_i - E(u_i))^2\right\}$$
$$= E(u_i)^2$$
$$= \delta_u^2 \Rightarrow cons\tan t$$

This assumption is called the assumption of homoscedasticity or constant variance of $u's$. If it is not satisfied in any particular case, we say that the $u's$ are heteroscedastic.

The meaning of the assumption of heteroscedasticity is that the variation of each u_i around its zero mean does not depend on the values of X. The variance of each u_i remains the same irrespective of small or large values of the explanatory variable, i.e., δ_u^2 *is not a function of* X_i

$$\Rightarrow \delta_u^2 \neq f(X_i).$$

If δ_u^2 is not constant, but its values depend on the values of X, we may write it as:

$$\delta_u^2 = f(x_i)$$

The case of Heteroscedasticity is shown by increasing or decreasing dispersion of the observations from the regression line. The pattern of the observations on a scatter diagram depends on the form of Heteroscedasticity, i.e., on the form of the relationship between δ_{ui}^2 and $X_{i.}$

Generally, we can encounter three types of Heteroscedasticity:

i. ***Increasing Heteroscedasticity:*** *This is the case of (monotonically) increasing variance of the stochastic term,* u, i.e., as X increases, so

does the variance of u. This is the common form of heteroscedasticity assumed in econometric applications.

ii. ***Decreasing Heteroscedasticity*:** As X assumes higher values the deviation of the observations from the regression line decreases, i.e., the variance of the random variable changes in the opposite direction to the explanatory variable.

iii. ***Cyclical Heteroscedasticity:*** The variance of u decreases initially as X assumes higher values, but after a certain level of X, the variance of u increases with X.

Generally, it should be clear that the pattern of heteroscedasticity depends on the signs and values of the coefficients of the relationship between δ_{ui}^2 and $X_{i.}$

$$\Rightarrow \delta_{ui}^2 = f(x_i).$$

However, since the $u's$ are not observable, we do not know the true pattern of heteroscedasticity. In applied research econometricians usually make the assumption that Heteroscedasticity, if exists, will take the following form

$$\delta_{ui}^2 = K^2 X_i^2$$

Where, $K_i's$ are constants to be estimated from the model.

6.2.2 Plausibility of the Assumption of Homoscedasticity.

Dear students, do you think that the variance of the stochastic term is constant in real world applications? ----------

---

In many econometric applications the assumption of constant variance of the random variable may well be expected not to hold. Why?

1. As u expresses the influence of measurement errors on the values of the dependent variable, there is a cogent reason for expecting the variance of u to vary over time, in most cases. For example as Y increases, errors of measurement tend to increase, because it becomes more difficult to collect data, and check its consistency and reliability. Furthermore, the errors of measurement tend to be cumulative over time, so that their size tends to increase. In this case the variance of u increases with increasing values of X.
2. The sampling techniques and data collection methods may continuously improve over time, and thus errors of measurement may decrease, in which case δ_{ui}^2 decreases over time.
3. Many of the variables omitted from the model tend to change in the same direction with X. Thus, they cause an increase in the variation of the observations of the dependent variable from the regression line, which implies an increase in the variance of u as the values of X increase.
4. On a priori grounds there are reasons to believe that the assumption of Heteroscedasticity may often be violated in practice. For example, in

estimating the savings function from a cross-section of family budget 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. High income families tend to stick to a certain standard of living and when their income falls they cut down their savings rather than cutting down their consumption expenditure. On the other hand, low income families save for certain purposes, and thus their saving patterns are more regular. This implies that at high incomes the $u's$ will be high, while at low incomes the $u's$ will be small.

6.2.3 Consequences of Heteroscedasticity

1. The usual formulae of the variances of the coefficients are not appropriate to conduct tests of significance and construct confidence intervals. The tests are inapplicable.

2. If u is heteroscedastic, the OLS estimates do not have the minimum variance property in the class of unbiased estimators, that is, the OLS estimates are inefficient.

3. The coefficient estimates would still be statistically unbiased, albeit the $u's$ are heteroscedastic. This is because the unbiasedness property of the least squares estimates does not require that the variance of the $u's$ be constant.

4. The prediction (of Y for a given value of X) based on OLS estimates from the original data would have a high variance, and hence the prediction would be inefficient.

6.2.4 Tests for Heteroscedasticity

From the discussion we had in section 6.2.3, we have seen that if there is heteroscedasticity in the model, estimation, statistical inferences, predictions etc will be futile. Hence, before proceeding with any statistical or econometric analysis, we have to figure out whether there is the problem of heteroscedasticity. The most commonly used tests to detect the possible existence of heteroscedasticity are discussed as follows.

A. The Spearman rank-correlation test

This is the simplest test of heteroscedasticity. It can be applied to small samples as well as large samples. The steps to conduct this test are outlined as follows.

Step 1. Formulate the null hypothesis of homoscedasticity against the alternative hypothesis of heteroscedasticity as:

$$H_0 : \rho^s = 0$$
$$H_1 : \rho^s \neq 0$$

Where, ρ^s is the population rank correlation coefficient and its sample counterpart is r^s

Step 2. Regress Y on X as shown below

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad \text{.....6.2.1}$$

Then, obtain the residuals, $e's$ which are the estimates of the $u's$ from equation 6.2.1.

Step 3. Order the $e's$ (ignoring their sign) and the X value in ascending or descending order and compute the rank correlation coefficient between e and X.

The Spearman rank-correlation coefficient, r^s, is given as:

$$r^s_{e,x} = 1 - \frac{6\sum D_i^2}{n(n^2-1)} \quad \text{.....6.2.2}$$

Where, D_i is the difference between the ranks of corresponding pairs of e and X and n is the no of observations in the sample.

Step 4. Compute the value of t ($t^s{}_{calculated}$) form equation 6.2.3.

Assuming that the null hypothesis is true (i.e., population rank correlation coefficient, ρ^s is zero) and $n > 8$, the significance of sample rank correlation coefficient, r^s, can be tested by using t test. The formula used to obtain the value of t for this purpose is given as:

$$t^s{}_{calculated} = \frac{r^s\sqrt{n-2}}{\sqrt{1-(r^s)^2}} \quad \text{.....6.2.3}$$

Step 5. Make decision by using the following rule

Decision Rule

Reject the null hypothesis of homoscedasticity, if the value of t obtained from equation 6.2.3 ($t^s{}_{calculated}$) is greater that the value of t obtained from t tables for

degree of freedom of $n-2$, and conclude that there is the problem of heteroscedasticity.

Note that if we have a model with more than one explanatory variable, we may compute the spearman's rank correlation coefficient between e and each one of the explanatory variables separately.

B. The Goldfeld and Quandt test

This test is mainly applicable when the sample size is large. To apply this test the number of observations on the variables included in the model must be at least twice as many as the parameters to be estimated from the model.

The following steps can be followed to conduct the **Goldfeld and Quandt test.**

Step 1. Formulate the hypotheses to be tested.

The customary hypotheses used in the Goldfeld and Quandt test are

H_0 : The $u's$ are homoscedastic

H_1 : The $u's$ are heteroscedastic

Step 2. Order the observations according to the magnitude of the explanatory variable X.

Step 3. Select arbitrarily a certain number, c, of central observations to be omitted from the analysis.

Generally, for samples larger than 30 ($n > 30$), the optimum number of central observations to be omitted from the test is approximately a quarter of the total observations. For example, we omit 8 central observations for $\mathrm{n} = 30$, 16 central observations for $n = 60$ and so on.

Then, divide the remaining $(n-c)$ observations into two sub-samples of equal size, $\frac{n-c}{2}$; where one includes the small values of X and the other includes the large values of X.

Step 4. Fit separate regressions to each sub-samples in step 2 above and obtain the sum of squared residuals ($\sum e^2$) from each regressions.

Let $\sum e_1^2$ be the sum of residuals obtained from the sub-sample of low values of X, with $\left[\frac{n-c}{2}\right]-k$ degrees of freedom and $\sum e_2^2$ be the sum of residuals from the sub-sample of high values of X, with the same degree of freedom, $\left(\frac{n-c}{2}\right)-k$, where k is the total number of parameters in the model.

Then, if each of these sums of squared residuals is divided by the appropriate degrees of freedom, we obtain the estimates of the variances of the $u's$ in the two sub-samples. Furthermore, the ratio of the two variances given as

$$F=\frac{\sum e_2^2 \Big/ \left(\frac{n-c}{2}\right)-k}{\sum e_1^2 \Big/ \left(\frac{n-c}{2}\right)-k}=\frac{\sum e_2^2}{\sum e_1^2} \quad \text{.....6.2.4}$$

has an $F-$distribution with $v_1=v_2=\left(\frac{n-c}{2}\right)-k$ degrees of freedom.

If the two variances are equal (i.e., if the $u's$ are homoscedastic) the value of F in equation 6.2.4 will tend to 1. on the other hand, if the two variances differ, the the F will have a large value, given that by the design of the test $\sum e_2^2 > \sum e_1^2$).

Step 4. Make decision

To make decision on the status of the null hypothesis, compare the observed value of F obtained from equation 6.2.4 with the theoretical value of F ($F_{tabulated}$) for $v_1 = v_2 = \left(\frac{n-c}{2}\right) - k$ degrees of freedom and the chosen level of significance.

Decision Rule

Reject the H_0 of homoscedasticity (i.e., no difference in the variances of the two groups), if $F > F_{tabulated}$; otherwise we can conclude that there is heteroscedasticity problem in the model.

