Bonga University

Undergraduate Program

Department of Agricultural Economics

Course title: **Econometrics** Course Code**:** AgEc

Program**: BSc in Agricultural Economics** ECTS credit **(CP)**: 5

Instructor**: Mr. Fami. A** Semester**: 2nd**

Course delivery system**: Parallel**

Course objectives

The primary objective of the course is to equip students with the basic tools that are useful to conduct researches in the areas of agriculture and natural resources. At the end of the course, students will be able to:

- explain main goals of Econometrics and its purpose;
- discuss and apply the methodology of Econometrics for researches;
- develop a mathematical model which will be used to explain the relationship between variables presented in the model;
- make predictions and forecasting using simple and multiple linear regression analysis;
- apply Econometric methods to the investigation of economic relationships and processes;
- make generalizations and policy implications based on relationships analysed using multiple regression models;

- get acquainted with some non-linear model and time series analysis,

Course content

Unit 1: What are Econometrics – What it is all about?

- ✓ Definition and Scope
- ✓ Goals of Econometrics
- ✓ Methodology of Econometrics
- ✓ Elements of Econometrics
- ✓ Types of Econometrics

Unit 2: Correlation Theory

- ✓ Basic concepts of Correlation
- ✓ Coefficient of Linear Correlation
- ✓ Types of Correlation Coefficient

Unit 3: Simple Linear Regression Models

- ✓ Basic Concepts and Assumptions
- ✓ Least Squares Criteria
- ✓ Normal Equations of OLS
- ✓ Estimation of Elasticities from regression equations
- ✓ Coefficient of Correlation and Determination
- ✓ Hypothesis Testing

Unit 4: Multiple Regression Analysis

- ✓ Model with two Explanatory Variables
- ✓ Notations and Assumptions
- ✓ Estimation of Partial regression coefficient
- ✓ Partial Correlation Coefficients
- ✓ Analysis of Variance
- ✓ Hypothesis Testing
- ✓ Other functional forms

Unit 5: Dummy Variable Regression Analysis

- ✓ Definitions of Dummy Variables
- ✓ ANOVA Models
- ✓ ANCOVA models

Unit 6: Econometric Problems

- ✓ Non-normality
- ✓ Multicollinearity
- ✓ Heteroskedasticity
- ✓ Autocorrelation

Unit 7: Non-linear Regression and Time Series Econometrics

- ✓ Non-linear regression models: Overview
- ✓ Time series Analysis

UNIT ONE: WHAT IS ECONOMETRICS?

Definition and scope of econometrics

Simply stated Econometric means economic measurement. The "metric" part of the word signifies measurement and econometrics is concerned with the measuring of economic relationships.

- It is a social science in which the tools of economic theory, mathematics and statistical inference are applied to the analysis of economic phenomena (A.S. Goldberger, 1964).
- In the words of Maddala econometrics is "the application of statistical and mathematical methods to the analysis of economic data, with a purpose of giving empirical content to economic theories and verifying them or refuting them." .
- It is a special type of economic analysis and research in which the general economic theory formulated in mathematical form (i.e. mathematical economics) is combined with empirical measurement (i.e. statistics) of economic phenomena.
- In short, econometrics may be considered as the integration of economics, mathematics, and statistics for the purpose of providing numerical values for the parameters of economic relationships and verifying economic theories.
- The field of knowledge which helps us to carry out such an evaluation of economic theories in empirical terms

The purpose of studying the relationships among economic variables are attempting to answer questions of

- ✓ If one variable changes in a certain magnitude, by how much will another variable change?

- ✓ If we know the value of one variable; can we forecast or predict the corresponding value of another?
- ✓ It also helps us to understand the real economic world we live in.

A distinction between econometrics and mathematical economics, econometrics and statistics, econometrics and economic model

Econometrics vs. mathematical economics

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

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

Example: Other things remaining constant (ceteris paribus) a reduction in the price of a commodity is expected to increase the quantity demanded. And Economic theory postulates an

inverse relationship between price and quantity demanded of a commodity. But the theory does not provide numerical value as the measure of the relationship between the two. Here comes the task of the econometrician to provide the numerical value by which the quantity will go up or down as a result of changes in the price of the commodity.

Econometrics vs. statistics

Econometrics differs from both mathematical statistics and economic statistics. An economic statistician gathers empirical data, records them, tabulates them or charts them, presenting economic data (descriptive statistics). and attempts to describe the pattern in their development over time and perhaps detect some relationship between various economic magnitudes. Economic statistics is mainly a descriptive aspect of economics. It does not provide explanations of the development of the various variables and it does not provide measurements the coefficients of economic relationships.

Mathematical (or inferential) statistics deals with the method of measurement which are developed on the basis of controlled experiments. But statistical methods of measurement are not appropriate for a number of economic relationships because for most economic relationships controlled or carefully planned experiments cannot be designed due to the fact that the nature of relationships among economic variables are stochastic or random. Yet the fundamental ideas of inferential statistics are applicable in econometrics, but they must be adapted to the problem economic life. Econometric methods are adjusted so that they may become appropriate for the measurement of economic relationships which are stochastic. The adjustment consists primarily in specifying the stochastic (random) elements that are supposed to operate in the real world and enter into the determination of the observed data. Mathematical economics do provide much of the tools used in Econometrics. But Econometrics needs special methods to deal with economic data which are never experimental data.

Examples: Errors of measurement, problem of multicollinearity, problem of serial correlation are only econometric problems and are not concerns of mathematical statistics.

Econometrics utilizes these data to estimate quantitative economic relationships and to test hypothesis about them. The Econometrician is called upon to develop special methods of analysis and deal with such kinds of Econometric problems.

Economic models vs. econometric models

i) Economic models:

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

1. A set of variables

2. A list of fundamental relationships and

3. A number of strategic coefficients

ii) Econometric models:

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

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

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

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

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

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

Methodology of Econometrics

Econometric research is concerned with the measurement of the parameters of economic relationships and with the predication of the values of economic variables. The relationships of economic theory which can be measured with econometric techniques are relationships in which some variables are postulated as causes of the variation of other variables. Starting with the postulated theoretical relationships among economic variables, econometric research or inquiry generally proceeds along the following lines/stages.

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

1. Specification of the model

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

- Determine dependent and independent (explanatory) variables to be included in the model,
- Determine a priori theoretical expectations about the size and sign of the parameters of the function, and
- Determine mathematical form of the model (number of equations, specific form of the equations, etc.

Specification of the econometric model will be based on economic theory and on any available information related to the phenomena under investigation. Thus, specification of the econometric model presupposes knowledge of economic theory and familiarity with the particular phenomenon being studied. Specification of the model is the most important and the most difficult stage of any econometric research. It is often the weakest point of most econometric applications. In this stage there exists enormous degree of likelihood of committing errors or incorrectly specifying the model. The most common errors of specification are:

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

Such misspecification errors may associate to one or more reasons. Some of the common reasons for incorrect specification of the econometric models are:

- imperfections, looseness of statements in economic theories
- limited knowledge of the factors which are operative in any particular case
- formidable obstacles presented by data requirements in the estimation of large models

2. Estimation of the model

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

i. Gathering of the data on the variables included in the model.

ii. Examination of the identification conditions of the function (especially for simultaneous equations models).

iii. Examination of the aggregations problems involved in the variables of the function.

iv. Examination of the degree of correlation between the explanatory variables (i.e. examination of the problem of multicollinearity).

v. Choice of appropriate economic techniques for estimation, i.e. to decide a specific econometric method to be applied in estimation; such as, OLS, MLM

3. Evaluation of the estimates

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

i. Economic a priori criteria: These criteria are determined by economic theory and refer to the size and sign of the parameters of economic relationships.

ii. Statistical criteria (first-order tests): These are determined by statistical theory and aim at the evaluation of the statistical reliability of the estimates of the parameters of the model. Correlation coefficient test, standard error test, t-test, F-test, and R^2-test are some of the most commonly used statistical tests.

iii. Econometric criteria (second-order tests): These are set by the theory of econometrics and aim at the investigation of whether the assumptions of the

econometric method employed are satisfied or not in any particular case. The econometric criteria serve as a second order test (as test of the statistical tests) i.e. they determine the reliability of the statistical criteria; they help us establish whether the estimates have the desirable properties of unbiasedness, consistency, etc. Econometric criteria aim at the detection of the violation or validity of the assumptions of the various econometric techniques.

4) Evaluation of the forecasting power of the model

Forecasting is one of the aims of econometric research. However, before using an estimated model for forecasting by some way or another, the predictive power and other requirements of the model need to be checked. It is possible that the model may be economically meaningful and statistically and econometrically correct for the sample period for which the model has been estimated. Yet it may not be suitable for forecasting due to various factors (reasons). Therefore, this stage involves the investigation of the stability of the estimates and their sensitivity to changes in the size of the sample. Consequently, we must establish whether the estimated function performs adequately outside the sample data which require model performance test under extra sample.

1.1.Goals of Econometrics

Basically, there are three main goals of Econometrics. They are:

i) Analysis i.e. testing economic theory

ii) Policy making i.e. obtaining numerical estimates of the coefficients of economic relationships for policy simulations.

iii) Forecasting i.e. using the numerical estimates of the coefficients in order to forecast the future values of economic magnitudes.

Elements of Econometrics

Data

Collecting and coding the sample data, the raw material of econometrics. Most economic data is *observational*, or *non-experimental*, data (as distinct from *experimental* data generated under controlled experimental conditions).

⇒ Cross-section data

- ✓ many units observed at one point in time
- ✓ Generally obtained through official records of individual units, surveys, questionnaires (data collection instrument that contains a series of questions designed for a specific purpose)

⇒ Time-series data

- ✓ Same unit observed at many points in time (usually equally spaced)
- ✓ Chronological ordering
- ✓ Frequency of time series data: hour, day, week, month, year
- ✓ Time length between observations is generally equal

⇒ Pooled data

- ✓ Observations both over time and across units, but not necessarily the same units in each time period
- ✓ consists of cross-sectional data sets that are observed in different time periods and combined together
- ✓ At each time period (e.g., year) a different random sample is chosen from population
- ✓ Individual units are not the same
- ✓ For example, if we choose a random sample 400 firms in 2002 and choose another sample in 2010 and combine these cross-sectional data sets we obtain a pooled cross-section data set.
- ✓ Cross-sectional observations are pooled together over time.

⇒ Panel (longitudinal) data

- ✓ Special case of pooled data

- ✓ Multiple (same) units observed at multiple points in time
- ✓ Consists of a time series for each cross-sectional member in the data set.
- ✓ The same cross-sectional units (firms, households, etc.) are followed over time.
- ✓ For example: wage, education, and employment history for a set of individuals followed over a ten-year period.
- ✓ Another example: cross-country data set for a 20 year period containing life expectancy, income inequality, real GDP per capita and other country characteristics.

Specification

Specification of the *econometric model* that we think (hope) generated the sample data -- that is, specification of the data generating process (or DGP).

An *econometric model* consists of <u>two</u> components:

1. An *economic model*: specifies the *dependent* or *outcome* variable to be explained and the *independent* or *explanatory* variables that we think are related to the dependent variable of interest.
   - Often suggested or derived from economic theory.
   - Sometimes obtained from informal intuition and observation.

2. A *statistical model*: specifies the statistical elements of the relationship under investigation, in particular the *statistical properties* of the *random* variables in the relationship.

Estimation

Consists of using the assembled *sample data* on the *observable* variables in the model to compute *estimates* of the numerical values of all the unknown parameters in the model.

Eg. how big a change in one variable tends to be associated with a unit change in another?

Testing hypotheses

 Is the income elasticity of the demand for money equal to one?

Inference

Consists of using the parameter estimates computed from sample data to test hypotheses about the numerical values of the unknown *population* parameters that describe the behaviour of the population from which the sample was selected.

Eg. How many more students would Reed need to admit in order to fill its class if tuition were $1000 higher?

Desirable properties of an econometric model

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

1. Theoretical plausibility. The model should be compatible with the postulates of economic theory. It must describe adequately the economic phenomena to which it relates.

2. Explanatory ability. The model should be able to explain the observations of the actual world. It must be consistent with the observed behavior of the economic variables whose relationship it determines.

3. Accuracy of the estimates of the parameters. The estimates of the coefficients should be accurate in the sense that they should approximate as best as possible the true parameters of the structural model. The estimates should if possible, possess the desirable properties of unbiasedness, consistency and efficiency.

4. Forecasting ability. The model should produce satisfactory predictions of future values of the dependent (endogenous) variables.

5. Simplicity. The model should represent the economic relationships with maximum simplicity. The fewer the equations and the simpler their mathematical form, the better the model is considered, ceteris paribus (that is to say provided that the other desirable properties are not affected by the simplifications of the model).

Types of Econometrics

Econometrics may be divided into two broad categories: theoretical econometrics and applied econometrics.

Theoretical econometrics is concerned with the development of appropriate methods for measuring economic relationships specified by econometric models. In this aspect, econometrics leans heavily on mathematical statistics. For example, one of the methods used extensively in this book is least squares. Theoretical econometrics must spell out the assumptions of this method, its properties, and what happens to these properties when one or more of the assumptions of the method are not fulfilled.

In applied econometrics we use the tools of theoretical econometrics to study some special field(s) of economics and business, such as the production function, investment function, demand and supply functions, portfolio theory, etc.

UNIT TWO: CORRELATION THEORY

2.1. Basic Concepts of correlation

Economic variables have a great tendency of moving together and very often data are given in pairs of observations in which there is a possibility that the change in one variable is on average accompanied by the change of the other variable. Correlation may be defined as the degree of relationship existing between two or more variables. Correlation Analysis is a statistical technique used to indicate the nature and degree of relationship existing between one variable and the other(s).

Types of Correlation

- *Positive and negative correlation*. The correlation is said to be positive correlation if the values of two variables changing with same direction. Ex. Pub. Exp. & sales, Height & weight. The correlation is said to be negative correlation when the values of variables change with opposite direction. Ex. Price & qty. demanded. For example, the correlation between price of a commodity and its quantity supplied is positive since as price rises, quantity supplied will be increased and vice versa. Correlation is said to negative if an increase or a decrease in one variable is accompanied by a decrease or an increase in the other in which both are changed with opposite direction. For example, the correlation between price of a commodity and its quantity demanded is negative since as price rises, quantity demanded will be decreased and vice versa.

- *Simple and multiple correlations*. The degree of relationship existing between two variables is called simple correlation. The degree of relationship connecting three or more variables is called multiple correlations. A correlation is also said to be partial if it studies the degree of relationship between two variables keeping all other variables connected with these two are constant.

- *Linear and non-linear correlation*. Correlation may be linear, when all points (X, Y) on scatter diagram seem to cluster near a straight, and nonlinear, when all points seem to lie near a curve. In other words, correlation is said to be linear if the change in one variable brings a constant change of the other. It may be non-linear if the change in one variable brings a different change in the other.

2.2. Methods of Measuring Correlation

In correlation analysis there are two important things to be addressed. These are the type of co-variation existed between variables and its strength. And the types of correlation mentioned before do not show to us the strength of co-variation between variables. correlation can be measuring by:

A, The Scattered Diagram or Graphic Method

B, The Simple Linear Correlation coefficient

C, The coefficient of Rank Correlation

D, partial correlation coefficient

The Scattered Diagram or Graphic Method

The scatter diagram is a rectangular diagram which can help us in visualizing the relationship between two phenomena. It puts the data into X-Y plane by moving from the lowest data set to the highest data set. It is a non-mathematical method of measuring the degree of co-variation between two variables. Scatter plots usually consist of a large body of data. The closer the data points come together and make a straight line, the higher the correlation between the two variables, or the stronger the relationship.

If the data points make a straight line going from the origin out to high x- and y-values, then the variables are said to have a positive correlation. If the line goes from a high-value on the y-axis down to a high-value on the x-axis, the variables have a negative correlation.

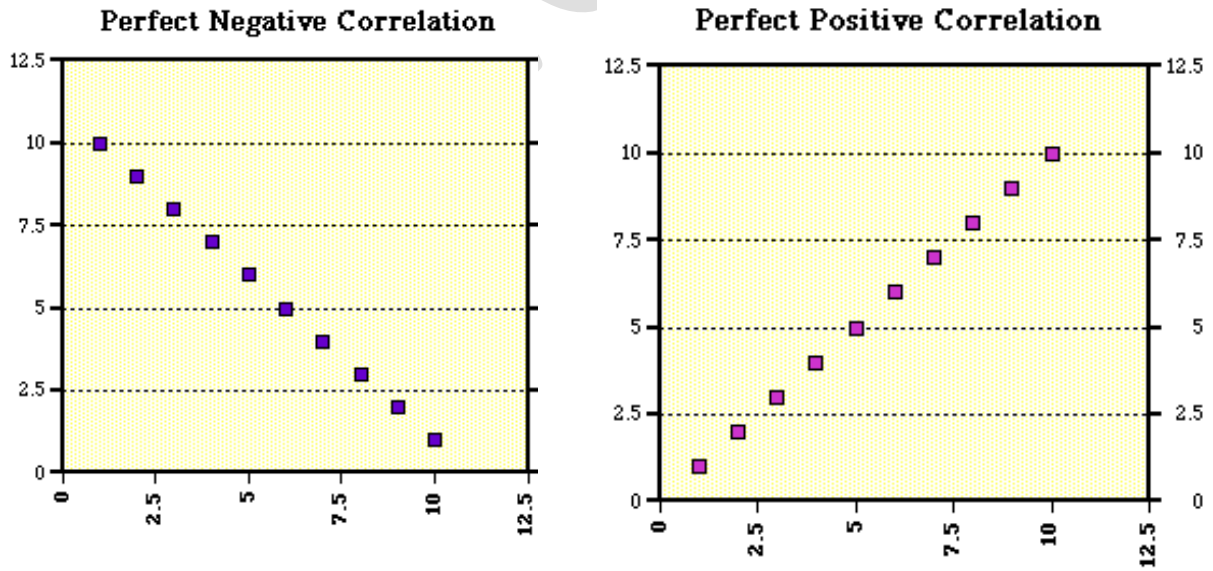


Figure 1: Perfect linear correlations

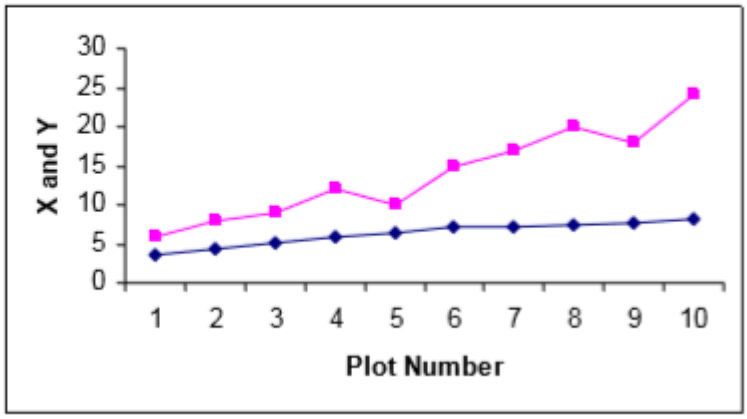
A perfect positive correlation is given the value of 1. A perfect negative correlation is given the value of -1. If there is absolutely no correlation present the value given is 0. The closer the number is to 1 or -1, the stronger the correlation, or the stronger the relationship between the variables. The closer the number is to 0, the weaker the correlation.

Two variables may have a positive correlation, negative correlation, or they may be uncorrelated. This holds true both for linear and nonlinear correlation. Two variables are said to be positively correlated if they tend to change together in the same direction, that is, if they tend to increase or decrease together. Such positive correlation is postulated by economic theory for the quantity of a commodity supplied and its price. When the price increases the quantity supplied increases. Conversely, when price falls the quantity supplied decreases. Negative correlation: Two variables are said to be negatively correlated if they tend to change in the opposite direction: when X increases Y decreases, and vice versa. For example, saving and household size are negatively correlated. When price increases, demand for the commodity decreases and when price falls demand increases.

Example 2. Find out graphically, if there is any correlation between price yield per plot (qtls); denoted by Y and quantity of fertilizer used (kg); denote by X.

Plot No.:	1	2	3	4	5	6	7	8	9	10
Y:	3.5	4.3	5.2	5.8	6.4	7.3	7.2	7.5	7.8	8.3
X:	6	8	9	12	10	15	17	20	18	24

Solution: The Correlogram of the given data is shown in Figure 4-3



Shows that the two curves move in the same direction and, moreover, they are very close to each other, suggesting a close relationship between price yield per plot (qtls) and quantity of fertilizer used (kg)

Remark: Both the Graphic methods - scatter diagram and correlation graph provide a 'feel for' of the data – by providing visual representation of the association between the variables. These are readily comprehensible and enable us to form a fairly good, though rough idea of the nature and degree of the relationship between the two variables. However, these methods are unable to quantify the relationship between them. To quantify the extent of correlation, we make use of algebraic methods - which calculate correlation coefficient.

The Population linear Correlation Coefficient 'ρ' and its Sample Estimate 'r'

(Karl Pearson's Coefficient of Correlation)

For a precise quantitative measurement of the degree of correlation between Y and X we use a parameter which is called the correlation coefficient and is usually designated by the Greek letter ρ. Having as subscripts the variables whose correlation it measures, ρ refers to the correlation of all the values of the population of X and Y. Its estimate from any particular sample (the sample statistic for correlation) is denoted by r with the relevant subscripts. For example, if we measure the correlation between X and Y the population correlation coefficient

is represented by ρ_{xy} and its sample estimate by r_{xy}. The simple correlation coefficient is used to measure relationships which are simple and linear only. It cannot help us in measuring non-linear as well as multiple correlations.

Karl Pearson's measure, known as Pearsonian correlation coefficient between two variables X and Y, usually denoted by r(X,Y) or rxy or simply r is a numerical measure of linear relationship between them and is defined as the ratio of the covariance between X and Y, to the product of the standard deviations of X and Y.

Symbolically

$$r_{xy} = \frac{Cov(X,Y)}{S_x . S_y} \qquad \text{.....(4.1)}$$

when, $(X_1, Y_1); (X_2, Y_2); \ldots\ldots\ldots\ldots (X_n, Y_n)$ are N pairs of observations of the variables X and Y in a bivariate distribution,

$$Cov(X,Y) = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{N} \qquad \text{.....(4.2a)}$$

$$S_x = \sqrt{\frac{\sum (X - \overline{X})^2}{N}} \qquad \text{.....(4.2b)}$$

and $$S_y = \sqrt{\frac{\sum (Y - \overline{Y})^2}{N}} \qquad \text{.....(4.2c)}$$

Thus by substituting *Eqs. (4.2)* in *Eq. (4.1)*, we can write the Pearsonian correlation coefficient as

$$r_{xy} = \frac{\frac{1}{N}\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\frac{1}{N}\sum (X - \overline{X})^2}\sqrt{\frac{1}{N}\sum (Y - \overline{Y})^2}}$$

$$r_{xy} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2}\sqrt{\sum (Y - \overline{Y})^2}} \qquad \text{.....(4.3)}$$

If we denote, $d_x = X - \overline{X}$ and $d_y = Y - \overline{Y}$

Then $$r_{xy} = \frac{\sum d_x d_y}{\sqrt{\sum d_x^{\,2} d_y^{\,2}}} \qquad \text{.....(4.3a)}$$

We can further simply the calculations of *Eqs. (4.2)*

We have

$$Cov(X,Y) = \frac{1}{N}\sum (X-\overline{X})(Y-\overline{Y})$$

$$= \frac{1}{N}\sum XY - \overline{X}\overline{Y}$$

$$= \frac{1}{N}\sum XY - \frac{\sum X}{N}\frac{\sum Y}{N}$$

$$= \frac{1}{N^2}\left[N\sum XY - \sum X \sum Y\right] \qquad \text{.....(4.4)}$$

and $$S_x^2 = \frac{1}{N}\sum (X-\overline{X})^2$$

$$= \frac{1}{N}\sum X^2 - (\overline{X})^2$$

$$= \frac{1}{N}\sum X^2 - \left(\frac{\sum X}{N}\right)^2$$

$$= \frac{1}{N^2}\left[N\sum X^2 - \left(\sum X\right)^2\right] \qquad \text{.....(4.5a)}$$

Similarly, we have

$$S_y^2 = \frac{1}{N^2}\left[N\sum Y^2 - \left(\sum Y\right)^2\right] \qquad \text{.....(4.5b)}$$

So Pearsonian correlation coefficient may be found as

$$r_{xy} = \frac{\frac{1}{N^2}\left[N\sum XY - \sum X\sum Y\right]}{\sqrt{\frac{1}{N^2}\left[N\sum X^2 - \left(\sum X\right)^2\right]}\sqrt{\frac{1}{N^2}\left[N\sum Y^2 - \left(\sum Y\right)^2\right]}}$$

or $$r_{xy} = \frac{N\sum XY - \sum X\sum Y}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}} \qquad \text{.....(4.6)}$$

Simple correlation coefficient is defined by the formula

$$r = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{n\sum X_i^2 - (\sum X_i)^2}\sqrt{n\sum Y_i^2 - (\sum Y_i)^2}}$$

Or

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

$x_i = X_i - \bar{X}$, $y_i = Y_i - \bar{Y}$

Example 2.1: The following table shows the quantity supplied for a commodity with the corresponding price values. Determine the type of correlation that exists between these two variables.