Note: The higher the observed value of F, the stronger would be the heteroscedasticity problem.

C. The Glejser test

To detect the possible existence of heteroscedasticity using the Glejser test, we commonly use the following steps.

Step 1. Regress of the dependent variable, Y, on all the explanatory variables and compute the regression residuals, $e's$

Step2. Then regress the absolute values of $e's$, ($|e|$), on the explanatory variable with which δ_{ui}^2 is thought, on a priori grounds, to be associated.

The actual form of this regression is usually not known, so that one may experiment with various formulations, containing various powers of X,
For example, we can regress

$$|e| = a_0 + a_1 X_i^2$$

$$|e| = a_o + a_1 X_i^{-1} = a_0 + a_1 \frac{1}{X_i}$$

$$|e| = a_0 + a_1 X^{1/2}{}_i = a_0 + a_1 \sqrt{X_i} \text{ and so on.}$$

Finally, among these regressions, we choose the model that gives the best fit in the light of the correlation coefficient and the standard errors of the coefficients a_0 and a_1. Then, by using the coefficients (a_0 and a_1) of the chosen model, we can decide on the existence of heteroscedasticity problem as follows.

(i) If $a_0 = 0$, while $a_1 \neq 0$, it implies that there is pure heteroscedasticity.

(ii) If both $a_0 \neq 0$ and $a_1 \neq 0$, it implies that there is mixed heteroscedasticity

Note that heteroscedasticity is judged in the light of the statistical significance of a_0 and a_1. In other words, we perform the usual standard test of significance for these coefficients, and if they are found to be significantly different from zero we reject the null hypotheses of homoscedasticity.

The Glejser test has the advantage that it gives also information on the form of heteroscedasticity, i.e., it gives information on the particular way in which δ_{ui}^2 is connected with X. This information is crucial for the correction of the heteroscedasticity of the disturbance term.

6.2.5 Remedial Measures for Heteroscedasticity

When the existence of the problem of heteroscedasticity is established on the basis of any of the tests we have discussed in section 6.2.4, the appropriate solution is to transform the original model in such a way as to obtain a form in which the transformed disturbance term has constant variance. We, then, apply the method of classical least squares to the transformed model.

The transformation of the model reduces to the adjustment of the original data. However, the transformation of the model depends on the particular form of

heteroscedasticity, i.e., on the form of the relationship between the variance of u_i, δ_{ui}^2, and the values of the explanatory variable(s). ($\delta_{ui}^2 = f(x_i)$)

In general, the transformation of the original model consists dividing through the original relationship by the square root of the term which is responsible for the problem of heteroscedasticity.

To illustrate, assume that original model is

$$Y_i = \beta_0 + \beta_1 X_i + u_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 6.2.1^*$$

Where, u_i is heteroscedastic, but satisfies all other assumption of the stochastic term.

Case 1. Suppose the heteroscedasticity is of the form

$$E(u_i^2) = \delta_{ui}^2 = k^2 X_i^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 6.2.5$$

This equation implies that the variance of u_i increases proportional with X_i^2.

Then, solving for the constant factor of proportionality, k^2, from equation 6.2.5, we get

$$k^2 = \delta_{ui}^2 \Big/ X_i^2 .$$

This suggests that the appropriate transformation of the original model consists of the division of the original relationship by $\sqrt{X_i^2} = X_i$, which means that the appropriate transformation version is

$$\frac{Y_i}{X_i} = \frac{\beta_0}{X_i} + \beta_1 + \frac{u_i}{X_i}$$

$$Y_i^* = \beta_o \, 1 \Big/ X_i + \beta_1 + u_i^* \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 6.2.6$$

Where, $Y_i^* = \frac{Y_i}{X_i}$ and $u_i^* = \frac{u_i}{X_i}$

Note that u_i^* in the transformed model given in equation 6.2.6 is homoscedastic. Why?

Proof

The variance of u_i^* can be given as

$$E(u^{*2}) = E(\frac{u_i}{X_i})^2 = 1/X^2{}_i \, E(u_i{}^2)$$

$$= \frac{1}{X_i^2}. k^2 \; X_i^2 \quad \text{Since from equation 6.2.5, } \delta_{ui}^2 = k^2 \; X_i^2$$

$$= k^2, \text{ which is constant.}$$

Therefore, we can conclude that through the above transformation, we solved the problem of heteroscedasticity. Consequently, we can apply classical least squares to the transformed version of the model.

Case 2. Assume the form of Heteroscedasticity to be

$$E\left(u_i^2\right) = \delta_{ui}^2 = k^2 X_i$$

Then, the transformation of the original model consists of the division of the original relationship given in equation 6.2.1* by $\sqrt{X_i}$ as follows

$$\Rightarrow \frac{Y_i}{\sqrt{X_i}} = \frac{\beta_0}{\sqrt{X_i}} + \beta_1 \frac{X_i}{\sqrt{X_i}} + \frac{u_i}{\sqrt{X_i}}. \qquad 6.2.7$$

Note that the transformed random term $u_i/\sqrt{X_i}$ in the transformed model given in equation 6.2.7 is homoscedastic with constant variance equal to k^2.

Proof

The variance of the stochastic term in the transformed model can be given as

$$E\left(\frac{u_i}{\sqrt{X_i}}\right)^2 = E\left(\frac{u_i^2}{X_i}\right)$$

$$= \frac{1}{X_i} E\left(u_i^2\right)$$

$$= \frac{1}{X_i} k^2 X_i$$

$$= k^2$$

Note that k^2 is constant. Hence, we can conclude that with the above transformation we solved the problem of heteroscedasticity, and hence we can estimate the transformed model with OLS.

Case 3. Assume the form of heteroscedasticity to be

$$E\left(u_i^{\,2}\right) = \delta_{ui}^2 = k^2 \left(a_0 + a_1 X_i\right)^2$$

Then, the appropriate transformation will be to divide the original model by

$$\sqrt{\left(a_0 + a_1 X_i\right)^2} = a_0 + a_1 X_i$$

$$\Rightarrow \frac{Y_i}{a_0 + a_1 X_i} = \frac{\beta_0}{a_o + a_1 X_i} + \frac{\beta_1}{a_0 + a_1 X_i} X_i + u_i \Big/ a_0 + a_1 X_i \quad \text{……..6.2.8}$$

Now we can show that the stochastic term of the transformed model given in equation 6.2.8, has constant variance.

Proof

The variance of the stochastic term in the transformed model is given as

$$E\left[u_i \Big/ a_0 + a_1 X_i\right]^2 = \frac{1}{\left(a_0 + a_1 X_i\right)^2} E\left(u_i\right)^2$$

$$= \frac{1}{\left(a_0 + a_1 X_i\right)^2} . \delta_{ui}^2$$

$$= \left(\frac{1}{a_0 + a_1 x_i} \right)^2 k^2 \left(a_{0_1} + a_1 X_i \right)^2$$

$$= k^2$$

Hence, with the above transformation we solved the problem of heteroscedasticity, and hence we can estimate the transformed model with OLS.

In general, if heteroscedasticity is of the form

$$E\left(u_i^2\right) = \delta_{ui}^2 = k^2 f(X_i)$$

Then the solution will be the transformation of the original model by dividing it through by $\sqrt{f(X_i)}$.

6.3 Multi-collinearity: What happens if the regressors are correlated?

6.3.1 Definition

The assumption of no perfect multicollinearity among the explanatory variables is a crucial condition for the application of the least squares method to estimate the parameters of a model. The term multicollinearity is used to denote the presence of linear relationship among the explanatory variables. That means the term does not rule out non-linear relationships among regressors.

For example,

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$$

does not violate the assumption of no-multicollinearity. Furthermore, the assumption of no perfect multicollinearity implies that the explanatory variables (regressors) should not be perfectly linearly correlated, i.e., $r_{x_i, x_j} \neq 1$. This is because as shown in section 6.3.3 below, if the explanatory variables are perfectly linearly related, the parameters become indeterminate; it is impossible

to obtain numerical values for each parameter separately and the method of least squares breaks down.

At the other extreme if the explanatory variables are not intercorrelated at all (that is if the correlation coefficient between them is equal to zero), then the variables are called orthogonal. In such cases there is no need to perform multiple linear regression analysis, and thus, each parameter can be estimated by a simple regression of Y on the corresponding regressor.

In practice neither of the above cases is often met. In most cases there is some degree of intercorrelation among the explanatory variables due to the interdependence of many economic magnitudes over time. In this event the simple correction coefficient for each pair of explanatory variables will have a value between zero and unity. As this value approaches unity, multicollinearity gets stronger and it impairs the accuracy and stability of the parameter estimates. Finally, note that multicollinearity is not a condition that either exists or does not exist in economic functions, but rather it is a phenomenon inherent in most relationships due to the nature of economic magnitudes.

6.3.2 Sources of multicollinearity.

i) A tendency of economic variables to move together over time

ii) The use of lagged values of some explanatory variables as separate independent factors in the relationship.

6.3.3 Consequences of Multicollinearity.

The consequence of multicollinearity can be observed under two categories of multicollinearity.

1. If the intercorrelation between the explanatory variables is perfect, i.e., if $r_{x_i,x_j} = 1.$

Then, the consequences of this type of multicollinearity are:

a) The estimates of the coefficients will be indeterminate.