Table 1: Data for computation of correlation coefficient

Time period(in days)	Quantity supplied Y_i (in tons)	Price X_i (in shillings)
1	10	2
2	20	4
3	50	6
4	40	8
5	50	10
6	60	12
7	80	14
8	90	16
9	90	18

10	120	20

To estimate the correlation coefficient we, compute the following results.

Table 2: Computations of inputs for correlation coefficients

Y	X	$x_i = X_i - \bar{X}$	$y_i = Y_i - \bar{Y}$	x^2	y^2	$x_i y_i$	XY	X^2	Y^2
10	2	-9	-51	81	2601	459	20	4	100
20	4	-7	-41	49	1681	287	80	16	400
50	6	-5	-11	25	121	55	300	36	2500
40	8	-3	-21	9	441	63	320	64	1600
50	10	-1	-11	1	121	11	500	100	2500
60	12	1	-1	1	1	-1	720	144	3600
80	14	3	19	9	361	57	1120	196	6400
90	16	5	29	25	841	145	1440	256	8100
90	18	7	29	49	841	203	1620	324	8100
120	20	9	59	81	3481	531	2400	400	14400
Sum=610	110	0	0	330	10490	1810	8520	1540	47700
Mean=61	11								

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}} = \frac{10(8520) - (110)(610)}{\sqrt{10(1540) - (110)(110)}\sqrt{10(47700) - (610)(610)}} = 0.975$$

Or using the deviation form (Equation 2.2), the correlation coefficient can be computed as:

$$r = \frac{1810}{\sqrt{330}\sqrt{10490}} = 0.975$$

This result shows that there is a strong positive correlation between the quantity supplied and the price of the commodity under consideration.

Example 2, Find the Pearsonian correlation coefficient between sales (in thousand units) and expenses (in thousand rupees) of the following 10 firms:

Firm:	1	2	3	4	5	6	7	8	9	10
Sales:	50	50	55	60	65	65	65	60	60	50
Expenses:	11	13	14	16	16	15	15	14	13	13

Solution, Let sales of a firm be denoted by X and expenses be denoted by Y

Firm	X	Y	$d_x = X - \bar{X}$	$d_y = Y - \bar{Y}$	d_x^2	d_y^2	$d_x d_y$
1	*50*	*11*	*-8*	*-3*	*64*	*9*	*24*
2	*50*	*13*	*-8*	*-1*	*64*	*1*	*8*
3	*55*	*14*	*-3*	*0*	*9*	*0*	*0*
4	*60*	*16*	*2*	*2*	*4*	*4*	*4*
5	*65*	*16*	*7*	*2*	*49*	*4*	*14*
6	*65*	*15*	*7*	*1*	*49*	*1*	*7*
7	*65*	*15*	*7*	*1*	*49*	*1*	*7*
8	*60*	*14*	*2*	*0*	*4*	*0*	*0*
9	*60*	*13*	*2*	*-1*	*4*	*1*	*-2*
10	*50*	*13*	*-8*	*-1*	*64*	*1*	*8*
	$\sum X$	$\sum Y$			$\sum d_x^2$	$\sum d_y^2$	$\sum d_x d_y$
	= 580	*= 140*			*=360*	*=22*	*=70*

$$r_{xy} = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 d_y^2}}$$
$$r_{xy} = \frac{70}{\sqrt{360x22}}$$
$$r_{xy} = \frac{70}{\sqrt{7920}}$$
$$r_{xy} = 0.78$$

The value of **rxy = 0.78,** indicate a high degree of positive correlation between sales and expenses.

Properties of Simple Correlation Coefficient

The simple correlation coefficient has the following important properties:

1. *The value of correlation coefficient always ranges between -1 and +1.*

- ✓ **Remarks:** (i) This property provides us a check on our calculations. If in any problem, the obtained value of r lies outside the limits + 1, this implies that there is some mistake in our calculations.
- ✓ (ii) The sign of r indicate the nature of the correlation. Positive value of r indicates positive correlation, whereas negative value indicates negative correlation. r = 0 indicate absence of correlation.
- ✓ (iii) The following table sums up the degrees of correlation corresponding to various values of r:

Value of *r*	Degree of correlation
±1	perfect correlation
±0.90 or more	very high degree of correlation
±0.75 to ±0.90	sufficiently high degree of correlation
±0.60 to ±0.75	moderate degree of correlation
±0.30 to ±0.60	only the possibility of a correlation
less than ±0.30	*possibly no correlation*
0	absence of correlation

2. *The correlation coefficient is symmetric.* That means $r_{xy} = r_{yx}$, where, r_{xy} is the correlation coefficient of X on Y and r_{yx} is the correlation coefficient of Y on X.

3. *The correlation coefficient is independent of change of origin and change of scale.* By change of origin we mean subtracting or adding a constant from or to every values of a variable and change of scale we mean multiplying or dividing every value of a variable by a constant.

4. *If X and Y variables are independent, the correlation coefficient is zero. But the converse is not true. In other words, two independent variables are uncorrelated but the converse is not true*

If X and Y are independent variables then

rxy = 0

However, the converse of the theorem is not true i.e., uncorrelated variables need not necessarily be independent. As an illustration consider the following bivariate distribution.

X	:	1	2	3	-3	-2	-1
Y	:	1	4	9	9	4	1

For this distribution, value of r will be 0.

Hence in the above example the variable X and Y are uncorrelated. But if we examine the data carefully we find that X and Y are not independent but are connected by the relation Y = X2. The above example illustrates that uncorrelated variables need not be independent.

Remarks: One should not be confused with the words uncorrelation and independence. rxy = 0 i.e., uncorrelation between the variables X and Y simply implies the absence of any linear (straight line) relationship between them. They may, however, be related in some other form other than straight line e.g., quadratic (as we have seen in the above example), logarithmic or trigonometric form.

5. The correlation coefficient has the same sign with that of regression coefficients.

6. The correlation coefficient is the geometric mean of two regression coefficients.

$$r = \sqrt{b_{yx} * b_{xy}}$$

7. The square of Pearsonian correlation coefficient is known as the coefficient of determination.

Coefficient of determination, which measures the percentage variation in the dependent variable that is accounted for by the independent variable, is a much better and useful measure for interpreting the value of r.

Though, correlation coefficient is most popular in applied statistics and econometrics, it has its own limitations.

The major limitations of the method are:

1. The correlation coefficient always assumes linear relationship regardless of the fact whether the assumption is true or not.

2. Great care must be exercised in interpreting the value of this coefficient as very often the coefficient is misinterpreted. For example, high correlation between lung cancer and smoking does not show us smoking causes lung cancer.

3. The value of the coefficient is unduly affected by the extreme values

4. The coefficient requires the quantitative measurement of both variables. If one of the two variables is not quantitatively measured, the coefficient cannot be computed.

The Rank Correlation Coefficient

The formulae of the linear correlation coefficient developed in the previous section are based on the assumption that the variables involved are quantitative and that we have accurate data for their measurement. However, in many cases the variables may be qualitative (or binary variables) and hence cannot be measured numerically. For example, profession, education, preferences for particular brands, are such categorical variables. Furthermore, in many cases precise values of the variables may not be available, so that it is impossible to calculate the value of the correlation coefficient with the formulae developed in the preceding section. For such cases it is possible to use another statistic, the rank correlation coefficient (or spearman's correlation coefficient.). We rank the observations in a specific sequence for example in order of size, importance, etc., using the numbers 1, 2, 3… n. Hence, the name of the statistic is given as rank correlation coefficient. If two variables X and Y are ranked in such way that the values are ranked in ascending or descending order, the rank correlation coefficient may be computed by the formula

$$r' = 1 - \frac{6\sum D^2}{n(n^2 - 1)} \qquad 2.3$$

Where,

D = difference between ranks of corresponding pairs of X and Y

n = number of observations.

The values that r may assume range from + 1 to – 1.

Two points are of interest when applying the rank correlation coefficient. Firstly, it does not matter whether we rank the observations in ascending or descending order. However, we must use the same rule of ranking for both variables. Second if two (or more) observations have the same value we assign to them the mean rank.

Example 2.2: A market researcher asks experts to express their preference for twelve different brands of soap. Their replies are shown in the following table.

Example for rank correlation coefficient

Brands of soap	A	B	C	D	E	F	G	H	I	J	K	L
Person I	9	10	4	1	8	11	3	2	5	7	12	6
Person II	7	8	3	1	10	12	2	6	5	4	11	9

The figures in this table are ranks but not quantities. We have to use the rank correlation coefficient to determine the type of association between the preferences of the two persons. This can be done as follows.

Computation for rank correlation coefficient

Brands of soap	A	B	C	D	E	F	G	H	I	J	K	L	Total
Person I	9	10	4	1	8	11	3	2	5	7	12	6	
Person II	7	8	3	1	10	12	2	6	5	4	11	9	
D_i	2	2	1	0	-2	-1	1	-4	0	3	1	-3	
D_i^2	4	4	1	0	4	1	1	16	0	9	1	9	50

The rank correlation coefficient (using Equation 2.3)

$$r' = 1 - \frac{6\sum D^2}{n(n^2-1)} = 1 - \frac{6(50)}{12(12^2-1)} = 0.827$$

This figure, 0.827, shows a marked similarity of preferences of the two persons for the various brands of soap.

In this case common ranks are assigned to the repeated items. These common ranks are the arithmetic mean of the ranks, which these items would have got if they were different from each other and the next item will get the rank next to the rank used in computing the common rank. For example, suppose an item is repeated at rank 4. Then the common rank to be assigned to each item is (4+5)/2, i.e., 4.5 which is the average of 4 and 5, the ranks which these observations would have assumed if they were different. The next item will be assigned the rank 6. If an item is repeated thrice at rank 7, then the common rank to be assigned to each value will be (7+8+9)/3, i.e., 8 which is the arithmetic mean of 7,8 and 9 viz., the ranks these observations would have got if they were different from each other. The next rank to be assigned will be 10.

If only a small proportion of the ranks are tied, this technique may be applied together with rank correlation formula. If a large proportion of ranks are tied, it is advisable to apply an adjustment or a correction factor that is explained as below:

Adding this equation

$$\frac{m(m^2-1)}{12}$$

to $\sum d^2$ *; where m is the number of times an item is repeated. This correction factor is to be added for each repeated value in both the series".*

Example, For a certain joint stock company, the prices of preference shares (X) and debentures (Y) are given below

X:	73.2	85.8	78.9	75.8	77.2	81.2	83.8
Y:	97.8	99.2	98.8	98.3	98.3	96.7	97.1

Use the method of rank correlation to determine the relationship between preference prices and debentures prices.

Solution: Calculations for Coefficient of Rank Correlation

X	Y	Rank of X (X_R)	Rank of Y (Y_R)	$d = X_R - Y_R$	d^2
73.2	97.8	7	5	2	4
85.8	99.2	1	1	0	0
78.9	98.8	4	2	2	4
75.8	98.3	6	3.5	2.5	6.25
77.2	98.3	5	3.5	1.5	2.25
81.2	96.7	3	7	-4	16
83.8	97.1	2	6	-4	16
				$\sum d = 0$	$\sum d^2 = 48.50$

In this case, due to repeated values of Y, we have to apply ranking as average of 2 ranks, which could have been allotted, if they were different values. Thus ranks 3 and 4 have been allotted as 3.5 to both the values of Y = 98.3. Now we also have to apply correction factor

$\frac{m(m^2-1)}{12}$ to $\sum d^2$, where m in the number of times the value is repeated, here $m = 2$.

$$\rho = \frac{6\left[\sum d^2 + \frac{m(m^2-1)}{2}\right]}{N(N^2-1)}$$

$$= \frac{6\left[48.5 + \frac{2(4-1)}{12}\right]}{7(7^2-1)}$$

$$= 1 - \frac{6 \times 49}{7 \times 48}$$

$$= 0.125$$

Hence, there is a very low degree of positive correlation, probably no correlation, between preference share prices and debenture prices.

Partial Correlation Coefficients

A partial correlation coefficient measures the relationship between any two variables, when all other variables connected with those two are kept constant. For example, let us assume that we want to measure the correlation between the number of hot drinks (X_1) consumed in a summer

resort and the number of tourists (X_2) coming to that resort. It is obvious that both these variables are strongly influenced by weather conditions, which we may designate by X_3. On a priori grounds we expect X_1 and X_2 to be positively correlated: when a large number of tourists arrive in the summer resort, one should expect a high consumption of hot drinks and vice versa. The computation of the simple correlation coefficient between X_1 and X_2 may not reveal the true relationship connecting these two variables, however, because of the influence of the third variable, weather conditions (X_3). In other words, the above positive relationship between number of tourists and number of hot drinks consumed is expected to hold if weather conditions can be assumed constant. If weather condition changes, the relationship between X_1 and X_2 may change to such an extent as to appear even negative. Thus, if the weather is hot, the number of tourists will be large, but because of the heat they will prefer to consume more cold drinks and ice-cream rather than hot drinks. If we overlook the weather and look only at X_1 and X_2 we will observe a negative correlation between these two variables which is explained by the fact that hot drinks as well as number of visitors are affected by heat. In order to measure the true correlation between X_1 and X_2, we must find some way of accounting for changes in X_3. This is achieved with the partial correlation coefficient between X_1 and X_2, when X_3 is kept constant. The partial correlation coefficient is determined in terms of the simple correlation coefficients among the various variables involved in a multiple relationship. In our example there are three simple correlation coefficients

r_{12} = correlation coefficient between X_1 and X_2

r_{13} = correlation coefficient between X_1 and X_3

r_{23} = correlation coefficient between X_2 and X_3

The partial correlation coefficient between X_1 and X_2, keeping the effect of X_3 constant is given by:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^{\ 2})(1-r_{23}^{\ 2})}} \qquad 2.4$$

Similarly, the partial correlation between X_1 and X_3, keeping the effect of X_2 constant is given by:

$$r_{13.2} = \frac{r_{13} - r_{12} * r_{23}}{\sqrt{(1 - r_{12}{}^2)(1 - r_{23}{}^2)}} \quad \text{and} \quad r_{23.1} = \frac{r_{23} - r_{12} * r_{13}}{\sqrt{(1 - r_{12}^2)(1 - r_{13}^2)}}$$

Example 2.3: The following table gives data on the yield of corn per acre(Y), the amount of fertilizer used(X_1) and the amount of insecticide used (X_2). Compute the partial correlation coefficient between the yield of corn and the fertilizer used keeping the effect of insecticide constant.

Table 3: Data on yield of corn, fertilizer and insecticides used

Year	**1971**	**1972**	**1973**	**1974**	**1975**	**1976**	**1977**	**1978**	**1979**	**1980**
Y	40	44	46	48	52	58	60	68	74	80
X_1	6	10	12	14	16	18	22	24	26	32
X_2	4	4	5	7	9	12	14	20	21	24

The computations are done as follows:

Table 4: Computation for partial correlation coefficients

Year	Y	X_1	X_2	Y	x_1	x_2	x_1y	x_2y	x_1x_2	x_1^2	x_2^2	y^2
1971	40	6	4	-17	-12	-8	204	136	96	144	64	289
1972	44	10	4	-13	-8	-8	104	104	64	64	64	169
1973	46	12	5	-11	-6	-7	66	77	42	36	49	121

1974	48	14	7	-9	-4	-5	36	45	20	16	25	81
1975	52	16	9	-5	-2	-3	10	15	6	4	9	25
1976	58	18	12	1	0	0	0	0	0	0	0	1
1977	60	22	14	3	4	2	12	6	8	16	4	9
1978	68	24	20	11	6	8	66	88	48	36	64	121
1979	74	26	21	17	8	9	136	153	72	64	81	289
1980	80	32	24	23	14	12	322	276	168	196	144	529
Sum	570	180	120	0	0	0	956	900	524	576	504	1634
Mean	57	18	12									

r_{yx1}=0.9854

r_{yx2}=0.9917

r_{x1x2}=0.9725

Then,

$$r_{yx_1.x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{(1-r_{yx_2}{}^2)(1-r_{x_1x_2}{}^2)}} = \frac{0.9854-(0.9917)(0.9725)}{\sqrt{(1-0.9917^2)(1-0.9725^2)}} = 0.7023$$

UNIT THREE: SIMPLE LINEAR REGRESSION

1.1.Concepts of linear regression

Economic theories are mainly concerned with the relationships among various economic variables. These relationships, when phrased in mathematical terms, can predict the effect of one variable on another. The functional relationships of these variables define the dependence of one variable upon the other variable (s) in the specific form. Regression analysis refers to estimating functions showing the relationship between two or more variables and corresponding tests. In other words, in regression analysis, we try to estimate or predict the average value of one variable on the basis of the fixed values of other variables.

1.2. Simple linear regression model

Simple regression includes estimating a simple linear function between two variables.

In the simple regression model, the population regression model or, simply, the population model is the following:

$$y = \beta_1 + \beta_2 x + u$$

We are going to consider that there are three types of variables in the model: y, x and u. In this model there is only one factor x to explain y. All the other factors that affect y are jointly captured by u.

In the literature the terms dependent variable and explanatory variable are described variously. A representative list is:

Dependent variable	Explanatory variable
⇳	⇳
Explained variable	Independent variable
⇳	⇳
Predictand	Predictor
⇳	⇳
Regressand	**Regressor**
⇳	⇳
Response	Stimulus
⇳	⇳
Endogenous	Exogenous
⇳	⇳
Outcome	Covariate
⇳	⇳
Controlled variable	Control variable

The variable u represents factors other than x that affect y. It is denominated error or random disturbance. The disturbance term can also capture measurement error in the dependent variable. The disturbance is an unobservable variable.

The parameters α and β are fixed and unknown.

On the right hand of we can distinguish two parts: the systematic component $\beta_1 + \beta_2 x$ and the random disturbance u.

Let's illustrate the distinction between stochastic (systematic component) and non-stochastic (disturbance component) relationships with the help of a supply function.

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

$$Q = f(P) = \alpha + \beta P$$

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

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

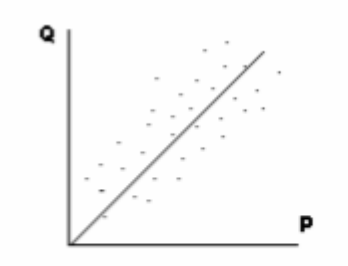


Fig 2.a The scatter diagram

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

a. Omission of variables from the function

b. Random behavior of human beings

c. Imperfect specification of the mathematical form of the model

d. Error of aggregation

e. Error of measurement

In order to take into account the above sources of errors we introduce in econometric functions a random variable which is usually denoted by the letter 'u' or 'ε ' and is called error term or random disturbance or stochastic term of the function, so called be cause u is supposed to 'disturb' the exact linear relationship which is assumed to exist between X and Y.

By introducing this random variable in the function, the model is rendered stochastic of the form:

$$Y_i = \alpha + \beta X + u_i .$$

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

2.3. Assumption of classical linear regression model

For the validity of a regression function and its attributes the data we use or the terms related to our regression function should fulfill the following conditions known as classical assumptions.

1. The mean value of error term 'U_i' is zero. This implies the sum of the individual disturbance terms is zero. The deviations of the values of some of the disturbance terms are negative; some are zero and some are positive and the sum or the average is zero. This is given by the following identities.

   $E(U_i) = \frac{\sum U_i}{n} = 0$. Multiplying both sides by (sample size 'n') we obtain the following.

   $\Rightarrow E(U_i) = \sum U_i = 0$.

   The same argument is true for sample regression function and so for residual terms given as follows $\sum e_i = 0$

   2. The disturbance terms have constant variance in each period. This is given as follows: $Var(U_i) = E((U_i - E(U_i))^2 = \delta^2 = \delta_u^2$. This assumption is known as the assumption of homoscedasticity**.** If this condition is not fulfilled or if the variance of the error terms varies as sample size changes or as the value of explanatory variables changes, then this leads to Heteroscedasticity problem.

2. The error terms 'U_i' are randomly distributed or the disturbance terms are not correlated. This means that there is no systematic variation or relation among the value of the error terms (U_i and U_j); Where i = 1, 2, 3, …….., j = 1, 2, 3, ……. and $i \neq j$.

Cov (U_i , U_j) = 0 for $i \neq j$. Note that the same argument holds for residual terms when we use sample data or sample regression function. Thus, Cov (e_i, e_j) = 0 for $i \neq j$. Otherwise, the error terms do not serve an adjustment purpose rather it causes an autocorrelation problem.

Algebraically,

$$Cov(u_i u_j) = \mathrm{E}\left[[(u_i - \mathrm{E}(u_i)][u_j - \mathrm{E}(u_j)]\right]$$

$$= E(u_i u_j) = 0 \ldots\ldots\ldots\ldots\ldots$$

3. The explanatory variable X_i is fixed in repeated samples. Each value of X_i does not vary for instance owing to change in sample size. This means the explanatory variables are non-random and hence distributional free variable.

4. Explanatory variables 'X_i' and disturbance terms 'U_i' are uncorrelated or independent. All the co-variances of the successive values of the error term are equal to zero. This condition is given by $Cov\left(U_i, X_i\right) = 0$. It is followed from this that the following identity holds true; $\sum e_i X_i = 0$. $\mathrm{cov}(XU) = \mathrm{E}\left[[(X_i - \mathrm{E}(X_i)][U_i - \mathrm{E}(U_i)]\right]$

$= \mathrm{E}[(X_i - \mathrm{E}(X_i)(U_i)]$ *given* $E(U_i) = 0$

$= \mathrm{E}(X_i U_i) - \mathrm{E}(X_i)\mathrm{E}(U_i)$

$= \mathrm{E}(X_i U_i)$

$= X_i \mathrm{E}(U_i)$, given that the x_i are fixed

$= 0$

5. Linearity of the model in parameters. The simple linear regression requires linearity in parameters; but not necessarily linearity in variables. The same technique can be applied to estimate regression functions of the following forms: Y = f (X); Y = f (X^2); Y = f (X^3)

6. Normality assumption-The disturbance term U_i is assumed to have a normal distribution with zero mean and a constant variance. This assumption is given as follows:

$U_i \sim N(0, \delta_u^2)$. This assumption is a combination of zero mean of error term assumption and homoscedasticity assumption.

2.4. Methods of estimation

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

1. Ordinary least square method (OLS)

2. Maximum likelihood method (MLM)

3. Method of moments (MM)

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

2.5. Obtaining the Ordinary Least Squares (OLS) Estimates

Before obtaining the least squares estimators, we are going to examine three alternative methods to illustrate the problem in hand.

7. The explanatory variables are measured without error - U absorbs the influence of omitted variables and possibly errors of measurement in the y's. i.e., we will assume that the regressors are error free, while y values may or may not include errors of measurement.

We can now use the above assumptions to derive the following

basic concepts.

A. The dependent variable Yi is normally distributed

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

Proof:

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

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

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

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

$$= E(u_i)^2$$

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

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

The shape of the distribution of i Y is determined by the shape of the distribution of iu which is normal by assumption 4. Since β and α, being constant, they don't affect the distribution of i y . Furthermore, the values of the explanatory variable, **xi** , are a set of fixed values by assumption 5 and therefore don't affect the shape of the distribution of **yi** .

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

B. successive values of the dependent variable are independent

$$Cov(Y_i, Y_j) = 0$$

Proof:

$$Cov(Y_i, Y_j) = E\{[Y_i - E(Y_i)][Y_j - E(Y_j)]\}$$

$$= E\{[\alpha + \beta X_i + U_i - E(\alpha + \beta X_i + U_i)][\alpha + \beta X_j + U_j - E(\alpha + \beta X_j + U_j)\}$$

$$\text{(Since } Y_i = \alpha + \beta X_i + U_i \text{ and } Y_j = \alpha + \beta X_j + U_j)$$

$$= E[(\alpha + \beta X_i + Ui - \alpha - \beta X_i)(\alpha + \beta X_j + U_j - \alpha - \beta X_j)], \text{Since } E(u_i) = 0$$

$$= E(U_i U_j) = 0 \quad \text{(from equation (2.5))}$$

Therefore, $Cov(Y_i, Y_j) = 0$.