To illustrate, suppose that the relationship to be estimated is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \quad \text{.....6.2.9}$$

and that X_1 and X_2 are related with the exact relationship given below

$$X_1 = kX_2 \quad \text{.....6.2.10}$$

Where, k is any arbitrary constant.

Now, recall that the OLS estimates of the parameters of equation 6.2.10 are given as

$$\hat{\beta}_1 = \frac{\sum x_{1i} y_i \sum x_{2i}^2 - \left(\sum x_{2i} y_i\right)\left(\sum x_{1i} x_{2i}\right)}{\sum x_{1i}^2 \sum x_{2i}^2 - \left(\sum x_{2i}\right)^2} \quad \text{.....6.2.11}$$

$$\hat{\beta}_2 = \frac{\sum x_{2i} y_i \sum x_{1i}^2 - \sum x_{1i} y_{1i} \left(\sum x_{1i} x_{2i}\right)}{\sum x_{1i}^2 \sum x_{2i}^2 - \left(\sum x_{1i} x_{2i}\right)^2} \quad \text{.....6.2.12}$$

Substituting $X_1 = kX_2 \Rightarrow x_1 = kx_2$ (where $x_1 = X_1 - \overline{X}_1$ and $x_2 = X_2 - \overline{X_2}$) into equation 6.2.11, we get

$$\Rightarrow \hat{\beta}_1 = \frac{\sum kx_{2i} y_i \sum x_{2i}^2 - \left(\sum x_{2i} y_i\right)\left(\sum (x_{2i}) kx_{2i}\right)}{\sum (kx_{2i})^2 \sum {x_{2i}}^2 - \left(\sum (kx_{2i})(x_{2i})\right)^2}$$

$$= \frac{k\sum x_{2i} y_i \sum x_{2i}^2 - k\left(\sum x_{2i} y_i\right)\sum x_{2i}^2}{k^2\left[\left(\sum x_i^2\right)^2 - k^2\left(\sum x_{2i}^2\right)\right]^2}$$

$$= \frac{k(0)}{k^2(0)}$$

$$= \frac{0}{0}$$

This shows that if there is perfect multicollinearity among the explanatory variables, the OLS estimate of β_1 would be indeterminate.

By Substituting $X_1 = kX_2 \Rightarrow x_1 = kx_2$ and following the procedures used to show the indeterminacy of the OLS estimate of β_1 above, we can show that the OLS estimate of β_2 would also be indeterminate. In a nutshell, this implies that in the presence of perfect multicollinearity, we can not find out the separate effect of individual variables independently.

Dear students, why do you think would the OLS estimates be indeterminate if there is perfect multicollinearity among the regressors?----------
---

In the previous chapters we discussed how to interpret the OLS estimates of the coefficients of a model like equation 6.2.9. For instance, $\hat{\beta}_1$ gives the rate of change in the average value of Y as X_1 changes by one unit, holding X_2 constant. But if X_1 and X_2 are perfectly collinear as shown in equation 6.2.10, there is no way X_2 can be kept constant: as X_1 changes so does X_2 by the factor $\frac{1}{k}$. What it means, then, is that in the case of perfect multicollinearity one cannot get a unique solution for the individual regression coefficients. What one

can, however, obtain is the combined effect of the explanatory variables on the dependent variable.

b) The standard errors of the estimates become infinitely large

It might be recalled from chapter 5, equations 5.50 and 5.51 that

$$Var(\hat{\beta}_1) = \delta_u^2 \left[\frac{\sum x_{1i}^2}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i}x_{2i}\right)^2} \right] \ldots\ldots\ldots\ldots\ldots\ldots\ldots 6.2.13$$

$$Var(\hat{\beta}_2) = \delta_u^2 \left(\frac{\sum x_{1i}^{\ 2}}{\sum x_{1i}^2 \left(\sum x_{1i}^2\right) - \left(\sum x_{1i}x_{2i}\right)^2} \right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots 6.2.14$$

Now, substituting $X_1 = kX_2 \Rightarrow x_1 = kx_2$ into equation 6.2.13, we get

$$Var(\hat{\beta}_1) = \sigma_u^2 \frac{\sum x_{2i}^2}{\left(\sum x_{1i}^2\right)\sum \left(kx_{2i}\right)^2 - \left(\sum x_{2i}kx_{2i}\right)^2}$$

$$= \sigma_u^2 \frac{\sum x^{2}_{2i}}{k^2\left[\left(\sum x_{2i}^2\right)^2 - \left(\sum x_{2i}^2\right)^2\right]}$$

$$= \frac{\sum x_{2i}^2}{0} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 6.2.15$$

By analogy, in presence of perfect multicollinearity, we can show that the variance of $\hat{\beta}_2$ is

$$Var(\hat{\beta}_2) = \frac{\sum x_{1i}^2}{0} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 6.2.16$$

It is evident from equations 6.2.15 and 6.2.16 that the variances of the estimates become infinite, if there is perfect multicollinearity among the explanatory variables unless $\sigma_u^2 = 0$. However, there is no a priori reason why σ_u^2 should tend to zero when intercorelation of the explanatory variables increases.

2. If there is high but less than perfect multicollinearity

The perfect multicollinearity situation we discussed above is a pathological extreme. Generally, there is no exact linear relationship among the explanatory variables, especially in data involving economic time series. In cases of near to perfect or high multicollinearity, one is likely to encounter the following consequences.

1) OLS estimates will remain BLUE
2) Although BLUE, the OLS estimators have large variances and covariances, making precise estimation difficult.
3) Because of consequence 2, the confidence intervals tend to be much wider, leading to the acceptance of the "zero null hypothesis" more rapidly.
4) Also because of consequence 2, the t-ratio of one or more coefficients tend to be statistically insignificant.
5) Although the t–ratio of one or more coefficients is statistically insignificant, the overall measure of goodness of fit, R^2 can be very high.
6) The OLS estimators and their standard errors can be sensitive to small changes in the data.

In general, the reason why classical linear regression model assumes that there is no multicollinearity among regressors is that if multicollinearity is perfect the regression coefficients of the explanatory variables are indeterminate and their standard errors are infinite. If multicollinearity is less than perfect, the regression coefficients, although determinate, possess large standard errors (in relation to the coefficients themselves), which means the coefficients can not be estimated with great precision or accuracy,

In econometrics we often encounter an important relationship between the variances of the estimators and the collinearity among the regressors of a model. This relationship can be established as follows.

We have already shown that

$$Var(\hat{\beta}_1) = \delta_u^2 \left[\frac{\sum x_{1i}^2}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i}x_{2i}\right)^2} \right]$$

Dividing both the denominator and numerator of this equation by $\sum x_{2i}^2$ gives,

$$Var(\hat{\beta}_1) = \delta_u^2 \left[\frac{\sum x_{2i}^2 \Big/ \sum x_{2i}^2}{\left[\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i}x_{2i}\right)^2\right] \Big/ \sum x_{2i}^2} \right]$$

$$= \frac{\sigma_u^2}{\sum x_{1i}^2 - \left(\sum x_{1i}x_{2i}\right)^2 \Big/ \sum x_{2i}^2} \quad \text{……6.2.17}$$

Multiplying the denominator of equation 6.2.17 by $\frac{\sum x_{1i}^2}{\sum x_{1i}^2}$, gives

$$Var(\hat{\beta}_1) = \frac{\sigma_u^2}{\sum x_{1i}^2 - \sum x_{1i}^2 \dfrac{\left(\sum x_{1i}x_{2i}\right)^2}{\sum x_{1i}^2 \sum x_{2i}^2}}$$

$$= \frac{\sigma_u^2}{\sum x_{1i}^2 \left[1 - \dfrac{\left(\sum x_{1i}x_{2i}\right)^2}{\sum x_{1i}^2 \sum x_{2i}^2}\right]}$$

$$= \frac{\sigma_u^2}{\sum x_{1i}^2 (1 - r_{x_1,x_2}^2)} \quad \text{……..6.2.18}$$

Since $r^2 = \frac{\left(\sum x_{1i}x_{2i}\right)^2}{\sum x_{1i}^2 \sum x_{2i}^2}$

It is evident from equation 6.2.18 that as collinearity increases, the variances of the estimators increase, and in the limit when $r_{x_1,x_2} = 1$, they are infinite. The speed with which variances and covariances of the OLS estimates increase can be seen from the **variance-inflating-factor(**VIF**)**, which is defined as $VIF = \frac{1}{1 - r_{x_1,x_2}^2}$. VIF shows how the variance of an estimator is inflated by the presence of multicollinearity. Therefore, the variance of $\hat{\beta}_1$ in equation 6.2.18, can be written as

$$\Rightarrow Var\ (\hat{\beta}_1) = \frac{\delta_u^2}{\sum x_{1i}^2}.VIF \quad \text{.....6.2.19}$$

The results given in equation 6.2.19, can be easily extended to the K explanatory variables model. In such a model, the variance of the K^{th} coefficient can be expressed as

$$\Rightarrow Var\ (\hat{\beta}_k) = \frac{\sigma_u^2}{\sum x_{ki}^2} VIF \quad \text{.....6.2.20}$$

$$Var\ (\hat{\beta}_k) = \frac{\delta_u^2}{\sum x_{ki}^2}(1 - R_k^2) \quad \text{.....6.2.21}$$

Where, R_k^2 is the R^2 in the regression of X_k on the remaining K-1 regressors.