There three methods in common is that they try to minimize the residuals as a whole.

Criterion 1 The first criterion takes as estimators those values of α and β that make the sum of all the residuals as near to zero as possible. According to this criterion, the expression to minimize would be the following:

$$\left|\sum_{i=1}^{3} \hat{u}_i = 0\right|$$

Criterion 2 In order to avoid the compensation of positive residuals with negative ones, the absolute values from the residuals are taken. In this case, the following expression would be minimized:

$$\text{Min} \sum_{i=1}^{n} |\hat{u}_i|$$

Criterion 3 A third procedure is to minimize the sum of the square residuals, that is to say

$$\text{Min } S = \text{Min} \sum_{i=1}^{n} \hat{u}_i^2$$

The estimators obtained are denominated least square estimators (LS), and they enjoy certain desirable statistical properties, which will be studied later on. On the other hand, as opposed to the first of the examined criteria, when we square the residuals their compensation is avoided, and the least square estimators are simple to obtain, contrary to the second of the criteria.

2.6. Ordinary Least Square Method

The (Ordinary) least square (OLS) method of estimating parameters or regression function is about finding or estimating values of the parameters ($\alpha \ and \ \beta$) of the simple linear regression function given below for which the errors or residuals are minimized. Thus, it is about minimizing the residuals or the errors.

$$Y_i = \alpha + \beta X_i + U_i \qquad 2.5$$

The above identity represents population regression function (to be estimated from total enumeration of data from the entire population). But, most of the time it is difficult to generate population data owing to several reasons; and most of the time we use sample data and we estimate sample regression function. Thus, we use the following sample regression function for the derivation of the parameters and related analysis.

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i \qquad 2.6$$

OLS method of Estimation

Estimating a linear regression function using the Ordinary Least Square (OLS) method is simply about calculating the parameters of the regression function for which the sum of square of the error terms is minimized. The procedure is given as follows. Suppose we want to estimate the following equation

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

Since most of the time we use sample (or it is difficult to get population data) the corresponding sample regression function is given as follows.

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

From this identity, we solve for the residual term $'e_i'$, square both sides and then take sum of both sides. These three steps are given (respectively as follows.

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \qquad 2.7$$

$$\sum e_i^2 = \sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2 \qquad 2.8$$

Where, $\sum e_i^2$ =RSS= Residual Sum of Squares.

The method of OLS involves finding the estimates of the intercept and the slope for which the sum squares given by the Equation is minimized. To minimize the residual sum of squares we take the first order partial derivatives of Equation 2.8 and equate them to zero.

That is, the partial derivative with respect to $\hat{\beta}_0$:

$$\frac{\partial \sum e_i^2}{\partial \hat{\beta}_0} = 2\sum\left(Y_i - \hat{\beta}_0 - \hat{\beta}X_i\right)(-1) = 0 \qquad 2.9$$

$$\Rightarrow \sum\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) = 0 \qquad 2.10$$

$$\Rightarrow \sum Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum X_i = 0 \qquad 2.11$$

$$\Rightarrow \sum Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum X_i \qquad 2.1$$

Where n is the sample size.

Partial derivative With respect to $\hat{\beta}_1$

$$\frac{\partial \sum e_i^2}{\partial \hat{\beta}_1} = 2\sum\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)(-X_i) = 0 \qquad 2.13$$

$$\sum\left(Y_i X_i - \hat{\beta}_0 X_i - \hat{\beta}_1 X_i^2\right) = 0 \qquad 2.14$$

$$\sum X_i Y_i = \hat{\beta}_0 \sum X_i + \hat{\beta}_1 \sum X_i^2 \qquad 2.15$$

Note that the equation $\sum(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$ is a composite function and we should apply a chain rule in finding the partial derivatives with respect to the parameter estimates.

Equations 2.12 and 2.15 are together called the system of **the normal equations**. Solving the system of normal equations simultaneously we obtain:

$$\hat{\beta}_1 = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$$

Or

$$\hat{\beta}_1 = \frac{\sum XY - n\bar{Y}\bar{X}}{\sum X_i^2 - n\bar{X}^2} \quad \text{and we have } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

2.6. Estimation of elasticity from regression equation

Elasticity is a measure of how responsive a dependent variable is to a small change in an independent variable(s).

Elasticity is defined as a ratio of the percentage change in the dependent variable to the percentage change in the independent variable.

We first measured own elasticity using mid-point elasticity estimation method as shown in the following equation

$$\mathbf{E_A = \%\Delta Q_A / \%\Delta P_A}$$

1

Upon algebraic rearrangement the above equation (1) can be expressed as follows

$$\mathbf{E_A = (\Delta Q_A / \Delta P_A)^*(\bar{P}_A / \bar{Q}_A)}$$

Example: Given the following sample data of three pairs of 'Y' (dependent variable) and 'X' (independent variable), find a simple linear regression function; Y = f(X).

Y_i	X_i
10	30
20	50
30	60

a) find a simple linear regression function; Y = f(X)

b) Interpret your result.

c) Predict the value of Y when X is 45.

Solution

To fit the regression equation, we do the following computations.

	Y_i	X_i	$Y_i X_i$	X_i^2
	10	30	300	900
	20	50	1000	2500
	30	60	1800	3600
Sum	60	140	3100	7000
Mean	$\bar{Y} = 20$	$\bar{X} = \frac{140}{3}$		

$$\hat{\beta}_1 = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2} = \frac{3(3100) - (140)(60)}{3(7000) - (140)^2} = 0.64$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 20 - 0.64(140/3) = -10$$

Thus the fitted regression function is given by: $\hat{Y}_i = -10 + 0.64\ X_i$

b) Interpretation, the value of the intercept term,-10, implies that the value of the dependent variable 'Y' is – 10 when the value of the explanatory variable is zero. The value of the slope coefficient ($\hat{\beta} = 0.64$) is a measure of the marginal change in the dependent variable 'Y' when the value of the explanatory variable increases by one. For instance, in this model, the value of 'Y' increases on average by 0.64 units when 'X' increases by one.

c) $\hat{Y}_i = -10 + 0.64\ X_i$ =-10+(0.64)(45)=18.8

That means when X assumes a value of 45, the value of Y on average is expected to be 18.8. The regression coefficients can also be obtained by simple formulae by taking the deviations between the original values and their means. Now, if

$x_i = X_i - \overline{X}$, and $y_i = Y_i - \overline{Y}$

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

Then, the coefficients can be represented by alternative formula given below ,

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

Find the regression equation for the data under Example 2.4, using the shortcut formula. To solve this problem, we proceed as follows.

	Y_i	**X_i**	**Y**	**x**	**xy**	**x^2**	**y^2**

	10	30	-10	-16.67	166.67	277.78	100
	20	50	0	3.33	0.00	11.11	0
	30	60	10	13.33	133.33	177.78	100
Sum	60	140	0	0	300.00	466.67	200
Mean	20	46.66667					

Then

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{300}{466.67} = 0.64$$

, and $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ =20-(0.64) (46.67) = -10 with results similar to previous case.

1.7. Properties of Ols Estimators

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

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

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

a. Linear: a linear function of the random variable, such as, the dependent variable Y.

b. Unbiased: its average or expected value is equal to the true population parameter.

c. Best: It has a minimum variance in the class of linear and unbiased estimators. An unbiased estimator with the least variance is known as an efficient estimator.

According to the Gauss-Markov theorem, the OLS estimators possess all the BLUE properties.

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

Proposition: β is linear in Y.

$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$ *but* $y_i = Y_i - \bar{Y}$ *and substitute this in place of* y_i

$\hat{\beta}_1 = \frac{\sum x_i (Y_i - \bar{Y})}{\sum x_i^2} = \frac{\sum x_i Y_i}{\sum x_i^2} - \frac{\bar{Y} \sum x_i}{\sum x_i^2}$ *but* $\sum x_i = 0$

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

Let $K_i = \frac{x_i}{\sum x_i^2}$

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

K_i *has the following properties;*

$\sum K_i = 0$. *This is because* $\sum K_i = \frac{\sum x}{\sum x_i^2} = \frac{0}{\sum x_i^2} = 0$

$\sum K_i X_i = \sum K_i x_i = 1$. *This is because* $\sum K_i X_i = \frac{\sum x_i X_i}{\sum x_i^2}$ *but* $x_i = X_i - \bar{X}$ *and substitute* int *o the formula*

$= \frac{\sum (X_i - \bar{X}) X_i}{\sum x_i^2} = \frac{\sum X_i^2}{\sum x_i^2} - \frac{\bar{X} \sum X_i}{\sum x_i^2}$. *But* $\sum X_i = n \bar{X}$ *and substitute in the formula*

$= \frac{\sum X_i^2}{\sum x_i^2} - \frac{n \bar{X}^2}{\sum x_i^2} = \frac{\sum X_i^2 - n \bar{X}^2}{\sum x_i^2}$. *But* $\sum X_i^2 - n \bar{X}^2 = \sum x_i^2 = \frac{\sum x_i^2}{\sum x_i^2} = 1$

$\sum K_i^2 = \frac{1}{\sum x_i^2}$. *This is because* $K_i = \frac{x_i}{\sum x_i^2}$ *and* $K_i^2 = \frac{x_i^2}{(\sum x_i^2)^2}$. *Therefore,* $\sum K_i^2 = \frac{\sum x_i^2}{(\sum x_i^2)^2} = \frac{1}{\sum x_i^2}$

Proof α or β_0 is linear to Y_i

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \,.But\, \hat{\beta}_1 = \sum K_i Y_i$$

$$= \bar{Y} - \bar{X} \sum K_i Y_i$$

$$= \sum (\frac{1}{n} - \bar{X} K_i) Y_i$$

$$= \sum (\frac{1}{n} - \bar{X} K_i)(\beta_0 + \beta_1 X_i + U_i)$$

$$= \sum (\frac{\beta_0}{n} + \frac{\beta_1 X_i}{n} + \frac{U_i}{n} - \bar{X}\beta_0 K_i - \bar{X}\beta_1 K_i X_i - \bar{X} K_i U_i)$$

$$= \frac{\sum \beta_0}{n} + \frac{\beta_1 \sum X_i}{n} + \frac{\sum U_i}{n} - \bar{X}\beta_0 \sum K_i - \bar{X}\beta_1 \sum K_i X_i - \bar{X} \sum K_i U_i$$

$$= \frac{n\beta_0}{n} + \beta_1 \bar{X} - \beta_1 \bar{X} - \bar{X} \sum K_i U_i$$

$$\hat{\beta}_0 = \beta_0 - \bar{X} \sum K_i U_i$$

b. Unbiasedness:

Proposition: $\hat{\alpha}$ & $\hat{\beta}$ are the unbiased estimators of the true parameters α & β From your statistics course, you may recall that if $\hat{\theta}$ is an estimator of θ then $E(\hat{\theta}) - \theta = the amount of bias$ and if $\hat{\theta}$ is the unbiased estimator of θ then bias =0 i.e. $E(\hat{\theta}) - \theta = 0 \Rightarrow E(\hat{\theta}) = \theta$

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

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

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

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

$$= \alpha\Sigma k_i + \beta\Sigma k_i X_i + \Sigma k_i u_i \quad ,$$

$$but \;\; \Sigma k_i = 0 \qquad and \quad \Sigma k_i X_i = 1$$

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

$$\Rightarrow \sum k_i = 0 \quad \textbf{.....(2.24)}$$

$$\Sigma k_i X_i = \frac{\Sigma x_i X_i}{\Sigma x_i^2} = \frac{\Sigma(X - \bar{X})Xi}{\Sigma x_i^2}$$

$$= \frac{\Sigma X^2 - \bar{X}\Sigma X}{\Sigma X^2 - n\bar{X}^2} = \frac{\Sigma X^2 - n\bar{X}^2}{\Sigma X^2 - n\bar{X}^2} = 1$$

$$\Rightarrow \sum k_i X_i = 1 \ldots\ldots\ldots\ldots\ldots \quad \textbf{.....(2.25)}$$

$$\hat{\beta} = \beta + \Sigma k_i u_i \Rightarrow \hat{\beta} - \beta = \Sigma k_i u_i \text{ ----------(2.26)}$$

$\mathrm{E}(\hat{\beta}) = E(\beta) + \Sigma k_i E(u_i)$, **Since** k_i **are fixed**

$\mathrm{E}(\hat{\beta}) = \beta$, **since** $\mathrm{E}(u_i) = 0$

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

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

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

$$\hat{\alpha} = \Sigma\left(\tfrac{1}{n} - \bar{X}k_i\right)Y_i$$

$= \Sigma\left[\left(\tfrac{1}{n} - \bar{X}k_i\right)(\alpha + \beta X_i + U_i)\right]$, **Since** $Y_i = \alpha + \beta X_i + U_i$

$= \alpha + \beta\,\tfrac{1}{n}\Sigma X_i + \tfrac{1}{n}\Sigma u_i - \alpha\bar{X}\Sigma k_i - \beta\bar{X}\Sigma k_i X_i - \bar{X}\Sigma k_i u_i$

$= \alpha + \tfrac{1}{n}\Sigma u_i - \bar{X}\Sigma k_i u_i$, $\Rightarrow \hat{\alpha} - \alpha = \tfrac{1}{n}\Sigma u_i - \bar{X}\Sigma k_i u_i$

$$= \sum (1/n - \bar{X}k_i)u_i \text{ .....(2.27)}$$

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

$$E(\hat{\alpha}) = \alpha \text{ ----------} \quad (2.28)$$

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

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

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

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

$$\text{var}(\beta) = E(\hat{\beta} - E(\hat{\beta}))^2 = E(\hat{\beta} - \beta)^2 \text{ .....(2.29)}$$

Substitute (2.26) in (2.29) and we get

$$\text{var}(\hat{\beta}) = E(\sum k_i u_i)^2$$

$$= E[k_1^2 u_1^2 + k_2^2 u_2^2 + \ldots\ldots\ldots + k_n^2 u_n^2 + 2k_1 k_2 u_1 u_2 + \ldots\ldots + 2k_{n-1} k_n u_{n-1} u_n]$$

$$= E[k_1^2 u_1^2 + k_2^2 u_2^2 + \ldots\ldots\ldots + k_n^2 u_n^2] + E[2k_1 k_2 u_1 u_2 + \ldots\ldots + 2k_{n-1} k_n u_{n-1} u_n]$$

$$= E(\sum k_i^2 u_i^2) + E(\Sigma k_i k_j u_i u_j) \qquad i \neq j$$

$$= \Sigma k_i^2 E(u_i^2) + 2\Sigma k_i k_j E(u_i u_j) = \sigma^2 \Sigma k_i^2 \text{ (Since } E(u_i u_j) \text{=0)}$$

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

$$\therefore \text{var}(\hat{\beta}) = \sigma^2 \Sigma k_i^2 = \frac{\sigma^2}{\Sigma x_i^2} \text{ ……..(2.30)}$$

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

$$\text{var}(\hat{\alpha}) = \text{E}\left((\hat{\alpha} - \text{E}(\alpha)\right)^2$$

$$= \text{E}\left(\hat{\alpha} - \alpha\right)^2 \text{ ----------(2.31)}$$

Substituting equation (2.27) in (2.31), we get

$$\text{var}(\hat{\alpha}) = \text{E}\left[\Sigma\left(\tfrac{1}{n} - \bar{X}k_i\right)^2 u_i^2\right]$$

$$= \sum\left(\tfrac{1}{n} - \bar{X}k_i\right)^2 \text{E}(u_i)^2$$

$$= \sigma^2 \Sigma(\tfrac{1}{n} - \bar{X}k_i)^2$$

$$= \sigma^2 \Sigma(\tfrac{1}{n^2} - \tfrac{2}{n}\bar{X}k_i + \bar{X}^2 k_i^2)$$

$$= \sigma^2 \Sigma(\tfrac{1}{n} - \tfrac{2\bar{X}}{n}\Sigma k_i + \bar{X}^2 \Sigma k_i^2) \text{ , Since } \sum k_i = 0$$

$$= \sigma^2 (\tfrac{1}{n} + \bar{X}^2 \Sigma k_i^2)$$

$$= \sigma^2 (\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2}) \text{ , Since } \Sigma k_i^2 = \frac{\Sigma x_i^2}{(\Sigma x_i^2)^2} = \frac{1}{\Sigma x_i^2}$$

Again:

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

$$\therefore \text{var}(\hat{\alpha}) = \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\Sigma x_i^2}\right) = \sigma^2\left(\frac{\Sigma X_i^2}{n\Sigma x_i^2}\right) \quad \text{.....(2.32)}$$

Therefore, Mean and Variance of Parameter Estimates have the following formula.

1. The mean of $\hat{\beta}_1 = E(\hat{\beta}_1) = \beta_1$

2. The variance of $\hat{\beta}_1 = Var(\hat{\beta}_1) = E((\hat{\beta}_1 - E(\hat{\beta}))^2 = \frac{\delta_U^2}{\sum x_i^2}$

3. The mean of $\hat{\beta}_0 = E(\hat{\beta}_0) = \hat{\beta}_0$

4. The variance of $\hat{\beta}_0 = E((\hat{\beta}_0 - E(\hat{\beta}_0))^2 = \frac{\delta_U^2 \sum X_i^2}{n\sum x_i^2}$

5. The estimated value of the variance of the error term $\hat{\delta_U}^2 = \frac{\sum e_i^2}{n-2}$

Hypothesis Testing of OLS Estimates

After estimation of the parameters there are important issues to be considered by the researcher. We have to know that to what extent our estimates are reliable enough and acceptable for further purpose. That means, we have to evaluate the degree of representativeness of the estimate to the true population parameter. Simply a model must be tested for its significance before it can be used for any other purpose. In this subsection we will evaluate the reliability of model estimated using the procedure we explained above.

The available test criteria are divided in to three groups: Theoretical a priori criteria, statistical criteria and econometric criteria. Priori criteria set by economic theories are in line with the consistency of coefficients of econometric model to the economic theory. Statistical criteria, also known as first order tests, are set by statistical theory and refer to evaluate the statistical reliability of the model. Econometric criteria refer to whether the assumptions of an

econometric model employed in estimating the parameters are fulfilled or not. There are two most commonly used statistical tests in econometrics. These are:

1. The square of correlation coefficient (r^2) which is used for judging the explanatory power of the linear regression of Y on X or on X's. The square of correlation coefficient in simple regression is known as the coefficient of determination and is given by R^2. The coefficient of determination measures the **goodness of fit** of the line of regression on the observed sample values of Y and X.

2. The standard error test of the parameter estimates applied for judging the statistical reliability of the estimates. This test measures the degree of confidence that we may attribute to the estimates.

i) The Coefficient of determination (R^2)

Let's see **Algebraic implications of the estimation** before going to see coefficient of determination.

The algebraic implications of the estimation are derived exclusively from the application of the OLS procedure to the simple linear regression model:

1. The sum of the OLS residuals is equal to 0:

$$\sum_{i=1}^{n} \hat{u}_i = 0$$

From the definition of residual

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i \qquad i = 1, 2, \cdots, n$$

If we sum up the n observations, we get

$$\sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0$$

which is precisely the first equation of the system of normal equations. Note that, if it holds, it implies that

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

and, dividing the above equation by n, we obtain

$$\bar{\hat{u}} = 0 \qquad \bar{y} = \bar{\hat{y}}$$

2. The OLS line always goes through the mean of the sample (, x y).

Effectively, dividing the first normal equation by n, we have:

$$\bar{y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{x}$$

3. The sample cross product between each one of the regressors and the OLS residuals is zero.

That is to say,

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0$$

We can see that is equal to the second normal equation

$$\sum_{i=1}^{n} x_i \hat{u}_i = \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0$$

4. The sample cross product between the fitted values (ˆ y) and the OLS residuals is zero. That is to say,

$$\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = 0$$

Proof

Taking into account the algebraic implications, we have

$$\sum_{i=1}^{n}\hat{y}_i\hat{u}_i=\sum_{i=1}^{n}(\hat{\beta}_1+\hat{\beta}_2x_i)\hat{u}_i=\hat{\beta}_1\sum_{i=1}^{n}\hat{u}_i+\hat{\beta}_2\sum_{i=1}^{n}x_i\hat{u}_i=\hat{\beta}_1\times 0+\hat{\beta}_2\times 0=0$$

Decomposition of the variance of y

By definition

$$y_i=\hat{y}_i+\hat{u}_i$$

Subtracting y mean on both sides of the previous expression (remember that ˆ y hated is equal to y), we have

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

Squaring both sides:

$$\left[y_i-\overline{y}\right]^2=\left[(\hat{y}_i-\overline{\hat{y}})+\hat{u}_i\right]^2=(\hat{y}_i-\overline{\hat{y}})^2+\hat{u}_i^2+2\hat{u}_i(\hat{y}_i-\overline{\hat{y}})$$

Summing for all i:

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

Taking into account the algebraic properties 1 and 4, the third term of the right-hand side is equal to 0. Analytically

$$\sum\hat{u}_i(\hat{y}_i-\overline{\hat{y}})=\sum\hat{u}_i\hat{y}_i-\overline{\hat{y}}\sum\hat{u}_i=0$$

Therefore, we have

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

In words,

Total sum of squares (TSS) = Explained sum of squares (ESS)+Residual sum of squares (RSS)

If there is no intercept in the fitted model, then in general the decomposition obtained will not be fulfilled.

This decomposition can be made with variances, by dividing both sides of by n:

$$\frac{\sum(y_i-\bar{y})^2}{n}=\frac{\sum(\hat{y}_i-\bar{\hat{y}})^2}{n}+\frac{\sum\hat{u}_i^2}{n}$$

In words,

Total variance=explained variance+ residual variance

The coefficient of determination is the measure of the amount or proportion of the total variation of the dependent variable

that is determined or explained by the model or the presence of the explanatory variable in the model. The total variation of the dependent variable is split in two additive components; a part explained by the model and a part represented by the random term. The total variation of the dependent variable is measured from its arithmetic mean.

$$Total\,\text{var}\,iation in Y_i=\sum(Y_i-\bar{Y})^2$$
$$Total\,\text{exp}\,lained\,\text{var}\,iation=\sum(\hat{Y}_i-\bar{Y})^2$$
$$Total\,un\,\text{exp}\,lained\,\text{var}\,iation=\sum e_i^2$$

The total variation of the dependent variable is given in the following form; TSS=ESS + RSS, which means total sum of square of the dependent variable is split into explained sum of square and residual sum of square.

The coefficient of determination is given by the formula 2.16

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum(y_i - \bar{y})^2 - \sum \hat{u}_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum \hat{u}_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS}$$

Therefore, is equal to 1 minus the proportion of the total sum of squares (TSS) that is non-explained by the regression (RSS).

Since $\hat{y} = \hat{\beta}x$.

Squaring and summing both sides give us

$\Sigma\hat{y}^2 = \hat{\beta}^2\Sigma x^2$ -

We can substitute in the equation and obtain

$$ESS / TSS = \frac{\hat{\beta}^2\Sigma x^2}{\Sigma y^2} \text{ .....}$$

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

$$= \frac{\Sigma xy}{\Sigma x^2}\frac{\Sigma xy}{\Sigma y^2} \text{ .....}$$

Comparing with the formula of the correlation coefficient:

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

Squaring will result in:

$$\boxed{r^2 = (\Sigma xy)^2 / (\Sigma x^2 \Sigma y^2).}$$

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

$R^2 = \frac{\hat{\beta}_1 \sum x_i y_i}{\sum y_i^2}$ the coefficient of determination can also be given as

The higher the coefficient of determination is the better the fit. Conversely, the smaller the coefficient of determination is the poorer the fit. That is why the coefficient of determination is used to compare two or more models. One minus the coefficient of determination is called the coefficient of non-determination, and it gives the proportion of the variation in the dependent variable that remained undetermined or unexplained by the model.

According to the definition of coefficient determination, the following must be accomplished

$$0 \le R^2 \le 1$$

Extreme cases:

a) If we have a perfect fit, then $\hat{u}_t = 0 \quad \forall i$. This implies that

$$\hat{y}_t = y_t \quad \forall i \quad \Rightarrow \sum (\hat{y}_i - \bar{\hat{y}})^2 = \sum (y_i - \bar{y})^2 \quad \Rightarrow R^2 = 1$$

b) If $\hat{y}_t = c \quad \forall i$, it implies that

$$\bar{\hat{y}} = c \quad \Rightarrow \hat{y}_i - \bar{\hat{y}} = c - c = 0 \quad \forall i \quad \Rightarrow \sum (\hat{y}_i - \bar{\hat{y}})^2 = 0 \quad \Rightarrow R^2 = 0$$

If R^2 is close to 0, it implies that we have a poor fit. In other words, very little variation in y is explained by x.

ii) Testing the significance of a given regression coefficient

Since the sample values of the intercept and the coefficient are estimates of the true population parameters, we have to test them for their statistical reliability.