This suggests that the variance of the K^{th} estimator depends on three ingredients;

1. σ_u^2
2. VIF
3. $\sum x_{ki}^2$

Note: the inverse of the VIF is called **tolerance (**TOL**)**, that is

$$TOL_k = \frac{1}{VIF_k}$$

$$= 1 - R_k^2 \quad \text{.....6.2.22}$$

Therefore, because of the intimate connection between VIF and TOL, one can use them interchangeably.

6.3.4 Detection of multicollinearity

Dear students, so far we have studied the nature and consequences of multicollinearity. *The next question, then, is: How does one know that collinearity is present in any given situation?*

To detect whether there is collinearity in any given situation, one can use the following tests of multicollinearity, i.e., if one or more of the following conditions occur we suspect the problem of multicollinearity to be sever in the given situation.

1. High R^2 but few significant t ratios
2. High pair-wise correlation coefficients among regressors

3. VIF and TOL.

   If $VIF > 10$, then it implies that the problem of multicollinearity to be sever.

4. Auxiliary Regressions

   These regressions are auxiliary to the main regression. They are obtained by regressing one explanatory variable (for example, K^{th} variable) on the remaining $K-1$ explanatory variables, and they are not regressions between the dependent variable and the explanatory variables. Then, after conducting auxiliary regressions, we compute

$$F_j = \frac{R_j^2 / k-3}{(1-\frac{R_j^2}{n-k+2})} \quad \text{……6.2.23}$$

Where, K is the number of explanatory variables in the original model and R_j^2 is the R^2 in the regression of X_j on the remaining K-1 regressors.

Finally, if $F_j > F_{critical}$, then the collinearity between X_j and the remaining K-1 regressors is strong; we expect the problem of multicollinearity to be sever.

Exercise

1. Discusses the consequences of serial correlation of among the values of the stochastic term *u*.
2. Given a sample of 100 observations and 4 explanatory variables, what can you say about autocorrelation if
   a. $d = 1.05$
   b. $d = 1.40$
   c. $d = 2.50$
3. Define the terms homoscedasticity and heteroscedasticity.
4. Explain the consequences of heteroscedasticity on the estimates of the parameters and their variances.
5. For a study of the relationship between consumption expenditure (C) and income (Y) in a hypothetical town, the data collected from 10 families are shown below.

C	100	120	80	100	500	120	50	600	150	250
Y	80	90	80	70	300	75	30	350	70	150

Using the Spearman's rank correlation coefficient, show whether there is heteroscedasticity problem or not.

6. What does multicollinearity mean?
7. Explain the effect of perfect multicollinearity on OLS estimates of the coefficients of a model?
8. Given the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \ ,$$

where *u,* satisfies the usual assumptions of classical linear regression model

a. Show that as degree of linear association between X_1 and X_2 (measured by r_{X_1,X_2}), increases, the variance of the OLS estimate for β_1 will increase keeping other things constant.

Explain the effect of an increase in r_{X_1,X_2} on the probability of accepting the hypothesis $H_0 : \beta_1 = 0$

CHAPTER SEVEN

MODEL SPECIFICATION

Over view

In previous chapters we have dealt with linear regression models and the techniques used to estimate the models. Furthermore, we have seen how to conduct significance tests to the estimates obtained for the coefficients of economic relationships. However, we have done all these based on the assumption that the regression model used in the analysis is "correctly" specified. In practice this is rarely true. Hence, if this assumption is not valid, we encounter the problem of model specification error or model specification bias. Consequently, this chapter is devoted to take a close and critical look at this problem.

At the end of this chapter, students will be able to

- Know how to find the correct model
- Identify different types of model specification errors
- Understand the consequences of specification errors
- Take remedial measures for specification errors
- Evaluate the performance of competing models.

7.1 Attributes of a Good Model

A model chosen for empirical analysis should satisfy the following criteria:

i) **Be data admissible**: this means that predictions made from the model must be logically possible.

ii) **Be consistent with theory**: The model must make good economic sense, i.e., it must be consistent with the economic theories

iii) **Have weakly exogenous regressors**: The explanatory variables used in the model must be uncorrelated with the error term.

iv) **Exhibit parameter constancy**: The values of the parameters obtained from the model should be stable.

v) **Exhibit data coherency**: The residuals estimated from the model must be purely random.

vi) **Be encompassing**: The model should encompass or include all the rival models in the sense that it is capable of explaining their results. In other words, other models cannot be an improvement over the chosen model.

7.2. TYPES OF SPECIFICATION ERRORS

Dear readers, in this part we will see the common types of specification errors. To illustrate the types and sources of specification errors, consider the familiar textbook example of the cubic total cost function given as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + U_{1i} \quad \text{.....7.1}$$

Where Y = total cost of production and X = output.

Suppose that on the basis of the criteria discussed above, we have verified that this model is accepted as a good model. Now, let us see the possible types of specification errors in turn in light of the model given under equation 7.1.

7.2.1. Omission of a relevant variable(s)

This type of error is the one that a researcher commits by omitting important variables from the model. Now, suppose for some reason a researcher decided to use the model in equation 7.2 instead of model 7.1.

$$Y_i = \alpha_0 + \alpha_1 X_i + \alpha_2 X_i^2 + U_{2i} \text{ .....7.2}$$

Since we assumed the model in equation 7.1 to be correct, adopting 7.2 would constitute a specification error. This is because the model in equation 7.2 omits a relevant variable. Consequently, the error term U_{2i} in 7.2 is in fact

$$U_{2i} = U_{1i} + \beta_3 X_i^3 \text{ .....7.3}$$

7.2.2. Inclusion of an unnecessary variable(s)

This type of error comes into existence, when a researcher includes unnecessary variables. To see this type of error, suppose that a researcher uses the following model:

$$Y_i = \lambda_0 + \lambda_1 X_i + \lambda_2 X_i^2 + \lambda_3 X_i^3 + \lambda_4 X_i^4 + U_{3i} \text{ .....7.4}$$

Given that the model in equation 7.1 is the correct one, this model also constitutes a specification error. In this case however, the researcher included **unnecessary or irrelevant variable.** Hence, the error term is in fact

$$U_{3i} = U_{1i} - \lambda_4 X_i^4 \text{ .....7.5.}$$

This is because $\lambda_4 = 0$ in the true model given in equation 7.1

7.2.3 Adopting the wrong functional form

These types of errors are usually committed at the stage of representing economic relationships by mathematical form. Some times a researcher may wrongly use linear model to represent non-linear relationships. For example a researcher may use equation 7.6 to represent the cubic relationship between cost of production (Y) and output produced (X) as given in equation 7.1.

$$\Rightarrow \ln Y_i = \gamma_0 + \gamma_1 X_i + \gamma_2 X_i^2 + \gamma_3 X_i^3 + U_{4i} \text{ .....7.6}$$

This model constitutes a specification bias, and the bias here occurred because of the use of the **wrong functional form.**

7.2.4. Errors of measurement

These errors occur when the measurement of variables involves errors. To illustrate these errors, suppose that a researcher used the following model instead of the model given in equation 7.1.

$$Y_i^* = \beta_0^* + \beta_1^* X_i^* + \beta_2^* X_i^{*2} + \beta_3^* X_i^{*3} + U_{1i}^* \quad \text{.....7.7}$$

Where $Y_i^* = Y_i + \varepsilon_i$, $X_i^* = X_i + w_i$, and ε_i and w_i being the errors of measurement. This model shows that instead of using the true values of the variables, the researcher used their proxies, Y_i^* and X_i^*; which may contain errors of measurement. Therefore, the researcher is bound to commit the **errors of measurement bias.**

7.2.5. Incorrect specification of the stochastic error term

These errors relate to the way the stochastic error term U_i enters the regression model. Consider for instance, the following regression model without the intercept term:

$$Y_i = \beta X_i U_i \quad \text{.....7.8}$$

Note that in this model the stochastic error term enters multiplicatively with the property that $\ln U_i$ satisfies the usual assumptions of the stochastic term. However, suppose that the true model is as given below.

$$Y_i = \alpha X_i + U_i \quad \text{.....7.9}$$

Note that in equation 7.9 the error term enters additively. Hence, although the variables are the same in the two models, the improper stochastic specification of the error term in equation 7.9 will constitute specification error.

7.3. CONSEQUENCES OF MODEL SPECIFICATION ERRORS

After discussing the possible types of specification errors, the next issue is related to their consequences. To keep the discussion simple, we will answer this question in the context of the first two specification errors discussed earlier in section 7.2.1 and 7.2.2 namely, omitting relevant variables, and including unnecessary variables.

7.3.1 Consequences of omitting relevant variables

i) If the omitted variable is correlated with one or more of the included variables, i.e., if the pair wise correlation coefficient between the two is nonzero, then the estimates will be biased and inconsistent.

ii) The estimates of the variance of the disturbance will be incorrect.

iii) The conventionally measured variance of the estimates of the parameters will be a biased estimator of the variance of the true estimators.

iv) The usual confidence interval and hypothesis-testing procedures are likely to give misleading conclusions about the statistical significance of the estimated parameters.

v) The forecasts based on the incorrect model and the forecast (confidence) intervals will be unreliable.

7.3.2. Consequences of Inclusion of an Irrelevant Variable

i) The OLS estimators of the parameters of the "incorrect" model are all unbiased and consistent.

ii) The variance of the stochastic term is correctly estimated.

iii) The usual confidence interval and hypothesis-testing procedures remain valid.

iv) The estimates of the parameters will be generally inefficient, that is, their variances will be generally larger than those obtained from the true model.