The significance of a model can be seen in terms of the amount of variation in the dependent variable that it explains and the significance of the regression coefficients.

There are different tests that are available to test the statistical reliability of the parameter estimates. The following are the common ones;

A) The standard error test

B) The standard normal test

C) The students t-test

Now, let us discuss them one by one.

A) The Standard Error Test

This test first establishes the two hypotheses that are going to be tested which are commonly known as the null and alternative hypotheses. The null hypothesis addresses that the sample is coming from the population whose parameter is not significantly different from zero while the alternative hypothesis addresses that the sample is coming from the population whose parameter is significantly different from zero. The two hypotheses are given as follows:

H_0: $\beta_i=0$

H_1: $\beta_i\neq 0$

The standard error test is outlined as follows:

1. Compute the standard deviations of the parameter estimates using the above formula for variances of parameter estimates. This is because standard deviation is the positive square root of the variance.

$$se(\hat{\beta}_1) = \sqrt{\frac{\delta_U^2}{\sum x_i^2}}$$

$$se(\hat{\beta}_0) = \sqrt{\frac{\delta_U^2 \sum X_i^2}{n\sum x_i^2}}$$

2. Compare the standard errors of the estimates with the numerical values of the estimates and make decision.

A) If the standard error of the estimate is less than half of the numerical value of the estimate, we can conclude that the estimate is statistically significant. That is, if $se(\hat{\beta}_i) < \frac{1}{2}(\hat{\beta}_i)$, reject the null hypothesis and we can conclude that the estimate is statistically significant.

B) If the standard error of the estimate is greater than half of the numerical value of the estimate, the parameter estimate is not statistically reliable. That is, if $se(\hat{\beta}_i) > \frac{1}{2}(\hat{\beta}_i)$, conclude to accept the null hypothesis and conclude that the estimate is not statistically significant.

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

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

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

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

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

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

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

B) The Standard Normal Test

This test is based on the normal distribution. The test is applicable if:

- The standard deviation of the population is known irrespective of the sample size

- The standard deviation of the population is unknown provided that the sample size is sufficiently large (n>30).

The standard normal test or Z-test is outline as follows;

1. Test the null hypothesis $H_0 : \beta_i = 0$ against the alternative hypothesis $H_1 : \beta_i \neq 0$
2. Determine the level of significant (α) in which the test is carried out. It is the probability of committing type I error, i.e. the probability of rejecting the null hypothesis while it is true. It is common in applied econometrics to use 5% level of significance.
3. Determine the theoretical or tabulated value of Z from the table. That is, find the value of $Z_{\alpha/2}$ from the standard normal table. $Z_{0.025} = 1.96$ from the table.
4. Make decision. The decision of statistical hypothesis testing consists of two decisions; either accepting the null hypothesis or rejecting it.

If $|Z_{cal}| < Z_{tab}$, accept the null hypothesis while if $|Z_{cal}| > Z_{tab}$, reject the null hypothesis. It is true that most of the times the null and alternative hypotheses are mutually exclusive. Accepting the null hypothesis means that rejecting the alternative hypothesis and rejecting the null hypothesis means accepting the alternative hypothesis.

Example: If the regression has a value of $\hat{\beta}_1$ =29.48 and the standard error of $\hat{\beta}_1$ is 36. Test the hypothesis that the value of $\beta_1 = 25$ at 5% level of significance using standard normal test.

Solution: We have to follow the procedures of the test.

$H_0 : \beta_1 = 25$
$H_1 : \beta_1 \neq 25$

After setting up the hypotheses to be tested, the next step is to determine the level of significance in which the test is carried out. In the above example the significance level is given as 5%.

The third step is to find the theoretical value of Z at specified level of significance. From the standard normal table we can get that $Z_{0.025} = 1.96$.

The fourth step in hypothesis testing is computing the observed or calculated value of the standard normal distribution using the following formula.

$Z_{cal} = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} = \frac{29.48 - 25}{36} = 0.12$. Since the calculated value of the test statistic is less than the tabulated value, the decision is to accept the null hypothesis and conclude that the value of the parameter is 25.

C) The Student t-Test

In conditions where Z-test is not applied (in small samples), t-test can be used to test the statistical reliability of the parameter estimates. The test depends on the degrees of freedom that the sample has. The test procedures of t-test are similar with that of the z-test. The procedures are outlined as follows;

1. Set up the hypothesis. The hypotheses for testing a given regression coefficient is given by:

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

2. Determine the level of significance for carrying out the test. We usually use a 5% level significance in applied econometric research.

3. Determine the tabulated value of t from the table with n-k degrees of freedom, where k is the number of parameters estimated.

4. Determine the calculated value of t. The test statistic (using the t- test) is given by:

$t_{cal} = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)}$

The test rule or decision is given as follows:

Reject H_0 if $|t_{cal}| \geq t_{\alpha/2,n-k}$

Numerical Example:

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

$$C = 100 + 0.70 + e$$
$$(75.5) \quad (0.21)$$

The values in the brackets are standard errors. We want to test the null hypothesis: H0: β=0 against the alternative H1: $\beta \neq 0$ using the t-test at 5% level of significance.

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

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

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

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

iii) Confidence Interval Estimation of the regression Coefficients

We have discussed the important tests that that can be conducted to check model and parameters validity. But one thing that must be clear is that rejecting the null hypothesis does not mean that the parameter estimates are correct estimates of the true population parameters. It means that

the estimate comes from the sample drawn from the population whose population parameter is significantly different from zero. In order to define the range within which the true parameter lies, we must construct a confidence interval for the parameter. Like we constructed confidence interval estimates for a given population mean, using the sample mean (in Introduction to Statistics), we can construct 100(1- α) % confidence intervals for the sample regression coefficients. To do so we need to have the standard errors of the sample regression coefficients. The standard error of a given coefficient is the positive square root of the variance of the coefficient.

Thus, we have discussed that the formulae for finding the variances of the regression coefficients are given as.

$$\text{var}(\hat{\beta}_0) = \delta_u{}^2 \frac{\sum X_i{}^2}{n\sum x_i{}^2}$$

Variance of the intercept is given by

$$\text{var}(\hat{\beta}_1) = \delta_u{}^2 \frac{1}{\sum x_i{}^2}$$

Variance of the slope $(\hat{\beta}_1)$ is given by

Where, $\delta_u{}^2 = \frac{\sum e_i{}^2}{n-k}$ is the estimate of the variance of the random term and k is the number of parameters to be estimated in the model. The standard errors are the positive square root of the variances and the 100 (1- α) % confidence interval for the slope is given by:

$$\hat{\beta}_1 - t_{\alpha/2}(n-k)(se(\hat{\beta}_1)) \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2}(n-k)(se(\hat{\beta}_1))$$

$$\beta_1 = \hat{\beta}_1 \pm t_{\alpha/2,n-k}(se(\hat{\beta}_1))$$

And for the intercept:

$$\beta_0 = \hat{\beta}_0 \pm t_{\alpha/2,n-k}(se(\hat{\beta}_0))$$

Numerical Example:

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

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

The values in the bracket are standard errors.

a. Construct 95% confidence interval for the slope of parameter

b. Test the significance of the slope parameter using constructed confidence interval.

Solution:

a.

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

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

$\hat{\beta} = 2.88$

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

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

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

The confidence interval is:

(1.09, 4.67)

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

Example: The following table gives the quantity supplied (Y in tons) and its price (X pound per ton) for a commodity over a period of twelve years.

Data on supply and price for given commodity

Y	69	76	52	56	57	77	58	55	67	53	72	64
X	9	12	6	10	9	10	7	8	12	6	11	8

Data for computation of different parameters

Time	Y	X	XY	X^2	Y^2	x	y	xy	x^2	y^2	$\hat{Y}$	e_i	e_i^2
1	69	9	621	81	4761	0	6	0	0	36	63.00	6.00	36.00
2	76	12	912	144	5776	3	13	39	9	169	72.75	3.25	10.56
3	52	6	312	36	2704	-3	-11	33	9	121	53.25	-1.25	1.56
4	56	10	560	100	3136	1	-7	-7	1	49	66.25	-10.25	105.06
5	57	9	513	81	3249	0	-6	0	0	36	63.00	-6.00	36.00
6	77	10	770	100	5929	1	14	14	1	196	66.25	10.75	115.56
7	58	7	406	49	3364	-2	-5	10	4	25	56.50	1.50	2.25
8	55	8	440	64	3025	-1	-8	8	1	64	59.75	-4.75	22.56
9	67	12	804	144	4489	3	4	12	9	16	72.75	-5.75	33.06
10	53	6	318	36	2809	-3	-10	30	9	100	53.25	-0.25	0.06
11	72	11	792	121	5184	2	9	18	4	81	69.50	2.50	6.25
12	64	8	512	64	4096	-1	1	-1	1	1	59.75	4.25	18.06
Sum	756	108	6960	1020	48522	0	0	156	48	894	756.00	0.00	387.00

Use the above Tables to answer the following questions

1. Estimate the Coefficient of determination (R^2)

2. Run significance test of regression coefficients using the following test methods

   A) The standard error test

   B) The students t-test

C) Fit the linear regression equation and determine the 95% confidence interval for the slope.

Solution

1. Estimate the Coefficient of determination (R^2)

Refer to Example above to determine how much percent of the variations in the quantity supplied is explained by the price of the commodity and what percent remained unexplained.

use data in the above table to estimate r^2 using the formula given below.

$$R^2 = 1 - \frac{\sum e_i^{\,2}}{\sum y_i^{\,2}} = 1 - \frac{387}{894} = 1 - 0.43 = 0.57$$

This result shows that 57% of the variation in the quantity supplied of the commodity under consideration is explained by the variation in the price of the commodity; and the rest 37% remain unexplained by the price of the commodity. In other word, there may be other important explanatory variables left out that could contribute to the variation in the quantity supplied of the commodity, under consideration.

2. Run significance test of regression coefficients using the following test methods

Fitted regression line for the data given is:

$\hat{Y}_i = 33.75 + 3.25X_i$

(8.3) (0.9) , where the numbers in parenthesis are standard errors of the respective coefficients.

A. Standard Error test

In testing the statistical significance of the estimates using standard error test, the following information needed for decision.

Since there are two parameter estimates in the model, we have to test them separately.

Testing for $\hat{\beta}_1$

We have the following information about $\hat{\beta}_1$ i.e $\hat{\beta}_1$ =3.25 and $se(\hat{\beta}_1) = 0.9$

The following are the null and alternative hypotheses to be tested.

$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

Since the standard error of $\hat{\beta}_1$ is less than half of the value of $\hat{\beta}_1$, we have to reject the null hypothesis and conclude that the parameter estimate $\hat{\beta}_1$ is statistically significant.

Testing for $\hat{\beta}_0$

Again, we have the following information about $\hat{\beta}_0$

$\hat{\beta}_0 = 33.75 \quad and \quad se(\hat{\beta}_0) = 8.3$

The hypotheses to be tested are given as follows;

$H_0 : \beta_0 = 0$
$H_1 : \beta_0 \neq 0$

Since the standard error of $\hat{\beta}_0$ is less than half of the numerical value of $\hat{\beta}_0$, we have to reject the null hypothesis and conclude that $\hat{\beta}_0$ is statistically significant.

B. The students t-test

In the illustrative example, we can apply t-test to see whether price of the commodity is significant in determining the quantity supplied of the commodity under consideration? Use α=0.05.

The hypothesis to be tested is:

$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

The parameters are known. $\hat{\beta}_1 = 3.25, \quad \text{se}(\hat{\beta}_1) = 0.8979$

Then we can estimate t_{cal} as follows

$$t_{cal} = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)} = \frac{3.25}{0.8979} = 3.62$$

Further tabulated value for t is 2.228. When we compare these two values, the calculated t is greater than the tabulated value. Hence, we reject the null hypothesis. Rejecting the null hypothesis means, concluding that the price of the commodity is significant in determining the quantity supplied for the commodity.

In this part we have seen how to conduct the statistical reliability test using t-statistic. Now let us see additional information about this test. When the degrees of freedom is large, we can conduct t-test without consulting the t-table in finding the theoretical value of t. This rule is known as "2t-rule". The rule is stated as follows;

The t-table shows that the values of t changes very slowly if the degrees of freedom (n-k) are greater than 8. For example, the value of $t_{0.025}$ changes from 2.30 (when n-k=8) to 1.96(when n-k=∞). The change from 2.30 to 1.96 is obviously very slow. Consequently, we can ignore the degrees of freedom (when they are greater than 8) and say that the theoretical value of t_{cal} is 2.0. Thus, a two-tail test of a null hypothesis at 5% level of significance can be reduced to the following rules.

1. If t_{cal} is greater than 2 or less than -2, we reject the null hypothesis

2. If t_{cal} is less than 2 or greater than -2, accept the null hypothesis.

C) Fit the linear regression equation and determine the 95% confidence interval for the slope.

$$\hat{Y}_i = 33.75 + 3.25X_i$$

Fitted regression model is indicated (8.3) (0.9) ,where the numbers in parenthesis are standard errors of the respective coefficients. To estimate confidence interval, we need standard error which is determined as follows

$$\delta_u^{\,2} = \frac{\sum e_i^{\,2}}{n-k} = \frac{387}{12-2} = \frac{387}{10} = 38.7$$

$$\text{var}(\hat{\beta}_1) = \delta_u^{\,2} \frac{1}{\sum x^2} = 38.7(\frac{1}{48}) = 0.80625$$

The standard error of the slope is $se(\hat{\beta}_1) = \sqrt{\text{var}(\hat{\beta}_1)} = \sqrt{0.80625} = 0.8979$

The tabulated value of t for degrees of freedom 12-2=10 and $\alpha/2$=0.025 is 2.228.

Hence the 95% confidence interval for the slope is given by:

$\hat{\beta}_1 = 3.25 \pm (2.228)(0.8979) = 3.25 \pm 2 = (3.25-2, 3.25+2) = (1.25, 5.25)$. The result tells us that at the error probability 0.05, the true value of the slope coefficient lies between 1.25 and 5.25

UNIT FOUR: MULTIPLE LINEAR REGRESSION MODELS

Concept and Notations of Multiple Regression Models

Simple linear regression model (also called the *two-variable model*) is extensively discussed in the previous section. Such models assume that a dependent variable is influenced by only one explanatory variable. However, many economic variables are influenced by several factors or variables. Hence, simple regression models are unrealistic. There is no more practicality of such models except simple to understand. Very good examples, for this argument, are demand and supply in which they have several determinants each.

Adding more variables to the simple linear regression model leads us to the discussion of multiple regression models i.e. models in which the dependent variable (or regressand) depends on two or more explanatory variables, or regressors.

- The multiple linear regression (population regression function) in which we have one dependent variable Y, and k explanatory variables, $X_1, X_2, ...X_k$ is given by

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + u_i \qquad 3.1$$

Where, $\beta_0 =$ the intercept = value of Y_i when all X's are zero

β_i = are partial slope coefficients

u_i = the random term

In this model, for example, β_1 is the amount of change in Y_i when X_1 changes by one unit, keeping the effect of other variables constant. Similarly, β_2 is the amount of change in Y_i when X_2 changes by one unit, keeping the effect of other variables constant. The other slopes are also interpreted in the same way.

Although multiple regression equation can be fitted for any number of explanatory variables (equation 3.1), the simplest possible regression model, and three-variable regression will be

presented for the sake of simplicity. It is characterized by one dependent variable (Y) and two explanatory variables (X_1 and X_2). The model is given by:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_i \qquad 3.2$$

β_0 = the intercept = value of Y when both X_1 & X_2 are zero

β_1 = the change in Y, when X_1 changes by one unit, keeping the effect of X_2 constant.

β_2 = the change in Y, when X_2 changes by one unit, keeping the effect of X_1 constant

Assumptions of the Multiple Linear Regression

Each econometric method that would be used for estimation purpose has its own assumptions..

***Assumption 1*:** Randomness of u_i - the variable u is a real random variable.

Assumption 2: Zero mean of u_i - the random variable u_i has a zero mean for each value of X_i i.e. $E(u_i) = 0$

Assumption 3: Homoscedasticity of the random term - the random term u_i has constant variance. In other words, the variance of each u_i is the same for all the X_i values.

$$E(u_i^2) = \delta_u^2 \quad Constant$$

***Assumption 4*:** Normality of u_i - the values of each u_i are normally distributed $u_i \approx N(0, \delta_u^2)$

***Assumption 5*:** No autocorrelation or serial independence of the random terms - the successive values of the random term are not strongly correlated. The values of u_i (corresponding to x_i) are independent of the values of any other u_j (corresponding to X_j).

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

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

***Assumption 7*:** No errors of measurement in the $X's$ - the explanatory variables are measured without error.

Assumption 8: No perfect multicollinearity among the $X's$ - the explanatory variables are not perfectly linearly correlated.

Assumption 9: Correct specification of the model - the model has no specification error in that all the important explanatory variables appear explicitly in the function and the mathematical form is correctly defined (linear or non-linear form and the number of equations in the model).

Estimation of Partial Regression Coefficients

The process of estimating the parameters in the multiple regression model is similar with that of the simple linear regression model. The main task is to derive the normal equations using the same procedure as the case of simple regression. Like in the simple linear regression model case, OLS and Maximum Likelihood (ML) methods can be used to estimate partial regression coefficients of multiple regression models. But, due to their simplicity and popularity, OLS methods can be used. The OLS procedure consists in so choosing the values of the unknown parameters that the residual sum of squares is as small as possible.

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

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

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

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

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

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

Now it is time to state how (3.3) is estimated. Given sample observation on Y, X1 and X2 , we estimate (3.3) using the method of least square (OLS)

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

is sample relation between Y, X_1 & X_2.

$$e_i = Y_i - \hat{Y} = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2 \quad \text{.....(3.7)}$$

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

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

$$\frac{\partial\left[\sum e_i^2\right]}{\partial\hat{\beta}_1} = -2\sum X_{1i}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_1 X_{1i}\right) = 0 \quad \text{..... (3.9)}$$

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

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

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

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

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

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

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

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

$$\sum X_{1i} Y_i = (\overline{Y} - \hat{\beta}_1 \overline{X}_1 - \hat{\beta}_2 \overline{X}_2)\Sigma X_{1i} + \hat{\beta}_1 \Sigma X_{1i}^{\ 2} + \hat{\beta}_2 \Sigma X_{2i}$$

$$\Rightarrow \sum X_{1i} Y_i - \hat{Y}\Sigma X_{1i} = \hat{\beta}_1(\Sigma X_{1i}^{\ 2} - \overline{X}_1 \Sigma X_{2i}) + \hat{\beta}_2(\Sigma X_{1i} X_{2i} - \overline{X}_2 \Sigma X_{2i})$$

$$\Rightarrow \sum X_{1i} Y_i - n\overline{Y}\,\overline{X}_{1i} = \hat{\beta}_2(\Sigma X_{1i}^{\ 2} - n\overline{X}_{1i}^{\ 2}) + \hat{\beta}_2(\Sigma X_{1i} X_2 - n\overline{X}_1\,\overline{X}_2) \text{ ------- (3.15)}$$

We know that

$$\sum\left(X_i - Y_i\right)^2 = (\Sigma X_i Y_i - n\overline{X}_i Y_i) = \Sigma x_i y_i$$

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

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

$$\sum x_1 y = \hat{\beta}_1 \Sigma x_1^{\,2} + \hat{\beta}_2 \Sigma x_1 x_2 \quad \text{.....(3.16)}$$

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

$$\sum x_2 y = \hat{\beta}_1 \Sigma x_1 x_2 + \hat{\beta}_2 \Sigma x_2^{\,2} \quad \text{.....(3.17)}$$

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

$$\sum x_1 y = \hat{\beta}_1 \Sigma x_1^{\,2} + \hat{\beta}_2 \Sigma x_1 x_2 \quad \text{.....(3.18)}$$

$$\sum x_2 y = \hat{\beta}_1 \Sigma x_1 x_2 + \hat{\beta}_2 \Sigma x_2^{\,2} \quad \text{.....(3.19)}$$

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

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

$$\begin{pmatrix} \sum x_1^{\,2} & \sum x_1 x_2 \\ \sum x_1 x_2 & \sum x_2^{\,2} \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \sum x_2 y \\ \sum x_{3y} \end{pmatrix} \quad \text{.....(3.20)}$$

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

$$\hat{\beta}_1 = \frac{\Sigma x_1 y \,.\, \Sigma x_2^{\,2} - \Sigma x_1 x_2 \,.\, \Sigma x_2 y}{\Sigma x_1^{\,2} \,.\, \Sigma x_2^{\,2} - \Sigma (x_1 x_2)^2} \quad \text{..... (3.21)}$$

$$\hat{\beta}_2 = \frac{\Sigma x_2 y \,.\, \Sigma x_1^{\,2} - \Sigma x_1 x_2 \,.\, \Sigma x_1 y}{\Sigma x_1^{\,2} \,.\, \Sigma x_2^{\,2} - \Sigma (x_1 x_2)^2} \quad \text{..... (3.22)}$$

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

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

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

Partial-correlation coefficients

In order to remove the influence of X2 on Y, we regress Y on X2 and find the residual e_1 = Y*. To remove the influence of X2 on XI, we regress XI on X2 and find the residual $e_2 = X_1^*$; Y And X_1*; then represent the variations in Y and XI, respectively, left unexplained after removing the influence of X2 from both Y and XI. Therefore, the partial correlation coefficient is merely the simple correlation coefficient between the residuals Y* and X^*; (that is, ryxl.x2 = $^{ry^*X^*}$;).

Partial correlation coefficient range in value from -1 to +1 ('just as in the case of simple correlation coefficients). For example, ryxl.x2 = -1 refers to the case where there is an exact or perfect negative linear relationship between Y and XI after removing the common influence of X2 from both Y and X1.

However, ryxl.x2 = 1 indicates a perfect positive linear net relationship between Y and XI. And ryxl.x2 = 0 indicates no linear relationship between Y and X, when the common influence of X, has been removed from both Y and XI. As a result, XI can be omitted from the regression.

The sign of partial correlation coefficients is the same as that of the corresponding estimated parameter. For example, for the estimated regression equation Y = bo + blXl + b2X2, ryx1.x2 has the same sign as bl and ryx2.xl has the same sign as b,.

Partial correlation coefficients are used in multiple regression analysis to determine the relative importance of each explanatory variable in the model. The independent variable with the highest partial correlation coefficient with respect to the dependent variable contributes most to the explanatory power of the model and is entered first in a stepwise multiple regression analysis. It should be noted, however, that partial correlation coefficients give an ordinal, not a cardinal, measure of net correlation, and the sum of the partial correlation coefficients between the dependent and all the independent variables in the model need not add up to 1.

To find $r_{YX_1 \cdot X_2}$, we need to find first r_{YX_1}, r_{YX_2}, and $r_{X_1X_2}$.

$$r_{YX_1} = \frac{\sum x_1 y}{\sqrt{\sum x_1^2}\sqrt{\sum y^2}}$$

$$r_{YX_2} = \frac{\sum x_2 y}{\sqrt{\sum x_2^2}\sqrt{\sum y^2}}$$

$$r_{X_1X_2} = \frac{\sum x_2 x_1}{\sqrt{\sum x_2^2}\sqrt{\sum x_1^2}}$$

$$r_{YX_1 \cdot X_2} = \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{\sqrt{1 - r_{X_1X_2}^2}\sqrt{1 - r_{YX_2}^2}}$$

$$r_{YX_2 \cdot X_1} = \frac{r_{YX_2} - r_{YX_1} r_{X_1X_2}}{\sqrt{1 - r_{X_1X_2}^2}\sqrt{1 - r_{YX_1}^2}}$$

If ryx2.xl exceeds the absolute value of ryxl .x2, we conclude that X2 contributes more than XI to the explanatory power of the model.

Variance and Standard errors of OLS Estimators

Estimating the numerical values of the parameters is not enough in econometrics if the data are coming from the samples. The standard errors derived are important for two main purposes: to establish confidence intervals for the parameters and to test statistical hypotheses. They are important to look into their precision or statistical reliability. An estimator cannot be used for any purpose if it is not a good estimator. The precision of an estimator is measured by observing the standard error of the estimator.