7.4. TESTS OF SPECIFICATION ERRORS

Very often specification biases arise inadvertently, perhaps from our inability to formulate the model as precisely as possible because the underlying theory is weak or because we do not have the right kind of data to test the model. Thus, the practical question is not why specification errors are made, but how to detect them. Detection of specification errors is a prerequisite to take remedial measures.

7.4.1. Detecting the Presence of Unnecessary Variables

Suppose a researcher develops a *k*-variable model to explain a certain phenomenon as follows:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + U_i \quad \text{.....7.10}$$

However, suppose that the researcher is not sure whether, say, variable X_k really belongs in the model. Then, how do you think will the researcher find this out?

The simple way to find this out is to test the significance of the estimated β_k with the usual *t* test. However, if he/she is not sure about the relevance of more than one variable, say X_3 and X_4, this can be easily ascertained by the *F* test.

Thus, detecting the presence of an irrelevant variable (or variables) is not a difficult task.

It is, however, very important to remember that in carrying out these tests of significance we have a specific model in mind, which is assumed to be "true". Given that specific model, then, we can find out whether one or more regressors are really relevant by the usual *t* and *F* tests. But note that the *t* and *F* tests should not be used to build a model *iteratively. In other words,* we should not say that initially *Y* is related to X_2 only because its coefficient is statistically significant and then expand the model to include X_3 and decide to keep that variable in the model if its coefficient turns out to be statistically significant, and so on.

7.4.2 Tests for Omitted Variables and Incorrect Functional Form

In regression analysis, we develop a model that we believe captures the essence of the subject under study based on theory prior empirical work. We then subject the model to empirical testing. After we obtain the results, we begin the postmortem, keeping in mind the criteria of a good model discussed earlier. It is at this stage that we come to know if the chosen model is adequate. To determine adequacy of a model, we look at some features of the results, such as the $\bar{R}^2$ value, the estimated *t* ratios, the signs of the estimated coefficients in relation to their prior expectations, the Durbin–Watson statistic, and the like. If these diagnostics are reasonably good, we claim that the chosen model is a fair representation of reality. Nonetheless, if the results do not look encouraging, for example the $\bar{R}^2$ value is too low, very few coefficients are statistically significant or have the correct signs, the Durbin–Watson *d* is too low, then we begin to worry about model adequacy and look for remedies: May be we have omitted an important variable, or have used the wrong functional form, or have not first differenced the time series (to remove serial correlation), and so on. To find out whether model inadequacy is on account of one or more of these problems, we can use some of the following methods.

i) Examination of residuals

In previous chapters we have seen how to use the residuals of a model to examine autocorrelation and hetroscedasticity problems. But these residuals can also be examined, especially in cross-sectional data, for model specification errors, such as omission of an important variable or incorrect functional form. We suspect our model for such errors, if the plot of the residuals exhibits distinct and noticeable patterns. This means that to test for misspecification errors first we have to estimate the model using OLS, obtain the residuals, the plot them and see the patterns of the plots.

ii) The Durbin–Watson test

In chapter 6 we have seen that the Durbin–Watson test can be used to test the autocorrelation problem in the regression model. This test can also be used to detect specification errors.

To use the Durbin–Watson test for detecting model specification error(s), we proceed as follows:

Step 1. From the assumed model, obtain the OLS residuals.

Step 2. If it is believed that the assumed model is mis-specified because it excludes a relevant explanatory variable, say, X_i from the model, order the residuals obtained in step 1 according to increasing values of X_i.

Step 3. Compute the *d* statistic from the residuals ordered in step 2 by the usual *d* formula that we developed in chapter 6, namely

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

Step 4. From the Durbin–Watson tables, if the estimated *d* value is significant, then one can accept the hypothesis of model misspecification.

If this test leads us to accept the hypothesis of misspecification, then inclusion of X_i in our will be the remedial measure.

iii. Ramsey's RESET Test

This is a test of specification error called RESET (regression specification error test) proposed by Ramsey (1969). To illustrate the simplest version of this test (the version that this module concentrates), let us assume that a researcher has modeled cost function as a linear function of output as given below.

$$Y_i = \beta_0 + \beta_1 X_i + U_i \quad \text{.....7.11}$$

where Y = total cost and X = output.

Steps in Ramsey's RESET test

Step 1. From the chosen model (in this case equation 7.11), obtain the estimated values of Y_i (i.e. $\hat{Y}_i$)

Step 2. Rerun to your original model in equation 7.11 and introduce $\hat{Y}_i$ in some form as an additional regressor(s) (i.e., $\hat{Y}_i$ or $\hat{Y}_i^{\,2}$ and so on), where some clue about the form of $\hat{Y}_i$ is obtained by plotting the estimated value of the error term $\hat{U}_i$ against $\hat{Y}_i$ .Suppose, the plot suggests a curvilinear relationship between $\hat{U}_i$ and $\hat{Y}_i$. Then, we run the following model.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 \hat{Y}_i^{\,2} + \beta_3 \hat{Y}_i^{\,3} + U_i \quad \text{.....7.12}$$

Step 3. Obtain R^2 from equations 7.11 and 7.12 independently and call them R^2_{old} and R^2_{new} respectively and calculate the *F* value as given below.

$$F = \frac{(R^2_{new} - R^2_{old}) / K}{(1 - R^2_{new}) / n - J} \quad \text{.....7.13}$$

Where, K is the number of new regressors, n is the sample size and J is number of parameters in the new model (equation 7.12). Then, find out if the increase in R^2 is statistically significant by using F test..

Step 4. Make decision

Decision rule

If the computed *F* value obtained from equation 7.13 is significant, at a given significance level, then accept the hypothesis that the model 7.11 is mis-specified

According to Gujirati, one advantage of RESET is that it is easy to apply, for it does not require one to specify what the alternative model is. But that is also its disadvantage because knowing that a model is mis-specified does not help us necessarily in choosing a better alternative.

Exercise

1. Suppose the "true" model is

$Y_i = \beta_1 X_i + u_i$ but instead of fitting this regression through the origin you routinely fit the usual intercept-present model as shown below

$$Y_i = \alpha_0 + \alpha_1 X_i + u_i$$

Assess the consequences of this specification error

2. Suppose the "true" model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i$$

But we add an irrelevant variable X2 the model and estimate the following model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

a. Would the R^2 and the adjusted R^2 from the second model be larger than that for model 1?
b. Are the estimates of β_0 and β_1 obtained from model 2 unbiased?

CHAPTER EIGHT

DUMMY VARIABLE REGRESSION MODELS

Overview

In empirical analysis there are four types of variables, such as ratio scale, interval scale, ordinal scale and nominal scale variables. In the preceding chapters we, one way or another, have discussed the ratio scale variables in detail. However, this does not mean that these variables are the only variables that the regression models can deal with. Regression analysis can deal with other variable as well. Hence, in this chapter we will mainly deal with nominal scale variables, which are also known **categorical/qualitative or dummy variables**.

At the end of this chapter, students will be able to:

- Define dummy variables
- Create dummy variables
- Know how to use dummy variables for the difference in intercept term
- Apply dummy variables to represent the difference in slope coefficients
- Use dummy variables to explain seasonal differences

8.1 Definition of Dummy Variables

It is customary that in regression analysis the dependent variable is influenced not only by ratio scale variables, but is also affected by nominal scale variables. These nominal variables may include sex, race, religion, nationality, strike, earthquakes, famine, war etc. This means that in regression analysis nominal scale variables are equally important to ratio scale rain. The importance of these variables is evidenced by several studies. For instance, as cited in Gujnati (2004) a study by Bruce and Julie (2000), found that female workers earn less than male workers, holding other factors constant. This clearly shows that qualitative variables can also influence the dependent variable and should explicitly be included among explanatory variables.

Dear readers, how do you think will these variables be included in regression analysis?

To include qualitative variables into empirical analysis, one has to change them into quantitative variables. One way to quantify nominal scale variables is by creating artificial variables that take on values of 1 or 0, where 1 indicates the presence of a certain attribute/quality and 0 indicates the absence of that attribute. This artificial representation is possible since nominal scale variables usually indicate the presence or absence of a certain quality or an attribute. For example, if an investigator encounters ”the sex of a family head” as a nominal variable in a regression analysis, he/she can use 1 to indicate that the family head is male and 0 to designate that the family head is female. By analogy dummy variables for other nominal variables can be created. Variables that assume 0 and 1 artificially are called **dummy variables**. Hence, dummy

variables can be considered as a device to classify data on qualitative variables into mutually exclusive categories such as presence or absence of an attribute. In passing not that dummy variables can also take other values than 0 and 1, where a lineal transformation function $Z = a + bD$ (where, a and b are constants, b $\neq$ 0 and D = 1 or 0) can be used to transform the pair (0,1) to other pairs. This is the reason why qualitative/dummy variables are called nominal scale variables; they have no natural scale of measurement.

Dummy variables can serve different purposes in regression analysis. They can be used to:

i. Allow for differences in intercept terms
ii. Allow for differences in slopes
iii. Allow for seasonal differences

8.2 Dummy Variables for Differences in Intercept Terms

Some times an investigator may deal with a regression analysis to assess the statistical significance of the relationship between quantitative dependent variable and qualitative explanatory variables. In such cases one often uses dummy variables models, with some assumptions about their use. For instance, the implicit assumption in this case is that the regression lines for the different groups differ only in the intercept term but have the same slope coefficients.