Like in the case of simple linear regression, the standard errors of the coefficients are vital in statistical inferences about the coefficients. We use standard the error of a coefficient to construct confidence interval estimate for the population regression coefficient and to test the significance of the variable to which the coefficient is attached in determining the dependent variable in the model. In this section, we will see these standard errors. The standard error of a coefficient is the positive square root of the variance of the coefficient. Thus, we start with defining the variances of the coefficients.

Variance of the intercept $\left(\hat{\beta}_0\right)$

$$Var\left(\hat{\beta}_0\right) = \hat{\delta}^2_{u_i}\left[\frac{1}{n} + \frac{\bar{X}_1^2 \sum x^2{}_2 + \bar{X}_2^2 \sum x_1^2 - 2\bar{X}_1\bar{X}_2 \sum x_1 x_2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2}\right] \quad 3.17$$

Variance of $\hat{\beta}_1$

$$Var\left(\hat{\beta}_1\right) = \delta_u^2\left[\frac{\sum x_2^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2}\right] \quad 3.18$$

Variance of $\hat{\beta}_2$

$$Var(\hat{\beta}_2) = \hat{\delta}^2_u\left[\frac{\sum x_1^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2}\right]$$

Where,

$$\hat{\delta}_u^2 = \frac{\sum e_i^2}{n-3}$$

Equation 3.20 here gives the estimate of the variance of the random term. Then, the standard errors are computed as follows:

Standard error of $\hat{\beta}_0$

$$SE(\hat{\beta}_0) = \sqrt{Var(\hat{\beta}_0)}$$

Standard error of $\hat{\beta}_1$

$$SE(\hat{\beta}_1) = \sqrt{Var(\hat{\beta}_1)} \qquad 3.22$$

Standard error of $\hat{\beta}_2$

$$SE(\hat{\beta}_2) = \sqrt{Var(\hat{\beta}_2)}$$

Note: The OLS estimators of the multiple regression model have properties which are parallel to those of the two-variable model.

Coefficient of Multiple Determinations

In simple regression model we have discussed about the coefficient of determination and its interpretation. In this section, we will discuss the coefficient of multiple determinations which has an equivalent role with that of the simple model. As coefficient of determination is the square of the simple correlation in simple model,

- Coefficient of multiple determinations is the square of multiple correlation coefficients.

The coefficient of multiple determinations (R^2) is the measure of the proportion of the variation in the dependent variable that is explained jointly by the independent variables in the model. One minus R^2 is called the coefficient of non-determination. It gives the proportion of the variation in the dependent variable that remains unexplained by the independent variables in

the model. As in the case of simple linear regression, R^2 is the ratio of the explained variation to the total variation. Mathematically:

$$R^2 = \frac{\sum \hat{y}^2}{\sum y^2} \qquad 3.24$$

Or R^2 can also be given in terms of the slope coefficients $\hat{\beta}_1 \, and \, \hat{\beta}_2 \, as:$

$$R^2 = \frac{\hat{\beta}_1 \sum x_1 y + \hat{\beta}_2 \sum x_2 y}{\sum y^2} \qquad 3.25$$

In simple linear regression, the higher the R^2 means the better the model is determined by the explanatory variable in the model. In multiple linear regression, however, every time we insert additional explanatory variable in the model, the R^2 increases irrespective of the improvement in the goodness-of- fit of the model. That means high R^2 may not imply that the model is good.

Thus, we adjust the R^2 as follows:

$$R^2_{ady} = 1-(1-R^2)\frac{(n-1)}{(n-k)} \qquad 3.26$$

Where, k = the number of explanatory variables in the model.

In multiple linear regression, therefore, we better interpret the adjusted R^2 than the ordinary or the unadjusted R^2. We have known that the value of R^2 is always between zero and one. But the adjusted R^2 can lie outside this range even to be negative.

In the case of simple linear regression, R^2 is the square of linear correlation coefficient. Again as the correlation coefficient lies between -1 and +1, the coefficient of determination (R^2) lies between 0 and 1. The R^2 of multiple linear regression also lies between 0 and +1. The adjusted R^2, however, can sometimes be negative when the goodness of fit is poor. When the adjusted

R^2 value is negative, we considered it as zero and interpret as no variation of the dependent variable is explained by repressors.

Confidence Interval Estimation

Confidence interval estimation in multiple linear regressions follows the same formulae and procedures that we followed in simple linear regression. You are, therefore, required to practice finding the confidence interval estimates of the intercept and the slopes in multiple regressions with two explanatory variables.

Please recall that 100(1-α) % confidence interval for β_i is given as $\hat{\beta}_i \pm t_{\alpha/2,n-k} se(\hat{\beta}_i)$ where k is the number of parameters to be estimated or the number of variables (both dependent and explanatory)

Interpretation of the confidence interval: Values of the parameter lying in the interval are plausible with 100(1-α) % confidence.

Hypothesis Testing in Multiple Regressions

Hypothesis testing is important to draw inferences about the estimates and to know how representative the estimates are to the true population parameter.

a) Testing hypothesis about an individual partial regression coefficient;

b) Testing the overall significance of the estimated multiple regression model (finding out if all the partial slope coefficients are simultaneously equal to zero);

Testing individual regression coefficients

The tests concerning the individual coefficients can be done using the standard error test or the t-test. In all the cases the hypothesis is stated as:

a) $H_0 : \hat{\beta}_1 = 0$ $H_1 : \hat{\beta}_1 \neq 0$

b) $H_0 : \hat{\beta}_2 = 0$ $H_1 : \hat{\beta}_2 \neq 0$

$H_0 : \hat{\beta}_K = 0$ $H_1 : \hat{\beta}_K \neq 0$

In a) we will like to test the hypothesis that X1 has no linear influence on Y holding other variables constant. In b) we test the hypothesis that X2 has no linear relationship with Y holding other factors constant. The above hypotheses will lead us to a two-tailed test however, one-tailed test might also be important. There are two methods for testing significance of individual regression coefficients.

a) Standard Error Test: Using the standard error test we can test the above hypothesis.

Thus the decision rule is based on the relationship between the numerical value of the parameter and the standard error of the same.

(i) If $S(\hat{\beta}_i) > \frac{1}{2}\hat{\beta}_i$, we accept the null hypothesis, i.e. the estimate of β_i is not statistically significant.

Conclusion: The coefficient $(\hat{\beta}_i)$ is not statistically significant. In other words, it does not have a significant influence on the dependent variable.

(ii) If $S(\hat{\beta}_i) < \frac{1}{2}\hat{\beta}_i$, we fail to accept H0, i.e., we reject the null hypothesis in favour of the alternative hypothesis meaning the estimate of βi has a significant influence on the dependent variable.

Generalisation: The smaller the standard error, the stronger is the evidence that the estimates are statistically significant.

(b) t-test

The more appropriate and formal way to test the above hypothesis is to use the t-test. As usual we compute the **t**-ratios and compare them with the tabulated t-values and make our decision.

Therefore: $t_{cal} = \frac{\hat{\beta}_i}{S(\hat{\beta}_i)} \approx t(n-1)$

Decision Rule: accept H0 if $-t_{\frac{\alpha}{2}} < t_{cal} < t_{\frac{\alpha}{2}}$

Otherwise, reject the null hypothesis. Rejecting H_0 means, the coefficient being tested is significantly different from 0. Not rejecting H_0, on the other hand, means we don't have sufficient evidence to conclude that the coefficient is different from 0.

Testing the Overall Significance of Regression Model

Here, we are interested to test the overall significance of the observed or estimated regression line, that is, whether the dependent variable is linearly related to all of the explanatory variables. Hypotheses of such type are often called joint hypotheses. Testing the overall significance of the model means testing the null hypothesis that none of the explanatory variables in the model significantly determine the changes in the dependent variable. Put in other words, it means testing the null hypothesis that none of the explanatory variables significantly explain the dependent variable in the model. This can be stated as:

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_1 : \beta_i \neq 0, \text{ at least for one i.}$$

The test statistic for this test is given by:

$$F_{cal} = \frac{\frac{\sum \hat{y}^2}{k-1}}{\frac{\sum e^2}{n-k}}$$

Where, k is the number of explanatory variables in the model.

The results of the overall significance test of a model are summarized in the analysis of variance (ANOVA) table as follows.

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	F_{cal}
Regression	$SSE = \sum \hat{y}^2$	$k-1$	$MSE = \frac{\sum \hat{y}^2}{k-1}$	$F = \frac{MSE}{MSR}$

Residual	$SSR = \sum e^2$	$n-k$	$MSR = \frac{\sum e^2}{n-k}$	
Total	$SST = \sum y^2$	$n-1$		

The values in this table are explained as follows:

$$SSE = \sum \hat{y}^2 = \sum (\hat{Y}_i - \bar{Y})^2 = \text{Explained Sum of Squares}$$
$$SSR = \sum y_i^2 = \sum (Y_i - \hat{Y})^2 = \text{Unexplained Sum of Squares}$$
$$SST = \sum y^2 = \sum (Y_i - \bar{Y})^2 = \text{Total Sum of Squares}$$

These three sums of squares are related in such a way that

$$SST = SSE + SSR$$

This implies that the total sum of squares is the sum of the explained (regression) sum of squares and the residual (unexplained) sum of squares. In other words, the total variation in the dependent variable is the sum of the variation in the dependent variable due to the variation in the independent variables included in the model and the variation that remained unexplained by the explanatory variables in the model. Analysis of variance (ANOVA) is the technique of decomposing the total sum of squares into its components. As we can see here, the technique decomposes the total variation in the dependent variable into the explained and the unexplained variations. The degrees of freedom of the total variation are also the sum of the degrees of freedom of the two components. By dividing the sum of squares by the corresponding degrees of freedom, we obtain what is called the **Mean Sum of Squares (MSS)**.

The Mean Sum of Squares due to regression, errors (residual) and Total are calculated as the Sum of squares and the corresponding degrees of freedom (look at column 3 of the above ANOVA table.

The final table shows computation of the test statistic which can be computed as follows:

$$F_{cal} = \frac{MSR}{MSE} \approx F_{\alpha}(k-1, n-k)$$ [The F statistic follows F distribution]

The test rule: Reject H_0 *if* $F_{cal} \geq F_{\alpha}(k-1, n-k)$ where $F_{\alpha}(k-1, n-k)$ is the value to be read from the F- distribution table at a given α level.

Relationship between F and R^2

You may recall that R^2 is given by $R^2 = \frac{\sum \hat{y}^2}{\sum y^2}$ and $\sum \hat{y}^2 = R^2 \sum y^2$

We also know that

$$R^2 = 1 - \frac{\sum e^2}{\sum y^2} \quad \text{Hence,} \quad \frac{\sum e^2}{\sum y^2} = 1 - R^2 \quad \text{which means} \quad \sum e^2 = (1 - R^2)\sum y^2$$

The formula for F is:

$$F_{cal} = \frac{\frac{\sum \hat{y}^2}{k-1}}{\frac{\sum e^2}{n-k}}$$

$$F_{cal} = \frac{\frac{R^2 \sum y^2}{k-1}}{\frac{(1-R^2)\sum y^2}{n-k}} = \frac{R^2 \sum y^2}{k-1} \cdot \frac{(n-k)}{(1-R^2)\sum y^2}$$

$$F_{cal} = \frac{(n-k)}{k-1} \cdot \frac{R^2}{(1-R^2)}$$

That means the calculated F can also be expressed in terms of the coefficient of determination.

THE MATRIX APPROACH TO LINEAR REGRESSION MODEL

The k-variable linear regression model: if we generalize the two- and three-variable linear regression models, the k-variable population regression model (prf) involving the dependent variable y and k−1 explanatory variables x2, x3, . . . , xk may be written as

$$\text{PRF: } Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i \qquad i = 1, 2, 3, \ldots, n$$

where β1 =the intercept, β2 to βk =partial slope coefficients, u=stochastic disturbance term, and i =ith observation, n being the size of the population. The PRF (C.1.1) is to be interpreted in the usual manner: It gives the mean or expected value of Y conditional upon the fixed (in repeated sampling) values of X2, X3, ..., Xk, that is, E(Y|X2i, X3i, ..., Xki).

shorthand expression of the above equation can be written as the following set of n simultaneous equations:

$$
\begin{aligned}
Y_1 &= \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + \cdots + \beta_k X_{k1} + u_1 \\
Y_2 &= \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + \cdots + \beta_k X_{k2} + u_2 \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
Y_n &= \beta_1 + \beta_2 X_{2n} + \beta_3 X_{3n} + \cdots + \beta_k X_{kn} + u_n
\end{aligned}
$$

Let us write the system of the above equations in an alternative but more illuminating way as follows

$$
\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ 1 & X_{22} & X_{32} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2n} & X_{3n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}
$$

$$
\underset{n \times 1}{\mathbf{y}} = \underset{n \times k}{\mathbf{X}} \quad \underset{k \times 1}{\boldsymbol{\beta}} + \underset{n \times 1}{\mathbf{u}}
$$

where $\mathbf{y} = n \times 1$ column vector of observations on the dependent variable Y
$\mathbf{X} = n \times k$ matrix giving n observations on $k-1$ variables X_2 to X_k, the first column of 1's representing the intercept term (this matrix is also known as the **data matrix**)
$\boldsymbol{\beta} = k \times 1$ column vector of the unknown parameters $\beta_1, \beta_2, \ldots, \beta_k$
$\mathbf{u} = n \times 1$ column vector of n disturbances u_i

Using the rules of matrix multiplication and addition, the above System that is known as the matrix representation of the general (k-variable) linear regression model. It can be written more compactly as

$$
\underset{n \times 1}{\mathbf{y}} = \underset{n \times k}{\mathbf{X}} \quad \underset{k \times 1}{\boldsymbol{\beta}} + \underset{n \times 1}{\mathbf{u}}
$$

Where there is no confusion about the dimensions or orders of the matrix X and the vectors y, β, and u, Eq. may be written simply as

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

OLS estimation

To obtain the OLS estimate of β, let us first write the k-variable sample regression (SRF):

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_k X_{ki} + \hat{u}_i$$

which can be written more compactly in matrix notation as

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

and in matrix form as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ 1 & X_{22} & X_{32} & \cdots & X_{k2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & X_{2n} & X_{3n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} + \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{bmatrix}$$

$$\underset{n \times 1}{\mathbf{y}} = \underset{n \times k}{\mathbf{X}} \quad \underset{k \times 1}{\hat{\boldsymbol{\beta}}} + \underset{n \times 1}{\hat{\mathbf{u}}}$$

where ˆ β is a k-element column vector of the OLS estimators of the regression coefficients and where ˆ u is an n×1 column vector of n residuals. As in the two- and three-variable models, in the k-variable case the OLS estimators are obtained by minimizing

$$\sum \hat{u}_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_k X_{ki})^2$$

where $\sum \hat{u}_i^2$ is the residual sum of squares (RSS). In matrix notation, this amounts to minimizing $\hat{\mathbf{u}}'\hat{\mathbf{u}}$ since

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = [\hat{u}_1 \quad \hat{u}_2 \quad \cdots \quad \hat{u}_n] \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{bmatrix} = \hat{u}_1^2 + \hat{u}_2^2 + \cdots + \hat{u}_n^2 = \sum \hat{u}_i^2$$

$$\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

Therefore,

$$\begin{aligned}\hat{\mathbf{u}}'\hat{\mathbf{u}} &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}\end{aligned}$$

where use is made of the properties of the transpose of a matrix, namely, $(\mathbf{X}\hat{\boldsymbol{\beta}})' = \hat{\boldsymbol{\beta}}'\mathbf{X}'$; and since $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$ is a scalar (a real number), it is equal to its transpose $\mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}}$.

In scalar notation, the method of OLS consists in so estimating β1, β2, ..., βk thatˆ u^2 i is as small as possible. This is done by differentiating the equation partially with respect to ˆ β1, ˆ β2, ..., ˆ βk and setting the resulting expressions to zero. This process yields k simultaneous equations in k unknowns, the normal equations of the least-squares theory. As shown in the following:

$$\begin{aligned}
n\hat{\beta}_1 + \hat{\beta}_2 \sum X_{2i} + \hat{\beta}_3 \sum X_{3i} + \cdots + \hat{\beta}_k \sum X_{ki} &= \sum Y_i \\
\hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 + \hat{\beta}_3 \sum X_{2i} X_{3i} + \cdots + \hat{\beta}_k \sum X_{2i} X_{ki} &= \sum X_{2i} Y_i \\
\hat{\beta}_1 \sum X_{3i} + \hat{\beta}_2 \sum X_{3i} X_{2i} + \hat{\beta}_3 \sum X_{3i}^2 + \cdots + \hat{\beta}_k \sum X_{3i} X_{ki} &= \sum X_{3i} Y_i \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\
\hat{\beta}_1 \sum X_{ki} + \hat{\beta}_2 \sum X_{ki} X_{2i} + \hat{\beta}_3 \sum X_{ki} X_{3i} + \cdots + \hat{\beta}_k \sum X_{ki}^2 &= \sum X_{ki} Y_i
\end{aligned}$$

In matrix form

$$\underbrace{\begin{bmatrix} n & \sum X_{2i} & \sum X_{3i} & \cdots & \sum X_{ki} \\ \sum X_{2i} & \sum X_{2i}^2 & \sum X_{2i} X_{3i} & \cdots & \sum X_{2i} X_{ki} \\ \sum X_{3i} & \sum X_{3i} X_{2i} & \sum X_{3i}^2 & \cdots & \sum X_{3i} X_{ki} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum X_{ki} & \sum X_{ki} X_{2i} & \sum X_{ki} X_{3i} & \cdots & \sum X_{ki}^2 \end{bmatrix}}_{(\mathbf{X}'\mathbf{X})} \underbrace{\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}}_{\hat{\boldsymbol{\beta}}} = \underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{21} & X_{22} & \cdots & X_{2n} \\ X_{31} & X_{32} & \cdots & X_{3n} \\ \cdots & \cdots & \cdots & \cdots \\ X_{k1} & X_{k2} & \cdots & X_{kn} \end{bmatrix}}_{\mathbf{X}'} \underbrace{\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix}}_{\mathbf{y}}$$

or, more compactly, as

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

Features of the ($X^{'}X$) matrix

1. It gives the raw sums of squares and cross products of the X variables, one of which is the intercept term taking the value of 1 for each observation. The elements on the main diagonal give the raw sums of squares, and those off the main diagonal give the raw sums of cross products (by raw we mean in original units of measurement).

2. It is symmetrical since the cross product between X2i and X3i is the same as that between X3i and X2i.

3. It is of order (k×k), that is, k rows and k columns.

Now using matrix algebra, if the inverse of ($X^{'}X$) exists, say, $(X^{'}X)^{-1}$, then pre multiplying both sides of of regression equation by this inverse, we obtain

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

But since $(\mathbf{X'X})^{-1}(\mathbf{X'X}) = \mathbf{I}$, an identity matrix of order $k \times k$, we get

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

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

As an illustration of the matrix methods developed so far, let us rework the consumption–income, For the two-variable case we have

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$$

$$(\mathbf{X'X}) = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ X_1 & X_2 & X_3 & \cdots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \\ & \cdots \\ 1 & X_N \end{bmatrix} = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}$$

And

$$\mathbf{X'y} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ X_1 & X_2 & X_3 & \cdots & X_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix}$$

Example, if

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 10 & 1700 \\ 1700 & 322000 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1110 \\ 205500 \end{bmatrix}$$

find estimators

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.97576 & -0.005152 \\ -0.005152 & 0.0000303 \end{bmatrix}$$

Therefore,

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 0.97576 & -0.005152 \\ -0.005152 & 0.0000303 \end{bmatrix} \begin{bmatrix} 1110 \\ 205500 \end{bmatrix}$$

$$= \begin{bmatrix} 24.4545 \\ 0.5079 \end{bmatrix}$$

In the two- and three-variable linear regression models an unbiased estimator of σ^2 was given by $\hat{\sigma}^2 = \sum \hat{u}_i^2/(n-2)$ and $\hat{\sigma}^2 = \sum \hat{u}_i^2/(n-3)$, respectively. In the k-variable case, the corresponding formula is

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-k} = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k}$$

where there are now $n-k$ df. (Why?)

Although in principle $\hat{\mathbf{u}}'\hat{\mathbf{u}}$ can be computed from the estimated residuals, in practice it can be obtained directly as follows. Recalling that $\sum \hat{u}_i^2$ $(= \text{RSS}) = \text{TSS} - \text{ESS}$, in the two-variable case we may write

$$\sum \hat{u}_i^2 = \sum y_i^2 - \hat{\beta}_2^2 \sum x_i^2$$

and in the three-variable case

$$\sum \hat{u}_i^2 = \sum y_i^2 - \hat{\beta}_2 \sum y_i x_{2i} - \hat{\beta}_3 \sum y_i x_{3i}$$

By extending this principle, it can be seen that for the k-variable model

$$\sum \hat{u}_i^2 = \sum y_i^2 - \hat{\beta}_2 \sum y_i x_{2i} - \cdots - \hat{\beta}_k \sum y_i x_{ki}$$

In matrix notation,

$$\text{TSS: } \sum y_i^2 = \mathbf{y}'\mathbf{y} - n\bar{Y}^2$$

$$\text{ESS: } \hat{\beta}_2 \sum y_i x_{2i} + \cdots + \hat{\beta}_k \sum y_i x_{ki} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\bar{Y}^2$$

where the term $n\bar{Y}^2$ is known as the correction for mean.[6] Therefore,

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$$

The coefficient of determination R^2 in matrix notation

The coefficient of determination R^2 has been defined as

$$R^2 = \frac{\text{ESS}}{\text{TSS}}$$

In the two-variable case,

$$R^2 = \frac{\hat{\beta}_2^2 \sum x_i^2}{\sum y_i^2}$$

and in the three-variable case

$$R^2 = \frac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}}{\sum y_i^2}$$

Generalizing we obtain for the k-variable case

$$R^2 = \frac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i} + \cdots + \hat{\beta}_k \sum y_i x_{ki}}{\sum y_i^2}$$

$$R^2 = \frac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\bar{Y}^2}{\mathbf{y}'\mathbf{y} - n\bar{Y}^2}$$

Find R^2 based on the first given example,

$$\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = [24.3571 \quad 0.5079]\begin{bmatrix} 1{,}110 \\ 205{,}500 \end{bmatrix}$$
$$= 131{,}409.831$$
$$\mathbf{y}'\mathbf{y} = 132{,}100$$

$$n\bar{Y}^2 = 123{,}210$$

R^2= 0.9924

Illustration: The following table shows a particular country's the value of imports (Y), the level of Gross National Product(X_1) measured in arbitrary units, and the price index of imported goods (X_2), over 12 years period.

Table 5: Data for multiple regression examples

Year	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971
Y	57	43	73	37	64	48	56	50	39	43	69	60
X_1	220	215	250	241	305	258	354	321	370	375	385	385
X_2	125	147	118	160	128	149	145	150	140	115	155	152

a) Estimate the coefficients of the economic relationship and fit the model.

To estimate the coefficients of the economic relationship, we compute the entries given in the following table

Year	Y	X_1	X_2	x_1	x_2	Y	X_1^2	x_2^2	x_1y	x_2y	x_1x_2	y^2
1960	57	220	125	-86.5833	-15.3333	3.75	7496.668	235.1101	-324.687	-57.4999	1327.608	14.0625
1961	43	215	147	-91.5833	6.6667	-10.25	8387.501	44.44489	938.7288	-68.3337	-610.558	105.0625
1962	73	250	118	-56.5833	-22.3333	19.75	3201.67	498.7763	-1117.52	-441.083	1263.692	390.0625
1963	37	241	160	-65.5833	19.6667	-16.25	4301.169	386.7791	1065.729	-319.584	-1289.81	264.0625
1964	64	305	128	-1.5833	-12.3333	10.75	2.506839	152.1103	-17.0205	-132.583	19.52731	115.5625
1965	48	258	149	-48.5833	8.6667	-5.25	2360.337	75.11169	255.0623	-45.5002	-421.057	27.5625
1966	56	354	145	47.4167	4.6667	2.75	2248.343	21.77809	130.3959	12.83343	221.2795	7.5625
1967	50	321	150	14.4167	9.6667	-3.25	207.8412	93.44509	-46.8543	-31.4168	139.3619	10.5625
1968	39	370	140	63.4167	-0.3333	-14.25	4021.678	0.111089	-903.688	4.749525	-21.1368	203.0625
1969	43	375	115	68.4167	-25.3333	-10.25	4680.845	641.7761	-701.271	259.6663	-1733.22	105.0625
1970	69	385	155	78.4167	14.6667	15.75	6149.179	215.1121	1235.063	231.0005	1150.114	248.0625
1971	60	385	152	78.4167	11.6667	6.75	6149.179	136.1119	529.3127	78.75022	914.8641	45.5625
Sum	639	3679	1684	0.0004	0.0004	0	49206.92	2500.667	1043.25	-509	960.6667	1536.25
Mean	**53.25**	**306.5833**	**140.3333**	**0**	**0**	**0**						

From the above table, we can take the following summary results.

$\sum Y = 639$ $\sum X_1 = 3679$ $\sum X_2 = 1684$ $n = 12$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{639}{12} = 53.25$$

$$\bar{X}_1 = \frac{\sum X_1}{n} = \frac{3679}{12} = 306.5833$$

$$\bar{X}_2 = \frac{\sum X_2}{n} = \frac{1684}{12} = 140.3333$$

The summary results in deviation forms are then given by:

$\sum x_1^2 = 4920.692$ $\sum x_2^2 = 2500.667$

$\sum x_1 y = 1043.25$ $\sum x_2 y = -509$

$\sum x_1 x_2 = 960.6667$ $\sum y^2 = 1536.25$

The coefficients are then obtained as follows.

$$\hat{\beta}_1 = \frac{(\sum x_1 y_1)(\sum x_2^2) - (\sum x_2 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} = \frac{(1043.25)(2500.667) - (-509)(960.6667)}{(49206.92)(2500.667) - (960.667)^2} = \frac{2608821 + 488979.4}{123050121 - 922880.51}$$

$$= \frac{3097800.2}{122127241} = 0.025365$$

$$\hat{\beta}_2 = \frac{(\sum x_2 y)(\sum x_1^2) - (\sum x_1 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} = \frac{(-509)(49206.92) - (1043.25)(960.6667)}{(49206.92)(2500.667) - (960.667)^2} = \frac{-2504632 - 1002216}{123050121 - 922880.51}$$

$$= \frac{-26048538}{122127241} = -0.21329$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 = 53.25 - (0.025365) - (-0.21329) = 75.40512$$

The fitted model is then written as: $\hat{Y}_i = 75.40512 + 0.025365X_1 - 0.21329X_2$

b) Compute the variance and standard errors of the slopes.

First, you need to compute the estimate of the variance of the random term as follows

$$\hat{\delta}_u^2 = \frac{\sum e_i^2}{n-3} = \frac{1401.223}{12-3} = \frac{1401.223}{9} = 155.69143$$

Variance of $\hat{\beta}_1$

$$Var\left(\hat{\beta}_1\right) = \delta_u^2 \left[\frac{\sum x_2^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2}\right] = 155.69143\left(\frac{2500.667}{12212724}\right) = 0.003188$$

Standard error of $\hat{\beta}_1$

$$SE(\hat{\beta}_1) = \sqrt{Var(\hat{\beta}_1)} = \sqrt{0.003188} = 0.056462$$

Variance of $\hat{\beta}_2$

$$Var(\hat{\beta}_2) = \hat{\delta}_u^2 \left[\frac{\sum x_1^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2}\right] = 155.69143\left(\frac{49206.92}{122127241}\right) = 0.0627$$

Standard error of $\hat{\beta}_2$

$$SE(\hat{\beta}_2) = \sqrt{Var(\hat{\beta}_2)} = \sqrt{0.0627} = 0.25046$$

Similarly, the standard error of the intercept is found to be 37.98177. The detail is left for you as an exercise.

c) Calculate and interpret the coefficient of determination.

We can use the following summary results to obtain the R^2.

$$\sum \hat{y}^2 = 135.0262$$

$$\sum e^2 = 1401.223$$

$$\sum y^2 = 1536.25$$

(The sum of the above two). Then,

$$R^2 = \frac{\hat{\beta}_1 \sum x_1 y + \hat{\beta}_2 \sum x_2 y}{\sum y^2} = \frac{(0.025365)(1043.25) + (-0.21329)(-509)}{1356.25} = 0.087894$$

$$R^2 = 1 - \frac{\sum e^2}{\sum y^2} = 1 - \frac{1401.223}{1356.25} = 0.087894$$

or

d) Compute the adjusted R^2.

$$R^2_{adj} = 1 - (1 - R^2)\frac{(n-1}{n-k)} = 1 - (1 - 0.087894)\frac{12-1}{12-3} = -0.114796$$

*e) Construct 95% confidence interval for the true population parameters (partial regression coefficients).[*Exercise: Base your work on Simple Linear Regression*]*

f) Test the significance of X_1 and X_2 in determining the changes in Y using t-test.

The hypotheses are summarized in the following table.

Coefficient	Hypothesis	Estimate	Std. error	Calculated t	Conclusion
β_1	H_0: $\beta_1=0$ H_1: $\beta_1\neq0$	0.025365	0.056462	$t_{cal} = \frac{0.025365}{0.056462} = 0.449249$	We do not reject H_0 since $t_{cal}<t_{tab}$
β_2	H_0: $\beta_2=0$ H_1: $\beta_2\neq0$	-0.21329	0.25046	$t_{cal} = \frac{-0.21329}{-0.21329} = -0.85159$	We do not reject H_0 since $t_{cal}<t_{tab}$

The critical value ($t_{0.05, 9}$) to be used here is 2.262. Like the standard error test, the t- test revealed that both X_1 and X_2 are insignificant to determine the change in Y since the calculated t values are both less than the critical value.

Exercise: Test the significance of X_1 and X_2 in determining the changes in Y using the standard error test.

*g) Test the overall significance of the model. (Hint: u*se $\alpha = 0.05$)

This involves testing whether at least one of the two variables X_1 and X_2 determine the changes in Y. The hypothesis to be tested is given by:

$H_0 : \beta_1 = \beta_2 = 0$
$H_1 : \beta_i \neq 0$, at least for one i.

The ANOVA table for the test is give as follows:

Source of variation	Sum of Squares	Degrees of freedom	Mean Sum of Squares	F_{cal}

Regression	$SSR = \sum \hat{y}^2 = 135.0262$	$k-1$=3-1=2	$MSR = \frac{\sum \hat{y}^2}{k-1} = \frac{135.0262}{2} =$ 67.51309	$F = \frac{MSR}{MSE}$ $= 0.433634$
Residual	$SSE = \sum e^2 = 1401.223$	$n-k$ =12-3=9	$MSE = \frac{\sum e^2}{n-k} = \frac{1401.223}{9} =$ 155.614	
Total	$SST = \sum y^2 = 1536.25$	$n-1$ =12-1=11		

The tabulated F value (critical value) is $F_{\alpha}(2, 11) = 3.98$

In this case, the calculated F value (0.4336) is less than the tabulated value (3.98). Hence, we do not reject the null hypothesis and conclude that there is no significant contribution of the variables X_1 and X_2 to the changes in Y.

h) Compute the F value using the R^2.

$$F_{cal} = \frac{(n-k)}{k-1} \cdot \frac{R^2}{(1-R^2)} = \frac{(12-3)}{3-1} \frac{0.087894}{1-0.087894} = 0.433632$$

Example 2. per capita personal consumption expenditure (ppce) and per capita personal disposable income (ppdi) in the united states, 1956–1970, in 1958 dollars

PPCE, Y	PPDI, X_2	Time, X_3	PPCE, Y	PPDI, X_2	Time, X_3
1673	1839	1 (= 1956)	1948	2126	9
1688	1844	2	2048	2239	10
1666	1831	3	2128	2336	11
1735	1881	4	2165	2404	12
1749	1883	5	2257	2487	13
1756	1910	6	2316	2535	14
1815	1969	7	2324	2595	15 (= 1970)
1867	2016	8			

In matrix notation, our problem may be shown as follows:

$$\begin{bmatrix} 1673 \\ 1688 \\ 1666 \\ 1735 \\ 1749 \\ 1756 \\ 1815 \\ 1867 \\ 1948 \\ 2048 \\ 2128 \\ 2165 \\ 2257 \\ 2316 \\ 2324 \end{bmatrix} = \begin{bmatrix} 1 & 1839 & 1 \\ 1 & 1844 & 2 \\ 1 & 1831 & 3 \\ 1 & 1881 & 4 \\ 1 & 1883 & 5 \\ 1 & 1910 & 6 \\ 1 & 1969 & 7 \\ 1 & 2016 & 8 \\ 1 & 2126 & 9 \\ 1 & 2239 & 10 \\ 1 & 2336 & 11 \\ 1 & 2404 & 12 \\ 1 & 2487 & 13 \\ 1 & 2535 & 14 \\ 1 & 2595 & 15 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} + \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \\ \hat{u}_4 \\ \hat{u}_5 \\ \hat{u}_6 \\ \hat{u}_7 \\ \hat{u}_8 \\ \hat{u}_9 \\ \hat{u}_{10} \\ \hat{u}_{11} \\ \hat{u}_{12} \\ \hat{u}_{13} \\ \hat{u}_{14} \\ \hat{u}_{15} \end{bmatrix}$$

$$\begin{matrix} \mathbf{y} \\ 15 \times 1 \end{matrix} = \begin{matrix} \mathbf{X} \\ 15 \times 3 \end{matrix} \quad \begin{matrix} \hat{\boldsymbol{\beta}} \\ 3 \times 1 \end{matrix} + \begin{matrix} \hat{\mathbf{u}} \\ 15 \times 1 \end{matrix}$$

From the preceding data we obtain the following quantities:

$$\bar{Y} = 1942.333 \qquad \bar{X}_2 = 2126.333 \qquad \bar{X}_3 = 8.0$$

$$\sum (Y_i - \bar{Y})^2 = 830{,}121.333$$

$$\sum (X_{2i} - \bar{X}_2)^2 = 1{,}103{,}111.333 \qquad \sum (X_{3i} - \bar{X}_3)^2 = 280.0$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ X_{21} & X_{22} & X_{23} & \cdots & X_{2n} \\ X_{31} & X_{32} & X_{33} & \cdots & X_{3n} \end{bmatrix} \begin{bmatrix} 1 & X_{21} & X_{31} \\ 1 & X_{22} & X_{32} \\ 1 & X_{23} & X_{33} \\ \vdots & \vdots & \vdots \\ 1 & X_{2n} & X_{3n} \end{bmatrix}$$

$$= \begin{bmatrix} n & \sum X_{2i} & \sum X_{3i} \\ \sum X_{2i} & \sum X_{2i}^2 & \sum X_{2i}X_{3i} \\ \sum X_{3i} & \sum X_{2i}X_{3i} & \sum X_{3i}^2 \end{bmatrix}$$

$$= \begin{bmatrix} 15 & 31{,}895 & 120 \\ 31{,}895 & 68{,}922.513 & 272{,}144 \\ 120 & 272{,}144 & 1240 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 29,135 \\ 62,905,821 \\ 247,934 \end{bmatrix}$$

Using the rules of matrix inversion given , one can see that

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 37.232491 & -0.0225082 & 1.336707 \\ -0.0225082 & 0.0000137 & -0.0008319 \\ 1.336707 & -0.0008319 & 0.054034 \end{bmatrix}$$

Therefore,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} 300.28625 \\ 0.74198 \\ 8.04356 \end{bmatrix}$$

The residual sum of squares can now be computed as

$$\begin{aligned} \sum \hat{u}_i^2 &= \hat{\mathbf{u}}'\hat{\mathbf{u}} \\ &= \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \\ &= 57,420,003 - [300.28625 \quad 0.74198 \quad 8.04356] \begin{bmatrix} 29,135 \\ 62,905,821 \\ 247,934 \end{bmatrix} \\ &= 1976.85574 \end{aligned}$$

whence we obtain

$$\hat{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{12} = 164.73797$$

The variance–covariance matrix for $\hat{\boldsymbol{\beta}}$ can therefore be shown as

$$\text{var-cov}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 6133.650 & -3.70794 & 220.20634 \\ -3.70794 & 0.00226 & -0.13705 \\ 220.20634 & -0.13705 & 8.90155 \end{bmatrix}$$

The diagonal elements of this matrix give the variances of ˆ β1, ˆ β2, and ˆ β3, respectively, and their positive square roots give the corresponding standard errors.

From the previous data, it can be readily verified that

$$\text{ESS: } \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\bar{Y}^2 = 828{,}144.47786$$

$$\text{TSS: } \mathbf{y}'\mathbf{y} - n\bar{Y}^2 = 830{,}121.333$$

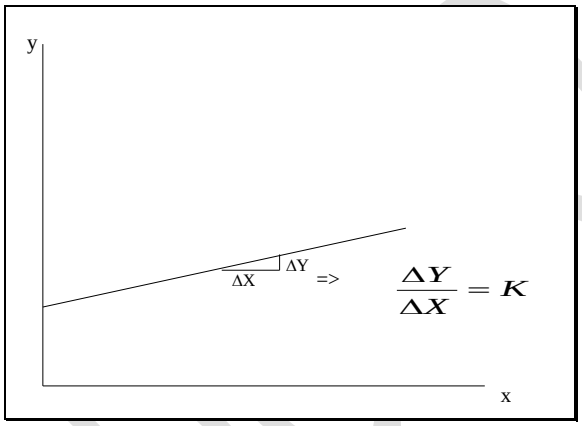
Therefore,

$$R^2 = \frac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\bar{Y}^2}{\mathbf{y}'\mathbf{y} - n\bar{Y}^2} = \frac{828{,}144.47786}{830{,}121.333} = 0.99761$$

Some Commonly Used Functional Forms in econometrics

a) ***The Linear-linear Form*:** It is based on the assumption that the slope of the relationship between the independent variable and the dependent variable is constant.

$\frac{\partial Y}{\partial X} = \beta_i$ i=1,2,...K



Suppose the dependent variable and independent variables of interest are both in linear (level) form. For example, suppose we want to know the effect of an additional year of schooling on wages and from available data we fit the following equation

$$W_i = \alpha + \beta_1 Ed_i + \beta_2 Age_i + \beta_3 Exp_i + \epsilon_i$$

In this case, the coefficient on the variable of interest can be interpreted as the marginal effect. The marginal effect is how the dependent variable changes when the independent variable changes by an additional unit holding all other variables in the equation constant (i.e. partial derivative) or

$$\frac{\partial W_i}{\partial Ed_i} = \beta_1$$

Therefore, βj can be interpreted as the change in wages from a one unit increase (or state change if dummy variable) of Xj holding all other independent variables constant.

Example: Suppose the fitted equation for (1) is

$$\hat{W}_i = 3.5 + 0.75Ed_i + 0.25Age_i + 0.30Exp_i + \epsilon_i$$

Based on the data used in this regression, an additional year of education corresponds to an increase in hourly wages of $0.75. Similarly, an additional year of experience is associated with a $0.30 per hour wage increase.

In this case elasticity is not constant.

$$N_{Y,X} = \frac{\partial Y / Y}{\partial X / X} = \frac{\partial Y}{\partial X} \cdot \frac{X}{Y} = \beta_i \frac{X}{Y}$$

If the hypothesized relationship between Y and X is such that the slope of the relationship can be expected to be constant and the elasticity can therefore be expected to be variable, then the linear functional form should be used.

Note: Economic theory frequently predicts only the sign of a relationship and not its functional form. Under such circumstances, the linear form can be used until strong evidence that it is inappropriate is found. Thus, unless theory, common sense, or experience justifies using some other functional form, one should use the linear model.

b) Log-log

Consider interpreting coefficients from a regression where the dependent and independent variable of interest are in log form. The coefficients can no longer be interpreted as marginal effects.

Suppose economic theory suggests estimation of our wage equation with the dependent variable in log form and inclusion of community volunteer hours per week (Comm) also in log form. The equation of interest is now

$$\log(W_i) = \alpha + \beta_1 Ed_i + \beta_2 Age_i + \beta_3 Exp_i + \beta_4 \log(Comm_i) + \epsilon_i$$

We would like to interpret the coefficient on the community volunteer variable (β4). To better understand the interpretation, consider taking the differential of (2) holding all independent variables constant except Commi.

$$d[\log(W_i)] = d[\log(Comm_i)]\beta_4$$
$$\frac{1}{W_i} dW_i = \frac{1}{Comm_i} dComm_i \beta_4$$

since $d[log(X)] = \frac{1}{X} d(X)$. This final equation can be rearranged such that,

$$\frac{\frac{100 * dW_i}{W_i}}{\frac{100 * dComm_i}{Comm_i}} = \beta_4$$

where the left hand side is the (partial) elasticity of W with respect to Comm. Elasticity is the ratio of the percent change in one variable to the percent change in another variable. The coefficient in a regression is a partial elasticity since all other variables in the equation are held constant. Therefore, β4 can be interpreted as the percent change in hourly wages from a one percent increase in community volunteer hours per week holding education, age and experience constant.

Example Suppose that the fitted equation is

$$\widehat{\log(W_i)} = 3.26 + 0.24 Ed_i + 0.08 Age_i + 0.16 Exp_i + 1.2 \log(Comm_i)$$

Based on these regression results, a one percent increase in community volunteer hours per week is associated with a 1.2% increase in hourly wages.

c) Log-linear Form

The log-linear functional form is a variant of the log-linear equation in which the dependent variables are expressed in terms of their logs. Such models expressed as: $\ln Y_i = \beta_0 + \beta_1 X_{1i} + U_i$

In equation (2), education, age and experience are in level terms while the dependent variable (wage) is in log terms. We would like to interpret the coefficients on these variables. First, consider education. Take the differential holding all other independent variables constant.

$$d[\log W_i] = dEd_i\beta_1$$
$$\frac{dW_i}{W_i} = dEd_i\beta_1$$

Multiply both sides by 100 and rearrange,

$$\frac{100 * dW_i}{W_i} = 100 * dEd_i\beta_1$$
$$100 * \beta_1 = \frac{\frac{100*dW_i}{W_i}}{dEd_i} = \frac{\%\Delta W_i}{\text{unit } \Delta Ed_i}$$

Therefore, 100∗β1 can be interpreted as the percentage change in Wi for a unit increase in Edi, holding all other independent variables constant. Similar derivations can derive the interpretation for the coefficients on age and experience.

Example: Consider the fitted equation

$$\widehat{\log(W_i)} = 3.26 + 0.24Ed_i + 0.08Age_i + 0.16Exp_i + 1.2\log(Comm_i)$$

Therefore, holding all other independent variables constant, an additional year of schooling is associated with a 24% increase in hourly wages. Similarly, an additional year of experience is associated with a 16% increase in hourly wages

d) Linear-log Form

The linear-log functional form is a variant of the linear-log equation in which the independent variables are expressed in terms of their logs. Such models expressed as: $Y_i = \beta_0 + \beta_1 \ln X_{1i} + U_i$

Consider a regression where the dependent variable is in linear or level terms and the independent variable of interest is in log terms. For example, consider the following equation

$$W_i = \alpha + \beta_1 Ed_i + \beta_2 Age_i + \beta_3 Exp_i + \beta_4 \log(Comm_i) + \epsilon_i$$

Recall from the section on level-level regressions that the coefficients on education, age and experience can be interpreted as marginal effects. We would like to interpret the coefficient on community volunteer hours (β4). Again, take the differential on both sides, holding all independent variables constant except community volunteer hours:

$$dW_i = d[log(Comm_i)]\beta_4$$
$$dW_i = \frac{1}{Comm_i} dComm_i \beta_4$$

Divide both sides by 100 and rearrange

$$\frac{\beta_4}{100} = \frac{dW_i}{\frac{100 * dComm_i}{Comm_i}} = \frac{\text{unit } \Delta W_i}{\%\Delta Comm_i}$$

Therefore, β4 100 can be interpreted as the increase in hourly wages from a one percent increase in community volunteer hours per week.

Example: Suppose that the fitted equation is

$$\hat{W}_i = 3 + 0.67 Ed_i + 0.28 Age_i + 0.34 Exp_i + 13.2 \log(Comm_i)$$

Therefore, holding education, age and experience constant, a one percent increase in community volunteer hours per week is associated with a $0.132 increase in hourly wages.

Generally interpreted as the following table

Model	If x increases by	then y will increase by
linear	1 unit	$\hat{\beta}_2$ units
linear-log	1%	$(\hat{\beta}_2/100)$ units
log-linear	1 unit	$(100\hat{\beta}_2)\%$
log-log	1%	$\hat{\beta}_2\%$

Summary of functional forms

MODEL	FORM	SLOPE	ELASTICITY
		$(\frac{dY}{dX})$	$\frac{dY}{dX}.\frac{X}{Y}$
Linear	$Y = B_1 + B_2 X$	B_2	$B_2(\frac{X}{Y})$
Log-linear	$\ln Y = B_1 + \ln X$	$B_2(\frac{Y}{X})$	B_2
Log-lin	$\ln Y = B_1 + B_2 X$	$B_2(Y)$	$B_2(X)$
Lin-log	$Y = B_1 + B_2 \ln X$	$B_2(\frac{1}{X})$	$B_2(\frac{1}{Y})$
Reciprocal	$Y = B_1 + B_2(\frac{1}{X})$	$-B_2(\frac{1}{X^2})$	$-B_2(\frac{1}{XY})$

UNIT FIVE: DUMMY VARIABLE REGRESSION MODELS

There are four basic types of variables we generally encounter in empirical analysis. These are: nominal, ordinal, interval and ratio scale variables. In preceding sections, we have encountered ratio scale variables. However, regression models do not deal only with ratio scale variables; they can also involve nominal and ordinal scale variables. In regression analysis, the dependent variable can be influenced by nominal variables such as sex, race, color, geographical region etc. models where all regressors are nominal (categorical) variables are called **ANOVA** (Analysis of Variance) models. If there is mixture of nominal and ratio scale variables, the models are called **ANCOVA** (Analysis of Covariance) models.

ANOVA MODEL

Models where all regressors are nominal (categorical) variables are called **ANOVA** (Analysis of Variance) models.

ANOVA models are used to assess the statistical significance of the relationship between a quantitative regressand and qualitative or dummy regressors. They are often used to compare the differences in the mean values of two or more groups or categories, and are therefore more general than the t test which can be used to compare the means of two groups or categories only.

Example: $Y_i = \alpha + \beta D_i + u_i$ ----------

where Y=annual salary of a college professor

$D_i = 1$ if male college professor

= 0 otherwise (i.e., female professor)

The two variable regression models encountered previously except that instead of a quantitative X variable we have a dummy variable D

Model may enable us to find out whether sex makes any difference in a college professor's salary, assuming, of course, that all other variables such as age, degree attained, and years of experience are held constant. Assuming that the disturbance satisfy the usually assumptions of the classical linear regression model, we obtain from the model.

Mean salary of female college professor: $E(Y_i / D_i = 0) = \alpha$ ----

Mean salary of male college professor: $E(Y_i / D_i = 1) = \alpha + \beta$

that is, the intercept term α gives the mean salary of female college professors and the slope coefficient β tells by how much the mean salary of a male college professor differs from the mean salary of his female counterpart, β +α reflecting the mean salary of the male college professor.

A test of the null hypothesis that there is no sex discrimination(H0: β=0) can be easily made by running regression model in the usual manner and finding out whether on the basis of the t test the estimated β is statistically significant

Illustration: The following model represents the relationship between geographical location and teachers' average salary in public schools. The data were taken from 50 states for a single year. The 50 states were classified into three regions: Northeast, South and West. The regression models looks like the following.

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + u_i$$

Where Y_i = the (average) salary of public school teachers in state i

D_{1i} = 1 if the state is in the Northeast

= 0 otherwise (i.e. in other regions of the country)

D_{2i} = 1 if the state is in the South

= 0 otherwise (i.e. in other regions of the country)

What does the model tell us? Assuming that the error term satisfies the usual OLS assumptions, on taking expectation on both sides, we obtain

Mean salary of public school teachers in the Northeast and North Central

$$E(Y_i \mid D_{2i} = 1, D_{3i} = 0) = \beta_1 + \beta_2$$

Mean salary of public school teachers in the South:

$$E(Y_i \mid D_{2i} = 0, D_{3i} = 1) = \beta_1 + \beta_3$$

You might wonder how we find out the mean salary of teachers in the West. If you guessed that this is equal to β1, you would be absolutely right, for

Mean salary of public school teachers in the West:

$$E(Y_i \mid D_{2i} = 0, D_{3i} = 0) = \beta_1$$

Example of the following summarized data.

$$\hat{Y}_i = 26{,}158.62 - 1734.47D_{1i} - 3264.62\beta_2 D_{2ii}$$

$$Se = (1128.52) \quad (1435.95) \quad (1499.62)$$

$$t = \quad (23.18) \quad (-1.21) \quad (-2.18)$$

$$p - value \quad (0.000) \quad (0.233) \quad (0.0349) \quad R^2 = 0.0901$$

From the above fitted model or regression analysis, we can see that mean salary of public school teachers in the West is about $26,158.62. The mean salary of teachers in the Northeast is lower by $1734.47 than those of the West and those in the South are lower by $3264.42. The actual mean salaries in the last two regions can be easily obtained by adding these differential salaries to the mean salary of teachers in the West. Doing this, we will find the average salaries in the latter two regions are about $24,424 and $22,894.