To exemplify the application of dummy variables with the above implicit assumption, let us suppose that an investigator wants to analyses the relationship between salary of public school teachers and their years of experience for two groups: **male and female public school teachers**.

This relationship can be represented with tow regression equations as follows:

$$Y_i = \begin{cases} \alpha_1 + \beta x_i + u_i - \ for\ male\ teachers \\ \alpha_2 + \beta x_i + u_i - \ for\ female\ teachers \end{cases} \quad \text{.....8.1}$$

Where Y_i denotes the salary of a public school teacher and X_i represents years of experience of a teacher.

Note that the slopes of the regression equations (β) for both groups are assumed to be the same but the intercepts $(\alpha_1 and\, \alpha_2)$ are assumed to be different.

Graphical representation of equation 8.1 is given as

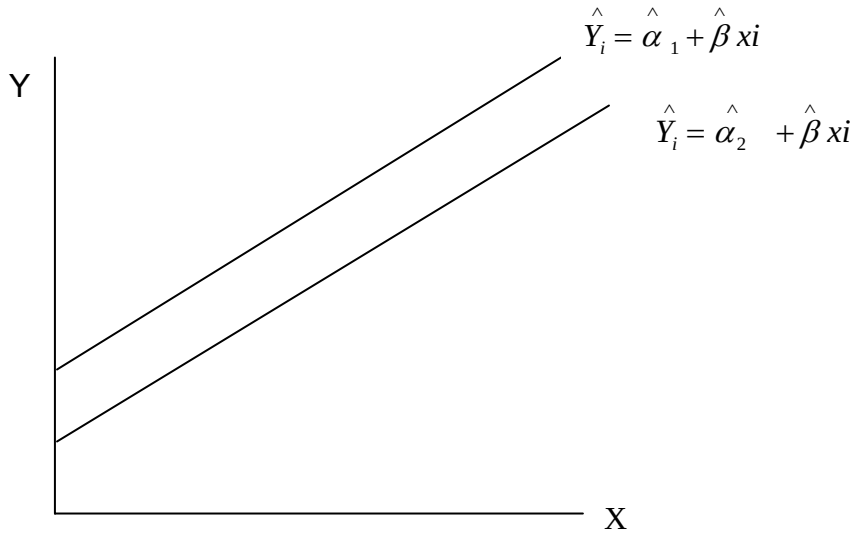


Figure 7.1: Public School Salary (Y) for Male and Female Teachers

Note that in the above figure it is assumed that the intercept terms are different, specifically $\hat{\alpha}_1 > \hat{\alpha}_2$

The two equations given in 8.1, can be combined into a single equation by using a dummy variable for the gender of the teachers as follows

$$Y_i = \alpha_1 + (\alpha_2 - \alpha_1)D + \beta x_i + ui \quad \text{.....8.2}$$

Where $D = \begin{cases} 1\text{ , } if\ \ the\ teacher\ is\ male \\ 0, \ if\ \ the\ teacher\ is\ \ female \end{cases}$

It is evident form equation 8.2 that the coefficient of the dummy variable measures the differences in the two intercept terms. Hence, it is called **differential coefficient**.

It is worth noting that the grouping/qualitative variable in equation 8.1 has two categories. However, the combined equation 8.2 has only one dummy variable. Similarly, if a qualitative variable has three categories, we will introduce only two dummy variables. For instance, suppose that investigator we considered above wants to see whether the salary of the public school teachers depends on the location of the school, where the public schools included in the investigation are from three mutually exclusive regions: the Northern, the Central and the Southern region. In this case we will have three equations representing the three groups as shown below

$$Y_i = \begin{cases} \alpha_1 + \beta x_i + u_i - \ for\ the\ Northern\ regition \\ \alpha_2 + \beta x_i + u_i - \ for\ the\ Centeral\ region \\ \alpha_3 + \beta x_i + u_i - \ for\ the\ Southern\ region \end{cases} \qquad \text{.....8.3}$$

Then, the combined form of equation 8.3 is given as:

$$Y_i = \alpha_1 + (\alpha_2 - \alpha_1)D_1 + (\alpha_3 - \alpha_1)D_2 + \beta x_i + u_i \qquad 8.4$$

Where

$D_1 = \begin{cases} 1, if\ \ the\ teacher\ is\ \ from\ the\ centeral\ region \\ 0, \ otherwise \end{cases}$

$D_2 = \begin{cases} 1, if\ \ the\ teacher\ is\ \ from\ the\ southern\ region \\ 0, \ otherwise \end{cases}$

By substituting the values for D_1 *and* D_2 in equation 8.4, we get the intercepts α_1, α_2 *and* α_3 respectively for the three regions.

Note that to combine the three equations in 8.3 into a sign equation 8.4, we used only two dummy variables D_1 *and* D_2. Generally, if there is a constant term in the regression equation, the number of dummies defined should be one less than the number of the categories of the variable. This is because the constant term is the intercept for the base group. As it can be seen from equation 8.4, the constant term, α_1, measures the intercept for the northern region. Furthermore, the constant term plus the coefficient of D_1 measures. The intercept for the central region and the constant term plus the coefficient of D_2 measures the increment for the third group. This interpretation holds as long as the base group is the northern region. But this should not connote that the base group is always the northern region. The choice of one of the groups as the base group is at the discretion of the investigator. Any one group may be chosen as the base group depending on the preference of the investigator.

On the other hand, if we do not introduce a constant term in the regression equation, we can define a dummy variable for each group, and in this case the coefficients of the dummy variables measure the intercepts for the respective groups. However, if we include both the constant term and dummies for all categories of a variable, we will encounter the problem of perfect multicollinearity and the regression program either will not run or will omit one of the dummies automatically.

In a nutshell, when we are introducing dummy variables, we have to follow the following rule. If the qualitative variable has m categories and the regression equation has a constant intercept, introduce only $m-1$ dummy variables, otherwise we will fall into what is called **the dummy variable trap**, that is, the problem of perfect multicollinearity.

As you might have noted, so far we have been dealing with regression analysis where there is only one qualitative variable. In practice, however, we may have more than one qualitative variable affecting the dependent variable. Then, how will you introduce more than one qualitative variable in your regression analysis? The introduction of qualitative variables in regression analysis is as direct as the introduction of quantitative explanatory variables. However, for each qualitative variable we introduce one on more than one dummy following the rule mentioned above. For example, suppose that we want to analyze the determinates of household consumption (C), and that we have data on the income of the household (Y), the sex of household head, the age of the household head (given in three categories such as less than 30 years, 30 to 60 years and greater than 60 years) and the education of the household head (given in three categories such as less than high school, greater than or equal to high school but less than collage degree and greater than or equal to college degree).

Following the rule of dummy variables, for each variable the number of dummies included in the regression analysis is one less than the number of its categories. Therefore, we can run the following regression equation.

$$C = \alpha + \beta Y + \gamma_1 D_1 + \gamma_2 D_2 + \gamma_3 D_3 D_4 + \gamma_5 + u \text{.....} 8.5$$

Where,

$$D_1 = \begin{cases} 1, \textit{ if sex is male} \\ 0, \textit{ otherwise} \end{cases}$$

$$D_2 = \begin{cases} 1, \textit{ if age} < 30 \textit{ years} \\ 0, \textit{ otherwise} \end{cases}$$

$$D_3 = \begin{cases} 1, \textit{ if age is between } 30 \textit{ and } 60 \textit{ years} \\ 0, \textit{ otherwase} \end{cases}$$

$$D_4 = \begin{cases} 1, \textit{ if educaiton is } 30 \textit{ and } 60 \textit{ years} \\ 0, \textit{ otherwise} \end{cases}$$

$$D_5 = \begin{cases} 1, \textit{ if educaiton is} \geq \textit{high school but} < \textit{college} \deg \textit{ree} \\ 0, \textit{ otherwise} \end{cases}$$

In equation 8.5, the intercept term for each individual is obtained by substituting the appropriate values for *D_1 through D_5.* Furthermore, the coefficients of the dummies are interpreted as the differences between the average consumption of the omitted (base) category and the category represented by the dummy under consideration, keeping other things constant. For instance, γ_1 in equation 8.5 above is interpreted as, keeping other things constant, the average consumption expenditure of male headed households is greater/less than that of their female headed counterparts.

The most important question, perhaps, is how to defer mite where these differences are statically significant. For this purpose we simply test whether the coefficients of the dummies as statistically significant or not by using the usual procedure of hypothesis testing, i.e., by using the standard error of the estimates or the students as test statistics.

Exercise

From equation 8.5, find the intercept if the household head is male whose age is less than 30 years and has collage education.

8.3 Dummy Variables for Changes in Slope Coefficients

So far we have considered how to use dummy variables to allow for differences in the intercept term, and we noted that these dummy variables assume the values 0 and 1. Nonetheless, this does not mean that all dummy variables are of this form. We can also use dummy variables to allow for differences in slope coefficients.

In this case the two regression equations in 8.1 can be written as

$$Y_i = \begin{cases} \alpha_1 + \beta_1 x_i + u_i - \textit{ for male teachers} \\ \alpha_2 + \beta_2 x_i + u_i - \textit{ for female teachers} \end{cases} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 8.6$$

These two equations can be combined together in a single equation as follows by using dummy variables.

$$Y_i = \alpha_1 + (\alpha_2 - \alpha_1)D_1 + \beta_1 x_i + (\beta_2 - \beta_1)D_2 + u_i \quad \text{.....8.7}$$

Where

$$D_1 = \begin{cases} 0, & \textit{for all male teachers} \\ 1, & \textit{otherwise} \end{cases}$$

$$D_2 = \begin{cases} 0, & \textit{for all male teachers} \\ x_i & \textit{otherwise} \end{cases}$$

In equation 8.7, the coefficient of D_1 measures the difference in the intercept terms and the coefficient of D_2 measures the difference in the slope. Estimation of equation 8.7 amounts to estimating the two equations in 8.7 separately, if we assume that the errors have an identical distribution. On the other hand, if we delete D_2 from equation 8.7, this amounts to allowing for different intercepts but not different slopes, and if we delete D_1, it amounts to allowing for different slopes but not different intercepts.

Dummy variable can also be used where there are changes in slopes and intercepts in different times. For instance, suppose we have data for three periods, where in the second period only the intercept has changed and in the third period the intercept and the slope have changed. Then, this can be writhen as;

$$\begin{aligned} Y_{1i} &= \alpha_1 + \beta_1 x_{1i} + u_{1i}, \quad \textit{for period } 1 \\ Y_{2i} &= \alpha_2 + \beta_2 x_{2i} + u_{2i}, \quad \textit{for period } 2 \\ Y_{3i} &= \alpha_3 + \beta_3 x_{3i} + u_{3i}, \quad \textit{for period } 3 \end{aligned} \qquad 8.8$$

Assuming that all the error terms in equation 8.8 have the same distribution, we can combine the equations and write the model as;

$$Y_i = \alpha_1 + (\alpha_2 - \alpha_1)D_1 + (\alpha_3 - \alpha_1)D_2 + \beta_1 x_i + (\beta_2 - \beta_1)D_3 + u_i \qquad 8.9$$

Where,

$$D_1 = \begin{cases} 1, & \textit{for observations in period } 2 \\ 0, & \textit{for other periods} \end{cases}$$

$$D_2 = \begin{cases} 1, & \textit{for observations in periods } 3 \\ 0, & \textit{for other periods} \end{cases}$$

$$D_3 = \begin{cases} 0, & \textit{for observations in periods } 1 \textit{ and } 2 \\ x_3, & \textit{for all observations in period } 3 \end{cases}$$

Finally, we can estimate equation 8.9 by using OLS and conduct significance tests by using the usual procedures.

8.4 Dummy Variables in Seasonal Analysis

Many economic time series based on monthly or quarterly data exhibit seasonal patterns, i.e., they show regular oscillatory movements. For example, household demand for money at holiday times shows seasonal pattern. Hence, it is desirable to remove the seasonal component from a time series so that one can concentrate on the other components, such as the trend. The process of removing the seasonal component from a time series is known as **deseasonalization**, and the series thus obtained is called **deseasonalized** time series. Usually the important economic series such as the unemployment rate, the consumer price index (*CPI)*, the producer price index (*PPI)*, and the industrial production index that we see in different reports are published in their seasonally adjusted form.

One of the methods used to **deseasonalize** a time series is a method of dummy variables. To illustrate the dummy variables technique of deseasonalization,

suppose that we have quarterly data on consumption (*C*) and income (*Y*). In order to deal with this, we fit the following regression equation

$$C = \alpha + \beta Y + \lambda_1 D_1 + \lambda_2 D_2 + \lambda_3 D_3 + u \quad \text{.....8.10}$$

Where, D_1, D_2 *and* D_3 are seasonal dummies defined as

$$D_1 = \begin{cases} 1, & for\ the\ first\ querter \\ 0, & for\ other\ querters \end{cases}$$

$$D_2 = \begin{cases} 1, & for\ the\ \text{sec}ond\ querter \\ 0, & for\ other\ querters \end{cases}$$

$$D_3 = \begin{cases} 1, & for\ the\ third\ querter \\ 0, & fro\ other\ querters \end{cases}$$

Note that in equation 8.10, we have treated the fourth quarter as the base category. Consequently, the coefficients attached to the various dummies are differential intercepts, showing by how much the average value of *C* in the quarter that receives a dummy value of 1 differs from that of the base category, holding the income level constant. In other words, the coefficients on the seasonal dummies will give seasonal increase or decrease in the average value of *C* relative to the base season, for a given value of *Y*. Furthermore, the coefficient of *Y* shows that, allowing for seasonal effects, if income (*Y*) changes by one unit, on average, consumption (*C*) will change by β units.

Dear readers, so far we have seen how to deseasonalize the dependent variable (for example *C* in equation 8.10). Now, an interesting question is: *Just as the dependent variable, C, exhibits seasonal patterns would not the explanatory variable, Y also exhibit seasonal patterns? How, then, do we take into account seasonality in the explanatory variable Y?*

Don't worry! Now you have got the key at your hand, that is, the dummy variables technique. The interesting thing about equation 8.10 is that the dummy variables

in the model not only remove the seasonality in *C* but also the seasonality in *Y,* if any. It is just like killing two birds with one stone.

Final Remarks in the Use of Dummy variables

1. Follow either of the two methods to introduce a dummy variable: (1) introduce a dummy for each category of a qualitative regressor and omit the intercept term or (2) include the intercept term and introduce only ($m - 1$) dummies, where m is the number of categories of the qualitative variable.
2. The category for which no dummy variable is assigned is known as the **base, benchmark, control, comparison, reference, or omitted category.** And all comparisons are made in relation to the benchmark category.
3. The value of the intercept represents the *mean value* of the benchmark category
4. The coefficients attached to the dummy variables are known as the **differential intercept coefficients** because they tell by how much the value of the intercept that receives the value of 1 differs from the intercept coefficient of the benchmark category.

Exercise

1. Explain the following concepts

a. Dummy variables

b. Seasonal dummy variables

c. Dummy variables trap

2. Suppose that the regression program of a graduating class economic student of Arbaminch University refused to estimate four seasonal coefficients when he enters the quarterly data including zero-one dummy variables for each quarter. What is he supposed to do?

3. If you have monthly data over 30 years, how many dummy variables will you introduce in your regression analysis to test the following hypotheses?

a. All the 12 months of the year exhibit seasonal patterns.

b. Only April, June and August and December exhibit seasonal patterns.

CHAPTER I

INTRODUCTION

Overview

In this chapter students will be introduced to the basic concepts of econometrics and how it is related to different branches of science. Furthermore, the chapter will give a brief description of the methodological stages to be followed in econometric researches. Finally, it will discuss the two types of econometrics: theoretical econometrics and applied econometrics.

At the end of this chapter students will be able to:

- Define econometrics and understand how it differs from other branches of science
- Understand the stages in methodology of econometric research
- Differentiate between theoretical and applied branches of econometrics

1.1 What Is Econometrics?

Literally, econometrics means “economic measurement”. However, although measurements are important part of econometrics, its scope is much broader. It is an integration of economics, mathematics and statistics for the purpose of providing numerical values for the parameters of economic relationships and verifying economic theories.

In econometrics general economic theory is formulated in mathematical terms and is combined with empirical measurement of economic phenomena. Thus, econometrics may be defined as a field of knowledge in which tools of economic theory, mathematics and statistics are applied to the analysis of economic phenomena.

1.2 Econometrics and Other Branches of Science

Dear readers, what do you think are the differences and similarities between econometrics and other branches of science such as mathematics, statistics and economics?----------
---

In this part you will be able to discern these nexuses between econometrics and other branches of science.

1.2.1 Econometrics and mathematical economics

Mathematical economics states economic theories in terms of mathematical symbols. Furthermore, it expresses the economic relationships in an exact form. It does not provide numerical values for the coefficients of economic

relationships. On the other hand, econometrics assumes that economic relationships are random or stochastic. In addition, econometrics provides numerical values for the coefficients of economic phenomena.

1.2.2 Econometrics and statistics

Econometrics differs from both mathematical statistics and economic statistics. In economic statistics, we gather empirical data, record them, tabulate them and then attempt to describe the pattern in their development over time (but no attempt is made to explain the phenomena). It is mainly a descriptive aspect of economics. In addition to this, it does not provide measurement of the parameters of economic relations ships. On the other hand, mathematical statistics, deals with methods of measurement which are developed on the basis of controlled experiments in laboratories. However, these statistical methods of measurement are not appropriate for economic relationships and can not be measured on the basis of evidence provided by controlled experiment. However, econometrics differs from both economic and mathematical statistics in that it only uses statistical methods after adapting them to the problems of economic life. The adjustment primary involves specifying the stochastic element.

1.3 Goals of Econometrics

Dear readers, why do you think economists use econometrics?----------
---

In the subsequent part you will learn the major goals of econometrics and the importance of econometrics in economic researches.

i) ***Analysis*:** This involves testing economic theories against economic reality. In other words, this goal deals with verification of economic theories and obtaining empirical evidence to test the explanatory power of economic theories.

ii) ***Policy formulation and decision making:*** This goal deals with obtaining numerical estimates of the economic relationships for policy simulation or decision making. For example the decision of the government about devaluating the currency of a country may depend on marginal propensity to import, and price elasticity of export and import, which would be obtained by applying econometric tools.

iii) ***Forecasting future values of economic magnitudes:*** Forecasts will enable policy makers to judge whether it is necessary to take any measure in order to influence the relevant economic variables in current period. For instance, government's decision on its current employment policy may be influenced by forecasted level of employment for the coming, say, ten years. Succinctly this means that to formulate current employment policy the government has to know (among others):

a) What is the current situation of employment?

b) What the level of employment will be, say in ten years' time if no measure is taken by the government?