But how do we know that these mean salaries are statistically different from the mean salary of teachers in the West, the comparison category?

In order to know that these mean salaries are statistically different from the mean salary of teachers in the West, we have to do is to find out if each of the"slope"coefficients is statistically significant.

From this regression, the estimated slope coefficient for Northeast and North Central **is not statistically** significant as its p value is 23 percent, where as that of the South is **statistically significant,** as the p value is only about 3.5 percent. Therefore, the overall conclusion is that

statistically the mean salaries of public school teachers in the West and the Northeast and North Central are about the same but the mean salary of teachers in the South is statistically significantly lower by about $3265.

Before proceeding further, note the following features of the dummy variable regression model considered previously.

1. To distinguish the two categories, male and female, we have introduced only one dummy variable Di. For if Di= 1 denotes a male, when Di=0 we know that it is a female since there are only two possible outcomes. Hence, one dummy variable suffices to distinguish two categories. The general rule is this: If a qualitative variable has 'm' categories, introduce only 'm-1' dummy variables. In our example, sex has two categories, and hence we introduced only a single dummy variable. If this rule is not followed, we shall fall into what might be called the **dummy variable trap**, that is, the situation of perfect multicollinearity.

2. The assignment of 1 and 0 values to two categories, such as male and female, is arbitrary in the sense that in our example we could have assigned D=1 for female and D=0 for male.

3. The group, category, or classification that is assigned the value of 0 is often referred to as the base, benchmark, control, comparison, reference, or omitted category. It is the base in the sense that comparisons are made with that category.

4. The coefficient α_2 attached to the dummy variable **D** can be called the ***differential intercept coefficien***t because it tells by how much the value of the intercept term of the category that receives the value of 1 differs from the intercept coefficient of the base category.

Anova models with two qualitative variables

Example: hourly wages in relation to marital status and region of residence

$$\hat{Y}_i = 8.8148 + 1.0997D_{2i} - 1.6729D_{3i}$$
$$\text{se} = (0.4015) \quad (0.4642) \quad (0.4854)$$
$$t = (21.9528) \quad (2.3688) \quad (-3.4462)$$
$$(0.0000)^* \quad (0.0182)^* \quad (0.0006)^*$$
$$R^2 = 0.0322$$

where Y = hourly wage ($)
D_2 = married status, 1 = married, 0 = otherwise
D_3 = region of residence; 1 = South, 0 = otherwise

* denotes the p values. In this example we have two qualitative regressors, each with two categories. Hence we have assigned a single dummy variable for each category.

Which is the benchmark category here? it is unmarried, non-South residence. In other words, unmarried persons who do not live in the South are the omitted category. Therefore, all comparisons are made in relation to this group. The mean hourly wage in this benchmark is about $8.81. Compared with this, the average hourly wage of those who are married is higher by about $1.10, for an **actual average** wage of $9.91 (=8.81+1.10). By contrast, for those who live in the South, the average hourly wage is lower by about $1.67, for an actual average hourly wage of $7.14.

Are the preceding average hourly wages statistically different compared to the base category? They are, for all the differential intercepts are statistically significant, as their p values are quite low.

ANCOVA MODEL

Regression models containing a mixture of quantitative and qualitative variables are called analysis of covariance (ANCOVA) models. ANCOVA models are an extension of the ANOVA models in that they provide a method of statistically controlling the effects of quantitative regressors, called covariates or control variables, in a model that includes both quantitative and qualitative, or dummy, regressors.

Ancova model with one quantitative variable and one qualitative variable with two classes, or categories

Consider the model

$$Y_i = \alpha_i + \alpha_2 D_i + \beta X_i + u_i.$$

Where: Y_i = annual salary of a college professor

X_i = years of teaching experience

D_i = 1 if male

=0 otherwise

Model contains one quantitative variable (years of teaching experience) and one qualitative variable (sex) that has two classes (or levels, classifications, or categories), namely, male and female. What is the meaning of this equation?

Assuming that $E(u_i) = 0$,

Mean salary of female college professor: $E(Y_i / X_i, D_i = 0) = \alpha_1 + \beta X_i$----

Mean salary of male college professor: $E(Y_i / X_i, D_i = 1) = (\alpha + \alpha_2) + \beta X_i$.

Geometrically, we have the situation shown in fig (for illustration, it is assumed that $0 > \alpha_1$). In words, model postulates that the male and female college professors' salary functions in relation to the ***years of teaching experience*** have the ***same slope*** β but ***different intercepts***. In other words, it is assumed that the **level** of the male professor's mean salary is different from that of the female professor's mean salary $\boldsymbol{\alpha_2}$ but the **rate of change** in the mean annual salary by years of experience is the same for both sexes.

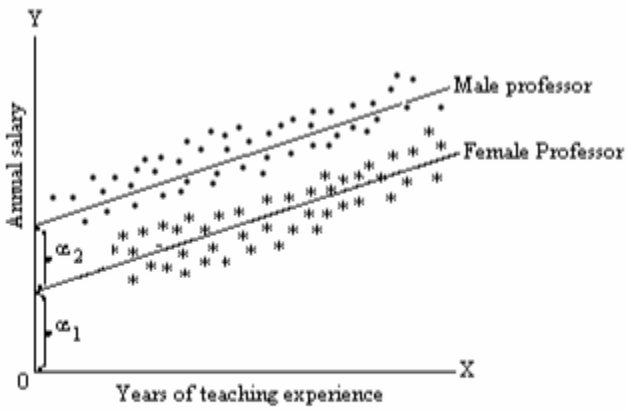


Figure 5.1 Hypothetical scattergram between annual salary and years of teaching experience of college professors.

Ancova model with one quantitative variable and one qualitative variable with more than two classes

Suppose that, on the basis of the cross-sectional data, we want to regress the annual expenditure on health care by an individual on the income and education of the individual. Since the variable education is qualitative in nature, suppose we consider three mutually exclusive levels of education: less than high school, high school, and college. Therefore, following the rule that the number of dummies be one less than the number of categories of the variable, we should introduce two dummies to take care of the three levels of education. Assuming that the three educational groups have a common slope but different intercepts in the regression of annual expenditure on health care on annual income,

We can use the following model:

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta X_i + u_i \quad ---$$

Where $Y_i =$ annual expenditure on health care

$X_i =$ annual expenditure

$D_2 = 1$ if high school education

$= 0$ otherwise

$D_3 = 1$ if college education

$= 0$ otherwise

Note that in the preceding assignment of the dummy variables we are arbitrarily treating the "less than high school education" category as the base category. Therefore, the intercept **α1** will reflect the intercept for this category. The differential intercepts **α2** and **α3** tell by how much the intercepts of the other two categories differ from the intercept of the base category, which can be readily checked as follows: Assuming $E(u_i) = 0$

$$E(Y_i \mid D_2 = 0,\ D_3 = 0, X_i) = \alpha_1 + \beta X_i$$

$$E(Y_i \mid D_2 = 1,\ D_3 = 0, X_i) = (\alpha_1 + \alpha_2) + \beta X_i$$

$$E(Y_i \mid D_2 = 0,\ D_3 = 1, X_i) = (\alpha_1 + \alpha_3) + \beta X_i$$

which are, respectively the mean health care expenditure functions for the three **levels** of education, namely, less than high school, high school, and college.

Geometrically, the situation is shown in fig (for illustrative purposes it is assumed that **α3 > α2**)

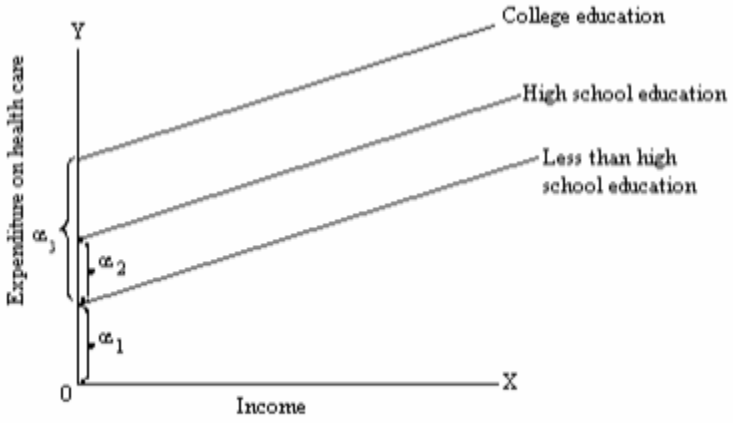


Figure 5.2 Expenditure on health care in realtion to income for three levels of education

ANCOVA model with one quantitative variable and two qualitative variables

The technique of dummy variable can be easily extended to handle more than one qualitative variable. Let us take college professors' salary regression on years of teaching experience and sex the skin color of the teacher. For simplicity, assume that color has two categories: black and white. We can now write

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta X_i + u_i$$

Where Yi = annual salary

Xi = years of teaching experience

D2 = 1 if male

=0 otherwise

D3 = 1 if white

= 0 otherwise

Notice that each of the two qualitative variables, sex and color, has two categories and hence needs one dummy variable for each. Note also that **the omitted**, or base, category now is ***"black female professor***."

Mean salary for black female professor:

$$E(Y_i \mid D_2 = 0, D_3 = 0, X_i) = \alpha_1 + \beta X_i$$

Mean salary for black male professor:

$$E(Y_i \mid D_2 = 1, D_3 = 0, X_i) = (\alpha_1 + \alpha_2) + \beta X_i$$

Mean salary for white female professor:

$$E(Y_i \mid D_2 = 0, D_3 = 1, X_i) = (\alpha_1 + \alpha_3) + \beta X_i$$

Mean salary for white male professor:

$$E(Y_i \mid D_2 = 1, D_3 = 1, X_i) = (\alpha_1 + \alpha_2 + \alpha_3) + \beta X_i$$

Once again, it is assumed that the preceding regressions differ only in the intercept coefficient but not in the slope coefficien**t β** .

OLS estimation will enable us to test a variety of hypotheses. Thus, if α_3 is statistically significant, it will mean that color does affect a professor's salary. Similarly, if α_2 is statistically significant, it will mean that sex also affects a professor's salary. If both these ***differential intercepts*** are ***statistically significant***, it would mean sex as well as color is an important determinant of professors' salaries

Let us reconsider Example of ANOVA by maintaining that the average salary of public school teachers may not be different in the three regions if we take into account any variables that cannot be standardized across the regions.

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 X_i + u_i$$

where Y_i = average annual salary of public school teachers in state ($)
X_i = spending on public school per pupil ($)
D_{2i} = 1, if the state is in the Northeast or North Central
= 0, otherwise
D_{3i} = 1, if the state is in the South
= 0, otherwise

Example: teacher's salary in relation to region and spending on public school per pupil

$$\hat{Y}_i = 13{,}269.11 - 1673.514D_{2i} - 1144.157D_{3i} + 3.2889X_i$$
$$\text{se} = (1395.056) \quad (801.1703) \quad (861.1182) \quad (0.3176)$$
$$t = (9.5115)^{*} \quad (-2.0889)^{*} \quad (-1.3286)^{**} \quad (10.3539)^{*}$$
$$R^2 = 0.7266$$

where* indicates p values less than 5 percent, and ** indicates p values greater than 5 percent. As these results suggest, ceteris paribus: as public expenditure goes up by a dollar, on average, a public school teacher's salary goes up by about $3.29. Controlling for spending on education, we now see that the differential intercept coefficient is significant for the Northeast and North-Central region, but not for the South.

UNIT SIX: ECONOMETRIC PROBLEMS

6.1. Assumptions Revisited

In many practical cases, two major problems arise in applying the classical linear regression model.

1) those due to assumptions about the specification of the model and about the disturbances and
2) those due to assumptions about the data

The following assumptions fall in either of the categories.

- The regression model is linear in parameters.
- The values of the explanatory variables are fixed in repeated sampling (non-stochastic).
- The mean of the disturbance (u_i) is zero for any given value of X i.e. $E(u_i) = 0$
- The variance of u_i is constant i.e. homoscedastic
- There is no autocorrelation in the disturbance terms
- The explanatory variables are independently distributed with the u_i.
- The number of observations must be greater than the number of explanatory variables.
- There must be sufficient variability in the values taken by the explanatory variables.
- There is no linear relationship (multicollinearity) among the explanatory variables.
- The stochastic (disturbance) term u_i are normally distributed i.e., $u_i \sim N(0, \sigma^2)$
- The regression model is correctly specified i.e., no specification error.

With these assumptions we can show that OLS are BLUE, and normally distributed. Hence it was possible to test Hypothesis about the parameters. However, if any of such assumption is relaxed, the OLS might not work.

6.2. Violations of Assumptions

The Zero Mean Assumption i.e. E(ui)=0

If this assumption is violated, we obtain a biased estimate of the intercept term. But, since the intercept term is not very important we can leave it. The slope coefficients remain unaffected even if the assumption is violated. The intercept term does not also have physical interpretation.

The Normality Assumption

This assumption is not very essential if the objective is estimation only. The OLS estimators are BLUE regardless of whether the u_i are normally distributed or not. In addition, because of the central limit theorem, we can argue that the test procedures – the t-tests and F-tests - are still valid asymptotically, i.e. in large sample.

Heteroscedasticity: The Error Variance is not Constant

The error terms in the regression equation have a common variance i.e., are Homoscedastic. If they do not have common variance we say they are **Hetroscedastic.** The basic questions to be addressed are:

- What is the nature of the problem?
- What are the consequences of the problem?
- How do we detect (diagnose) the problem?
- What remedies are available for the problem?

The Nature of the Problem

In the case of homoscedastic disturbance terms the spread around the mean is constant, i.e. = σ^2. But in the case of heteroscedasticity disturbance terms the variance changes with the explanatory variable. The problem of heteroscedasticity is likely to be more common in cross-sectional than in time-series data.

Causes of Heteroscedasticity

There are several reasons why the variance of the error term may be variable, some of which are as follows.

- Following the *error-learning* models, as people learn, their errors of behaviour become smaller over time where the standard error of the regression model decreases.

As income grows people have discretionary income and hence more scope for choice about the disposition of their income. Hence, the variance (standard error) of the regression is more likely to increase with income.

- Improvement in data collection techniques will reduce errors (variance).
- Existence of outliers might also cause heteroscedasticity.
- Misspecification of a model can also be a cause for heteroscedasticity.
- Skewness in the distribution of one or more explanatory variables included in the model is another source of heteroscedasticity.
- Incorrect data transformation and incorrect functional form are also other sources

Note: Heteroscedasticity is likely to be more common in cross-sectional data than in time series data. In cross-sectional data, individuals usually deal with samples (such as consumers, producers, etc) taken from a population at a given point in time. Such members might be of different size. In time series data, however, the variables tend to be of similar orders of magnitude since data is collected over a period of time.

Consequences of Heteroscedasticity

If the error terms of an equation are heteroscedastic, there are three major consequences.

a) The ordinary least square estimators are still linear since heteroscedasticity does not cause bias in the coefficient estimates. The least square estimators are still unbiased.

b) Heteroscedasticity increases the variance of the partial regression coefficients but it does affect the minimum variance property. Thus, the OLS estimators are inefficient. Thus the test statistics – t-test and F-test – cannot be relied on in the face of uncorrected heteroscedasticity.

Detection of Heteroscedasticity

There are no hard and fast rules (universally agreed upon methods) for detecting the presence of heteroscedasticity. But some rules of thumb can be suggested. Most of these methods are based on the examination of the OLS residuals, ei, since these are the once we observe and not the disturbance term u_i. There are informal and formal methods of detecting heteroscedasticity.

a) Nature of the problem

In cross-sectional studies involving heterogeneous units, heteroscedasticity is the rule rather than the exception.

Example: In small, medium and large sized agribusiness firms in a study of input expenditure in relation to sales, the rate of interest, etc. heteroscedasticity is expected.

b) Graphical method

If there is no a priori or empirical information about the nature of heteroscedasticity, one could do an examination of the estimated residual squared, ei^2 to see if they exhibit any systematic pattern. The squared residuals can be plotted either against Y or against one of the explanatory variables. If there appears any systematic pattern, heteroscedasticity might exist. These two methods are informal methods.

c) Park Test

Park suggested a statistical test for heteroscedasticity based on the assumption that the variance of the disturbance term (σi^2) is some function of the explanatory variable Xi.

Park suggested a functional form as: $\sigma_i^2 = \sigma^2 X_i^{\beta} e^{v_i}$ which can be transferred to a linear function using ln transformation. Hence, $Var(e_i) = \sigma^2 X_i^{\beta_i} e^{v_i}$ where v_i is the stochastic disturbance term.

$$\ln \sigma_i^2 = \ln \sigma^2 + \beta \ln X_i + v_i$$

$$\ln e_i^2 = \ln \sigma^2 + \beta \ln X_i + v_i$$ since σ^2 is not known.

The Park-test is a two-stage procedure: run OLS regression disregarding the heteroscedasticity question and obtain the ei and then run the above equation. The regression is run and if β turns out to be statistically significant, then it would suggest that heteroscedasticity is present in the data.

d) Spearman's Rank Correlation test

Recall that: $r_S = 1 - 6\left[\frac{\sum d_i^2}{N(N^2-1)}\right]$ d = difference between ranks

Step 1: Regress Y on X and obtain the residuals, ei.

Step 2: Ignoring the significance of ei or taking |ei| rank both ei and Xi and compute the rank correlation coefficient.

Step 3: Test for the significance of the correlation statistic by the t-test

$$t = \frac{r_S\sqrt{N-2}}{\sqrt{1-r_s^2}} \sim t(n-2)$$

If the computed t value exceeds the critical t value, we may accept the hypothesis of heteroscedasticity; otherwise we may reject it.

If more than one explanatory variable, compute the rank correlation coefficient between ei and each explanatory variable separately.

e) Goldfeld and Quandt Test

This is the most popular test and usually suitable for large samples. If it is assumed that the variance (σi^2) is positively related to one of the explanatory variables in the regression model and if the number of observations is at least twice as many as the parameters to be estimated, the test can be used.

Given the model

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

Suppose σi^2 is positively related to Xi as

$$\sigma_i^{\,2} = \sigma^2 X_i^{\,2}$$

Goldfeld and Quandt suggest the following steps:

1. Rank the observation according to the values of Xi in ascending order.
2. Omit the central **c** observations (usually the middle third of the recorded observations), or where **c** is specified a priori, and divide the remaining **(n-c)** observations into two groups, each with $\frac{(n-c)}{2}$ observations.ss
3. Fit separate regressions for the two sub-samples and obtain the respective residuals RSS, and RSS2 with $\frac{(n-c)}{2} - k$ df
4. Compute the ratio:

$$F = \frac{Rss_2 / df}{Rss_1 / df} \sim F_{V_1 V_2} \qquad v_1 = v_2 = \frac{(n-c-2k)}{2}$$

If in an application the **computed λ(= F)** is greater than the **critical F** at the chosen level of significance, we can reject the hypothesis of homoscedasticity, that is, we can say that **heteroscedasticity** is very likely.

If the two variances tend to be the same, then **F** approaches unity. If the variances differ we will have values for **F** different from one. The higher the F-ratio, the stronger the evidence of heteroscedasticity.

Note: There are also other methods of testing the existence of heteroscedasticity in your data. These are Glejser Test, Breusch-Pagan-Godfrey Test, White's General Test and Koenker-Bassett Test the details for which you are supposed to refer.

Remedial Measures

OLS estimators are still unbiased even in the presence of heteroscedasticity. But they are not efficient, not even asymptotically. This lack of efficiency makes the usual hypothesis testing procedure a dubious exercise. Remedial measures are, therefore, necessary. Generally the solution is based on some form of transformation.

a) The Weighted Least Squares (WLS)

Given a regression equation model of the form

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

The weighted least square method requires running the OLS regression to a transformed data. The transformation is based on the assumption of the form of heteroscedasticity.

Assumption One: Given the model $Y_i = \beta_0 + \beta_1 X_{1i} + U_i$

If $\text{var}(U_i) = \sigma_i^2 = \sigma^2 X_i^2$, then $E(U_i) = \sigma^2 X_{1i}^2$

Where σ^2 is a constant variance of a classical error term. So if as a matter of speculation or other tests indicate that the variance is proportional to the square of the explanatory variable **X**, we may transform the original model as follows:

$$\frac{Y_i}{X_{1i}} = \frac{\beta_0}{X_{1i}} + \frac{\beta_1 X_{1i}}{X_{1i}} + \frac{U_i}{X_{1i}}$$

$$= \beta_0 (\frac{1}{X_{1i}}) + \beta_1 + V_i$$

Now $E(V_i^2) = E(\frac{U_i}{X_{1i}}) = \frac{1}{X_{1i}^2} E(U_i^2) = \sigma^2$

Hence the variance of Ui is now homoscedastic and regress $\frac{Y}{X_{1i}}$ on $\frac{1}{X_{1i}}$.

Assumption Two: Again given the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + U_i$$

Suppose now $Var(U_i) = E(U_i^2) = \sigma_{U_i}^2 = \sigma^2 X_{1i}$

It is believed that the variance of Ui is proportional to X1i, instead of being proportional to the squared X1i. The original model can be transformed as follows:

$$\frac{Y_i}{\sqrt{X_{1i}}} = \frac{\beta_0}{\sqrt{X_{1i}}} + \beta_1 \frac{X_{1i}}{\sqrt{X_{1i}}} + \frac{U_i}{\sqrt{X_{1i}}}$$

$$= \beta_0 \frac{1}{\sqrt{X_{1i}}} + \beta_1 \sqrt{X_{1i}} + V_i$$

Where $V_i = U_i / \sqrt{X_{1i}}$ and where $X_{1i} > 0$

Thus, $E(V_i^2) = E(\frac{U_i}{\sqrt{X_{1i}}})^2 = \frac{1}{X_{1i}} E(U_i^2) = \frac{1}{X_{1i}} \cdot \sigma^2 X_{1i} = \sigma^2$

Now since the variance of Vi is constant (homoscedastic) one can apply the OLS technique to regress $\frac{Y}{\sqrt{X_{1i}}}$ on $\frac{1}{\sqrt{X_{1i}}}$ and $\sqrt{X_{1i}}$.

To go back to the original model, one can simply multiply the transformed model by $\sqrt{X_{1i}}$.

Assumption Three: Given the model let us assume that

$$E(U_i)^2 = \sigma^2[E(Y_i)]^2$$

The variance is proportional to the square of the expected value of Y.

Now, $E(Y_i) = \beta_0 + \beta_1 X_{1i}$

We can transform the original model as

$$\frac{Y_i}{E(Y_i)} = \frac{\beta_0}{E(Y_i)} + \beta_1 \frac{X_{1i}}{E(Y_i)} + \frac{U_i}{E(Y_i)}$$

$$= \beta_0 \frac{1}{E(Y_i)} + \beta_1 \frac{X_{1i}}{E(Y_i)} + V_i$$

Again it can be verified that $V_i = \dfrac{U_i}{E(Y_i)}$ gives us a constant variance σ^2

$$E(V_i) = E\left[\frac{U_i}{E(Y_i)}\right]^2 = \frac{1}{E(Y_i)^2} E(U_i)^2 = \frac{1}{[E(Y_i)]^2} \cdot \sigma^2[E(Y_i)]^2 = \sigma^2$$

The disturbance Vi is homoscedastic and the regression can be run.

Assumption Four: If instead of running the regression $Y_i = \beta_0 + \beta_1 X_{1i} + U_i$ one could run

$$\ln Y_i = \beta_0 + \beta_1 \ln X_{1i} + U_i$$

Then it reduces heteroscedasticity.

b) Other Remedies for Heteroscedasticity

Two other approaches could be adopted to remove the effect of heteroscedasticity.

- Include a previously omitted variable(s) if heteroscedasticity is suspected due to omission of variables.

- Redefine the variables in such a way that avoids heteroscedasticity. For example, instead of total income, we can use Income per capita.