Finally, if the forecast value of employment is higher than the expected labor force, inflation will follow. Hence, the government must take measures to reduce this problem. This implies the importance of forecasting future values of economic variables, which only became possible with the advent of econometrics.

1.4 Methodology of Econometric Research

Dear learners, in this part you will learn how econometricians may proceed with their analysis of economic problems and what methodology they may utilize in addressing economic problems via econometric application.

As far as the econometric methodology is concerned, there are several schools of thought. In this part, however, emphasis will be given to the classic (traditional) methodology, which still dominates empirical research in economics and other branches of science. This methodology generally, passes through the following stages:

Stage I. Statement of the theory or hypothesis

In this stage econometricians start by stating the existing economic theories. For example, an investigator may start with, say, the famous Keynesian theory of *marginal propensity to consume.* This theory states that *the marginal propensity to consume is greater than zero but less than one* (Keynes, 1936). Hence, his/her methodology starts by stating this postulate.

Stage II. Specification of the mathematical model

In this stage the investigator expresses economic relationships represented in economic theory in mathematical form. To do this s/he needs to determine:

i) Variables that should be included in the model and which of them are dependant and which of them are independent.
ii) The “a priori” expectation about the sign and size of the parameters.
iii) The mathematical form of the model, i.e., whether the model is linear in parameters, variables or both.
iv) Probabilistic assumptions about variables included in the model.

Returning to the economic theory given in stage one, we can see that the dependent variable is consumption expenditure (say C) and the independent

variable is income (say Y). For simplicity a mathematical economics might assume linearity between the two variables and suggest the following mathematical model.

$$C = \beta_0 + \beta_1 Y \quad \text{.....1.1}$$

Where, β_0 *and* β_1 are known as the parameters of the model.

Furthermore, from the postulate of Keynesian theory of *marginal propensity to consume*, we know that the slope coefficient β_1 lies between 0 and 1, i.e., $0 < \beta_1 < 1$.

This model is called single equation model, as it has only one equation. However, if the model has more than one equation, the model is known as multiple equations model.

Stage III. Specification of the econometric model

The mathematical model developed in stage II above assumes that the relationship between consumption expenditure and income is exact or deterministic. However, as it is common to economic variables this relationship is inexact and subject to individual variations, i.e., it is influenced by other variables which are not explicitly included in the model, such as family size, the age composition of the family, religion of the family etc. Thus, in real life the exact relationship is less likely to occur and hence, purely mathematical model is of little importance for econometricians.

Therefore, to allow for the inexact relationship between economic variables, an econometrician would modify the mathematical model developed in stage II as follows.

$$C = \beta_0 + \beta_1 Y + U \quad \text{.....1.2}$$

Where, U is known as the disturbance term, and it has well-defined probability properties. It represents all those factors that affect consumption but are not explicitly taken into account. Equation 1.2 is a simple example of econometric model, and it hypothesizes that consumption is linearly related to income but the relationship between the two economic variables is inexact: it is subject to individual variations.

Stage IV. Obtaining Data

As the econometrician would be mainly interested in obtaining numerical values to the parameters in equations 1.2 (i.e., the estimates of $\beta_0 \quad and \quad \beta_1$), s/he needs data on the variables included in the model. The data could be collected from either primary or secondary sources.

Note: Students who are not in a position to understand the different sources and types of data should refresh their knowledge on these concepts by referring to their modules for **Introduction to Business Statistics (Econ 242) and Statistics for Economists (Econ 321).**

Stage V. Estimation of the econometric model

In this stage attempts will be made to obtain numerical estimates of the coefficients of the model, ($\beta_0 \quad and \quad \beta_1$). This stage includes:

- Examination of the degree of correlation among explanatory variable (*see chapter 6)*
- Choice of appropriate estimation technique (*see chapter 2*)
- Examination of the assumptions of the chosen technique and their implication for the estimates of the coefficients (*see chapter 2)*
- Examination of the specification condition of the model (*see chapter 7*)

Stage VI. Evaluation of the estimates

Evaluation consists of deciding whether the estimates of the parameters are theoretically meaningful and statistically satisfactory. To evaluate the estimates, we can use the following criteria.

a) Economic criteria: These criteria use the principles of economic theory and refer to the sign and size of the parameters. For example Keynesian liquidity preference theory states that there is positive relationship between money demand and the level of income, and negative relationship between money demand and interest rate; which implies that the sign of the coefficients of income and interest rate in the money demand model is determined by this theory.

b) Statistical criteria: These criteria aim at the evaluation of statistical reliability of the estimates of the parameters. The most widely used statistical criteria are: *the square of the correlation coefficient and the standard error of the estimates:*

i) *The square of the correlation coefficient (r^2):* It shows the percentage of the total variation of the dependent variable explained by the changes of the explanatory variables. It is the measure of the extent to which the explanatory variables are responsible for the changes in the dependent variable (for detailed discussion see chapter 4).

ii) *The standard error of the estimates:* these are measures of the dispersion of the estimates around the true parameter. The larger values of the standard error of the estimates imply that the estimates are statistically less reliable.

Note: *The statistical criteria are secondary to the economic criteria*.

c) **Econometric Criteria:** These criteria aim at the investigation of whether the assumptions of the econometric methods employed are satisfied or not (valid or violated). They help us establish whether the estimates have desirable statistical properties or not.

Stage VII. Evaluation of the forecasting power of the model: This is the final stage and deals with the evaluation of the extra sample performance of the model. Furthermore, it evaluates the stability of the estimates and their sensitivity to changes in the size of the sample.

1.2. Types of Econometrics

Econometrics can broadly be divided into two categories: theoretical and applied econometrics. Theoretical econometrics is concerned with the development of appropriate methods for measuring economic relationships specified by econometric models. Whereas, applied econometrics uses the tools of theoretical econometrics to study some special fields of economics and business, such as production function, consumption function, investment function, demand and supply functions, portfolio theory etc.

Exercises

Attempt all of the following questions

1. Define econometrics

2. What is the difference between econometrics and
    i) Mathematical economics
    ii) Statistics

3. What is the difference between mathematical model and econometric model? Why do you think economists prefer econometric model to mathematical model to study the relationship between economic variables?

4. Explain the difference between theoretical and applied econometrics.

5. What types of criteria would you use to evaluate the results of an estimated relationship? Which of these criteria are more important? Why?

ARBAMINCH UNIVERSITY

FACULTY OF BUSINESS AND ECONOMICS

DEPARTMENT OF ECONOMICS

ASSIGNMENT FOR INTRODUCTION TO ECONOMICS

1. Explain the following concepts

a) Regression analysis and Correlation analysis

b) Population Regression Function and Sample Regression Function

c) Ordinary Least Squares Method

d) Coefficient of Determination

e) Autocorrelation

f) Dummy variables

2. Given the model $Y_i = \beta_0 + \beta_1 X_i + U_i$

a) Determine the distribution of the dependent variable(Y_i), given that all the assumption of the linear regression model are satisfied

b) State the assumptions you used to determine the distribution of Y_i

c) Econometricians are frequently heard saying that U_i in the above model is responsible for the omitted variables from the model. What do you think is/are the reason(s) that force econometricians to omit the variables from their model

3. Suppose the Marketing Department of firm A collected data from six different market centers on the quantity demanded of the commodity that the firm produces (say X) and its price, and the following data were collected.

Market center	I	II	II	IV	V	VI
Quantity sold(in tons)	8	3	4	7	8	6
Price(in Birr)	2	4	3	1	3	5

a) Estimate the demand function for commodity X, assuming linear relationship between price and quantity demanded. Interpret your results.
b) Are your estimates economically meaningful? Explain
c) Should the management accept the proposal of the marketing department? Show your steps clearly.
d) Test the hypothesis that the price of commodity X. does not significantly affect the demand for X.

4. Consider the consumption function for a certain locality to be

$$C_t = \beta_0 + \beta_1 Y_t + \beta_2 A_t + u_t$$

Where, C = consumption expenditure

Y= personal disposable income

A= liquid assets of consumers

$$E(u_t) = 0$$
$$Var(u_t) = \sigma_u^2 Y_t^2$$ (where, σ_u^2 is constant)

i. Transform the above model into one in which the disturbance term is homoscedastic

ii. Prove that the variance of the disturbance term of the transformed model is equal to σ_u^2 , and hence the transformed model is homoscedastic

5. Suppose that the following results were obtained from observations on 100 employees of XYZ Company.

$$\sum X_{1i} = 123$$

$$\sum X_{1i} Y_i = 1290$$

$$\sum Y_i = 460$$

$$\sum Y_i^2 = 539{,}500$$

$$\sum X_{2i} = 96$$

$\sum Y_i^2 = 3924$

$\sum X_{1i}^2 = 232$

$\sum X_{2i}^2 = 167$

$\sum X_{2i}Y_i = 615$

$\sum X_{2i}X_{1i} = 125$

$\sum X_{1i}Y_i = 870$

Given that the PRF is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

Where, Y, X_1 and X_2 denote hourly wage, education level in years of schooling and years of work experience, and u represents the error term that satisfies the usual assumptions of classical linear regression models. Using the information given above,

i. Form its matrix representation

ii. Using matrix approach, compute the OLS estimates of the coefficient of the model

iii. Compute the 90% confidence interval for the coefficients of X_1 and X_2

iv. Conduct the overall significance test using ANOVA table and *F* test