Autocorrelation: Error Terms are correlated

Autocorrelation or serial correlation refers to the case in which the error term in one time period is correlated with the error term in any other time period. If the error term in one time period is correlated with the error term in the previous time period, there is first-order autocorrelation. Most of the applications in econometrics involve first rather than second- or higher-order autocorrelation. Even though negative autocorrelation is possible, most economic time series exhibit positive

It is the problem which affects the assumption of the regression model is the non-existence of serial correlation (autocorrelation) between the disturbance terms, Ui.

$$Cov(U_i, V_j) = 0 \qquad i \neq j$$

Serial correlation implies that the error term from **one time period** depends in some **systematic** way on error terms from **other time** periods. Autocorrelation is more a problem of **time series** data than **cross-sectional data**. If by chance, such a correlation is observed in **cross-sectional** units, it is called ***spatial autocorrelation***. So, it is important to understand serial correlation and its consequences of the OLS estimators.

Nature of Autocorrelation

The classical model assumes that the disturbance term relating to any observation is not influenced by the disturbance term relating to any other disturbance term.

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

But if there is any interdependence between the disturbance terms then we have autocorrelation

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

Causes of Autocorrelation

Serial correlation may occur because of a number of reasons.

- Inertia (built in momentum) – a salient feature of most economic variables time series (such as GDP, GNP, price indices, production, employment etc) is inertia or sluggishness. Such variables exhibit **(business)** cycles.
- Specification bias – exclusion of important variables or incorrect functional forms
- Lags – in a time series regression, value of a variable for a certain period depends on the variable's previous period value.
- Manipulation of data – if the raw d/.

Autocorrelation can be negative as well as positive. The most common kind of serial correlation is the **first order serial** correlation. This is the case in which this period error terms are functions of the previous time period error term.

$$E_t = PE_{t-1} + U_t$$

This is also called the first order autoregressive model.

$$-1 < P < 1$$

The disturbance term **Ut** satisfies all the basic assumptions of the classical linear model.

$$E(U_t) = 0$$

$$E(U_t, U_{t-1}) = 0 \quad t \neq t-1$$

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

Consequences of serial correlation

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

1) The estimates of the parameters remain **unbiased** even in the presence of autocorrelation but the X's and the *u*'s must be uncorrelated.

2) Serial correlation **increases the variance** of the OLS estimators. The minimum variance property of the OLS parameter estimates is violated. That means the OLS are **no longer efficient.**

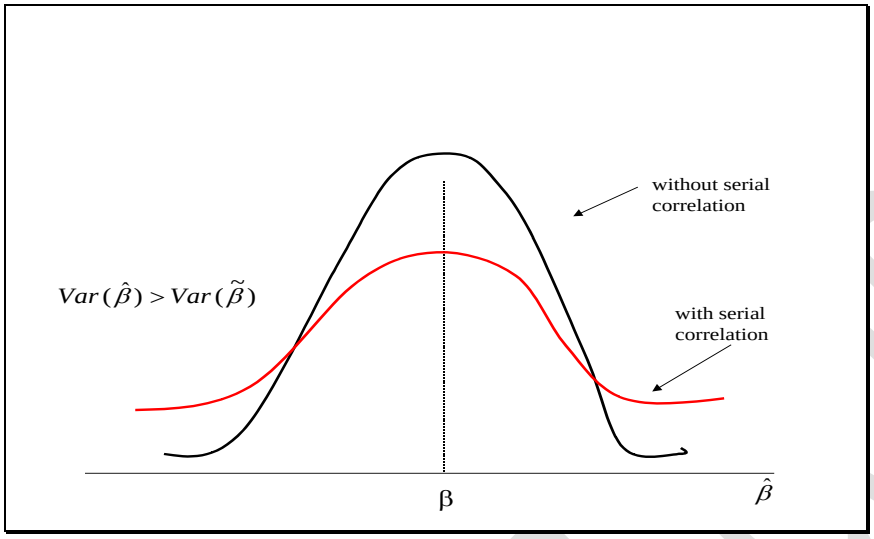


Figure 2: The distribution of with and without serial correlation.

3) Due to serial correlation the variance of the disturbance term, Ui may be **underestimated**.

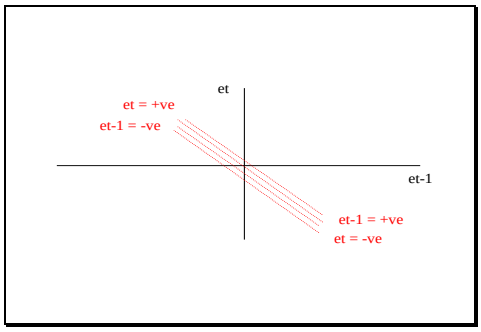
As a result, we are likely to overestimate R^2.

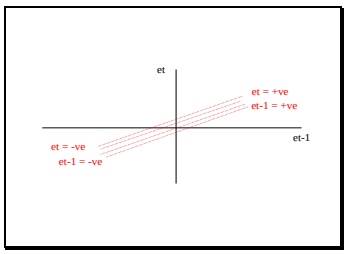
4. Therefore, the usual *t* and *F* tests of significance are no longer valid, and if applied, are likely to give seriously misleading conclusions about the statistical significance of the estimated regression coefficients.

4) If the Uis are autocorrelated, then prediction based on the ordinary least squares estimates will be inefficient. This is because of larger variance of the parameters. Since the variances of the OLS estimators are not minimal as compared with other estimators, the standard error of the forecast from the OLS will not have the least value.

Detecting Autocorrelation

Graphical Method- After running OLS regression, obtain the residuals and then we plotting the residuals either against their own lagged values or against time.





There are more accurate tests for the incidence of autocorrelation. The most common test of autocorrelation is the Durbin-Watson Test.

The Durbin-Watson d Test

The test for serial correlation that is most widely used is the Durbin-Watson d test. This test is appropriate only for the **first order autoregressive** scheme.

$U_t = PU_{t-1} + E_t$ then $E_t = PE_{t-1} + U_t$

The test may be outlined as

$$H_O : P = 0$$

$$H_1 : P \neq 0$$

This test is, however, applicable where the underlying assumptions are met:

- The regression model includes an intercept term
- The serial correlation is first order in nature

- The regression does not include the lagged dependent variable as an explanatory variable
- The error term is assumed to be normally distributed
- There are no missing observations in the data

The equation for the Durban-Watson d statistic is

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

Which is simply the ratio of the sum of squared differences in successive residuals to the RSS

Note that the numerator has one fewer observation than the denominator, because an observation must be used to calculate e_{t-1}. A great advantage of the d-statistic is that it is based on the estimated residuals. Thus, it is often reported together with R^2, t, etc.

The d-statistic equals **zero** if there is extreme positive serial correlation, **two** if there is no serial correlation, and **four** if there is extreme negative correlation.

1. Extreme positive serial correlation: $d \approx 0$

$e_t = e_{t-1}$ so $(e_t - e_{t-1}) \approx 0$ and $d \approx 0$.

2. Extreme negative correlation: $d \approx 4$

$e_t = -e_{t-1}$ and $(e_t - e_{t-1}) = (2e_t)$

thus $d = \frac{\sum (2e_t)^2}{\sum e_t^{\,2}}$ and $d \approx 4$

3. No serial correlation: d ≈ 2

$$d = \frac{\sum (e_t - e_{t-1})^2}{\sum e_t^{\,2}} = \frac{\sum e_t^{\,2} + \sum e_{t-1}^{\,2} - 2\sum e_t e_{t-1}}{\sum e_t^{\,2}} = 2$$

Since $\sum e_t e_{t-1} = 0$, because they are uncorrelated. Since $\sum e_t^{\,2}$ and $\sum e_{t-1}^{\,2}$ differ in only one observation, they are approximately equal.

The exact sampling or probability distribution of the d-statistic is not known and, therefore, unlike the t, X^2 or F-tests there are no unique critical values which will lead to the acceptance or rejection of the null hypothesis.

But Durbin and Watson have successfully derived the upper and lower bound so that if the computed value d lies outside these critical values, a decision can be made regarding the presence of a positive or negative serial autocorrelation.

Thus

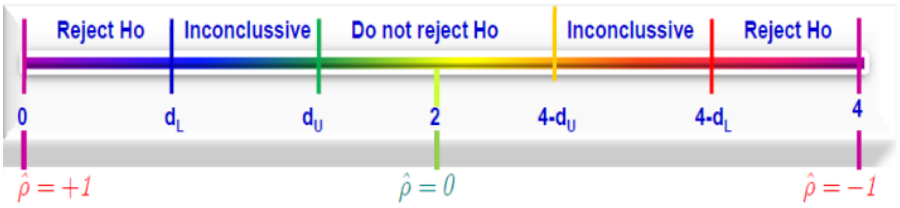
$$d = \frac{\sum (e_t - e_{t-1})^2}{\sum e_t^{\,2}} = \frac{\sum e_t^{\,2} + \sum e_{t-1}^{\,2} - 2\sum e_t e_{t-1}}{\sum e_t^{\,2}}$$

$$= 2(1 - \frac{\sum e_t e_{t-1}}{\sum e_{t-1}^{\,2}})$$

$$d = 2(1 - \hat{P}) \text{ since } \frac{\sum e_t e_{t-1}}{\sum e_{t-1}^{\,2}} = \hat{P}$$

But, since $-1 \le P \le 1$ the above identity can be written as: $0 \le d \le 4$

Therefore, the bounds of **d** must lie within these limits.



Some of the disadvantages of DW test are:

- It is used only for first – order serial correlation.
- It is biased towards 2 if there is a lagged dependent variable.
- It contains inconclusive region.

Thus if $\hat{P} = 0$ → d = 2, no serial autocorrelation.

if $\hat{P} = +1$ → d = 0, evidence of positive autocorrelation.

if $\hat{P} = -1$ → d = 4, evidence of negative autocorrelation.

Decision Rules for Durbin-Watson - d-test

Null hypothesis	Decision	If
No positive autocorrelation	Reject	$0 < d < d_L$
No positive autocorrelation	No decision	$d_L \le d \le d_U$
No negative autocorrelation	Reject	$4-d_L < d < 4$
No negative autocorrelation	No decision	$4-d_U \le d \le 4-d_L$
No autocorrelation	Do not reject	$d_U < d < 4-d_U$

\

Note: Other tests for autocorrelation include the **Runs test** and the **Breusch-Godfrey (BG) test**. There are so many tests of autocorrelation since there is no particular test that has been judged to be unequivocally best or more powerful in the statistical sense.

Remedial Measures for Autocorrelation

Since in the presence of serial correlation the OLS estimators are inefficient, it is essential to seek remedial measures.

1) The solution depends on the source of the problem.

- If the source is omitted variables, the appropriate solution is to include these variables in the set of explanatory variables.
- If the source is misspecification of the mathematical form the relevant approach will be to change the form.

2) If these sources are **ruled out** then the appropriate procedure will be to **transform the original data** so as to produce a model whose random variable satisfies the assumptions of non-autocorrelation. But the transformation depends on the pattern of autoregressive structure. Here we deal with first order autoregressive scheme.

$$\varepsilon_t = P\varepsilon_{t-1} + U_t$$

For such a scheme the appropriate transformation is to subtract from the original observations of each period the product of $\hat{P}$ times the value of the variables in the previous period.

$$Y_t^* = b_0^* + b_1 X_{1t}^* + \ldots + b_K X_{Kt}^* + U_t$$

where:

$$Y_t^* = Y_t - \hat{P}_{t-1}$$
$$X_{it}^* = X_{it} - PX_i(t-1)$$
$$V_t = U_t - PU_{t-1}$$
$$b_0^* = b_0 - Pb_0$$

Thus, if the structure of autocorrelation is known, then it is possible to make the above transformation. But often the structure of the autocorrelation is not known. Thus, we need to estimate **P** in order to make the transformation.

When ρ is not known

There are different ways of estimating the correlation coefficient, ρ , if it is unknown.

1) Estimation of ρ from the d-statistic

Recall that $d \approx 2(1-\hat{P})$ or $\hat{P} = 1 - \frac{d}{2}$

which suggest a simple way of obtaining an estimate of ρ from the estimated **d** statistic. Once an estimate of ρ is made available one could proceed with the estimation of the OLS parameters by making the necessary transformation.

2) Durbin's two step method

Given the original function as

$$Y_t - \hat{\hat{P}}Y_{t-1} = \beta_0(1-\hat{\hat{P}}) + \beta_1(X_t - \hat{\hat{\vartheta}}X_{t-1}) + U_t \; *^x$$

let $U_t = \vartheta U_{t-1} + V_t$

Step 1: start from the transformed model

$$(Y_t - PY_{t-1}) = \beta_0(1-P) + \beta_1(X_{1t} - PX_{1t-1}) + ... + \beta_K(X_{Kt}PX_{t-1}) + U_t \; *$$

rearranging and setting

$$\begin{aligned} &\beta_0(1-P) = a_0 \\ &\beta_1 = a_1 \\ &\beta_1\vartheta = a_2 \\ &etc. \end{aligned}$$

The above equation may be written as

$$Y_t = a_0 + PY_{t-1} + a_1X_{1t} + ... + a_KX_{Kt} + V_t$$

Applying OLS to the equation, we obtain an estimate ofρ, which is the coefficient of the lagged variable Y_{t-1}.

Step 2: We use this estimate, $\hat{\vartheta}$ to obtain the transformed variables.

$$(Y_t - PY_{t-1}) = Y_t^*$$
$$(X_{1t} - \vartheta X_{1t-1}) = X_{1t}^*$$
...
$$(X_{Kt} - PX_{Kt-1}) = X_{Kt}^*$$

We use this model to estimate the parameters of the original relationship.

$$Y_t^* = \beta_0 + \beta_1 X_{1t}^* + ... + \beta_K X_{Kt}^* + V_t$$

The methods discussed above to solve the problem of serial autocorrelation are basically two step methods. In step 1, we obtain an estimate of the unknown ρ and in step 2, we use that estimate to transform the variables to estimate the generalized difference equation.

Multicollinearity: Exact linear correlation between Regressors

One of the classical assumptions of the regression model is that the explanatory variables are uncorrelated. If the assumption that no independent variable is a perfect linear function of one or more other independent variables is violated, we have the problem of multicollinearity. If the explanatory variables are perfectly linearly correlated, the parameters become indeterminate. It is impossible to find the numerical values for each parameter and the method of estimation breaks.

If the correlation coefficient is 0, the variables are called *orthogonal;* there is no problem of multicollinearity. Neither of the above two extreme cases is often met. But some degree of inter-correlation is expected among the explanatory variables, due to the interdependence of economic variables.

Multicollinearity is not a condition that either exists or does not exist in economic functions, but rather a phenomenon inherent in most relationships due to the nature of economic magnitude. But there is no conclusive evidence which suggests that a certain degree of multicollinearity will affect seriously the parameter estimates.

Reasons for Existence of Multicollinearity

1. The data collection method employed- for example, sampling over a limited range of values of explanatory variables.

2. Constraints on the model or in the population being sampled.

3. Model specification

4. An over determined model-This happens when the model has more explanatory variables than the number of observations.

5. An additional reason for multicollinearity, especially in time series data, may be that the regressors included in the model share a common trend, that is, they all increase or decrease over time.

Consequences of Multicollinearity

Recall that, if the assumptions of the classical linear regression model are satisfied, the OLS estimators of the regression estimators are BLUE. As stated above if there is perfect multicollinearity between the explanatory variables, then it is not possible to determine the regression coefficients and their standard errors. But if collinearity among the **X**-variables is high, but not perfect, then the following might be expected.

1. Although BLUE, the OLS estimators have large variances and covariances, making precise estimation difficult.

2. Because of consequence 1:

a. the confidence intervals tend to be much wider, leading to the acceptance of the "zero null hypothesis" (i.e., the true population coefficient is zero) more readily.

b. the t ratio of one or more coefficients tends to be statistically insignificant

3. The OLS estimators and their standard errors can be sensitive to small changes in the data.

4. Although the t ratio of one or more coefficients is statistically insignificant, R2, the overall measure of goodness of fit, can be very high.

A high $\mathbf{R^2}$ but few significant t-ratios are expected in the presence of multicollinearity. So one or more of the partial slope coefficients are individually statistically insignificant on the basis of the

t-test. Yet the $\mathbf{R^2}$ may be so high. Indeed, this is one of the signals of multicollinearity, insignificant values but a high overall $\mathbf{R^2}$ and **F**-values. Thus because multicollinearity has little effect on the overall fit of the equation, it will also have little effect on the use of that equation for prediction or forecasting.

Detecting Multicollinearity

Having studied the nature of multicollinearity and the consequences of multicollinearity, the next question is how to detect multicollinearity. The main purpose in doing so is to decide how much multicollinearity exists in an equation, not whether any multicollinearity exists. So the important question is the degree of multicollinearity. But there is no one unique test that is universally accepted. Instead, we have some rules of thumb for assessing the severity and importance of multicollinearity in an equation. Some of the most commonly used approaches are the following:.

1) ***High pair-wise (simple) correlation coefficients among the regressors (explanatory variables).***

If the R's are high in absolute value, then it is highly probable that the **X**'s are highly correlated and that multicollinearity is a potential problem. The question is how high **r** should be to suggest multicollinearity. Some suggest that if **r** is in excess of 0.80, then multicollinearity could be suspected.

Another rule of thumb is that multicollinearity is a potential problem when the squared simple correlation coefficient is greater than the unadjusted R^2.

Two X's are severely multicollinear if $(r_{x_i x_j})^2 \geq R^2$.

A major problem of this approach is that although high zero-order correlations may suggest collinearity, it is not necessary that they be high to have collinearity in any specific case.

2) ***VIF and Tolerance***

VIF shows the speed with which the variances and covariances increase. It also shows how the variance of an estimator is influenced by the presence of multicollinearity. VIF is defined as followers:

$$VIF = \frac{1}{(1 - r^2_{23})}$$

Where r_{23} is the correlation between two explanatory variables. As r^2_{23} approaches 1, the VIF approaches infinity. If there no collinearity, VIF will be 1. As a rule of thumb, VIF value of 10 or more shows multicollinearity is sever problem. Tolerance is defined as the inverse of VIF.

Remedies for Multicollinearity

There is no automatic answer to the question "what can be done to minimize the problem of multicollinearity." The possible solution which might be adopted if multicollinearity exists in a function, vary depending on the severity of multicollinearity, on the availability of other data sources, on the importance of factors which are multicollinear, on the purpose for which the function is used. However, some alternative remedies could be suggested for reducing the effect of multicollinearity.

1) *Do Nothing*

Some writers have suggested that if multicollinearity does not seriously affect the estimates of the coefficients one may tolerate its presence in the function. In a sense, multicollinearity is similar to a non-life-threatening human disease that requires an operation only if the disease is causing a significant problem. A remedy for multicollinearity should only be considered if and when the consequences cause insignificant **t**-scores or widely unreliable estimated coefficients.

2) *Dropping one or more of the multicollinear variables*

When faced with severe multicollinearity, one of the simplest way to get rid of (drop) one or more of the collinear variables. Since multicollinearity is caused by correlation between the explanatory variables, if the multicollinear variables are dropped the correlation no longer exists.

Some people argue that dropping a variable from the model may introduce specification error or specification biases. According to them since OLS estimators are still BLUE despite near collinearity omitting a variable may seriously mislead us as to the true values of the parameters.

Example: If economic theory says that income and wealth should both be included in the model explaining the consumption expenditure, dropping the wealth variable would constitute specification bias.

3) Transformation of the variables

If the variables involved are all extremely important on theoretical grounds, neither doing nothing nor dropping a variable could be helpful. But it is sometimes possible to transform the variables in the equation to get rid of at least some of the multicollinearity.

Two common such transformations are:

(i) to form a linear combination of the multicollinear variables

(ii) to transform the equation into first differences (or logs)

The technique of forming a linear combination of two or more of the multicollinearity variables consists of:

- creating a new variable that is a function of the multicollinear variables
- using the new variable to replace the old ones in the regression equation (if X1 and X2 are highly multicollinear, a new variable, X3 = X1 + X2 or K1X1 + K2X2 might be substituted for both of the multicollinear variables in a re-estimation of the model)

The second kind of transformation to consider as possible remedy for severe multicollinearity is to change the functional form of the equation.

A first difference is nothing more than the change in a variable from the previous time-period.

$$\Delta X_t = X_t - X_{t-1}$$

If an equation (or some of the variables in an equation) is switched from its normal specification to a first difference specification, it is quite likely that the degree of multicollinearity will be significantly reduced for two reasons.

- Since multicollinearity is a sample phenomenon, any change in the definitions of the variables in that sample will change the degree of multicollinearity.
- Multicollinearity takes place most frequently in time-series data, in which first differences are far less likely to move steadily upward than are the aggregates from which they are calculated.

(4) Increase the sample size

Another solution to reduce the degree of multicollinearity is to attempt to increase the size of the sample. Larger data set (often requiring new data collection) will allow more accurate estimates than a small one, since the large sample normally will reduce somewhat the variance of the estimated coefficients reducing the impact of multicollinearity. But, for most economic and business applications, this solution is not feasible. As a result new data are generally impossible or quite expensive to find. One way to increase the sample is to pool cross-sectional and time series data.

UNIT SEVEN: NONLINEAR REGRESSION MODELS AND TIME SERIES ANALYSIS

Nonlinear regression models

some models may look nonlinear in the parameters but are inherently or intrinsically linear because with suitable transformation they can be made linear-in-the-parameter regression models.

NLRM means when the models cannot be linearized in the parameters , with suitable transformation.

Different example nonlinear regression model

- Consider now the famous Cobb–Douglas (C–D) production function. Letting Y =output, X2 =labor input, and X3 =capital input, we will write this function in three different ways

$$Y_i = \beta_1 X_{2i}^{\beta_2} X_{3i}^{\beta_3} e^{u_i}$$

$$\ln Y_i = \alpha + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + u_i$$

where α =lnβ1. Thus, in this format the C–D function is intrinsically linear

$$Y_i = \beta_1 X_{2i}^{\beta_2} X_{3i}^{\beta_3} u_i$$

Where we make transform,it can be

$$\ln Y_i = \alpha + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + \ln u_i$$

where α =lnβ1. This model too is linear in the parameters. But now consider the following version of the C–D function:

$$Y_i = \beta_1 X_{2i}^{\beta_2} X_{3i}^{\beta_3} + u_i$$

there is no way to transform the model so that it is intrinsically a nonlinear regression model

- Another well-known but intrinsically nonlinear function is the **constant elasticity of substitution (CES)** production function of which the Cobb– Douglas production is a special case. The CES production takes the following form:

$$Y_i = A\left[\delta K_i^{-\beta} + (1-\delta)L_i^{-\beta}\right]^{-1/\beta}$$

where Y = output, K = capital input, L = labor input, A= scale parameter, δ =distribution parameter ($0 < \delta < 1$), and β =substitution parameter ($\beta \geq -1$).

- **exponential regression** model and is often non-linear and used to measure the growth of a variable, such as population, GDP, or money supply.

$$Y_i = \beta_1 e^{\beta_2 X_i} + u_i$$

- A logistic growth model which is used to measure the growth of a population. where Y = population; t = time, measured chronologically; and the β's are the parameters.

$$Y_t = \frac{\beta_1}{1+\beta_2 e^{-\beta_3 t}} + u_t$$

- Notice an interesting thing about this model. Although there are only two variables, population and time, there are three unknowns, which shows that in a NLRM there can be more parameters than variables.

Limited Dependent Variable Models

The Linear Probability Model

It is among discrete choice models or dichotomous choice models. In this case the dependent variable takes only two values: 0 and 1. There are several methods to analyze regression models where the dependent variable is 0 or 1. The simplest method is to use the least squares method. In this case the model is called linear probability model. The other method is where there is an underlying or latent variable which we do not observe.

$$y = \begin{cases} 1 \ if \ \ y^* > 0 \\ 0 \ if \ \ y^* \le 0 \end{cases}$$

This is the idea behind Logit and Probit models

In this case the variable y is an indicator variable that denotes the occurrence and non occurrence of an event. For instance in the analysis of the determinants of unemployment, we have data on each person that shows whether or not the person is employed and we have some explanatory variables that determine employment.

In regression form that is written as:

$$y = x_i\beta + u_i \qquad 6.2$$

Where, $E(u_i) = 0$ and the conditional expectation $E(y_i / x_i) = x_i\beta$, which is the probability that the event will occur given x_i.

Since y_i takes only two values, 0 and 1, the regression in the above equation can take only two values,

$(1-\beta x_i)$ and $(-\beta x_i)$

The variance of u_i, $\text{var}(u_i) = \beta x_i (1 - \beta x_i)$

$$= E(y_i)\left[1 - E(y_i)\right]$$

Using OLS would result in heteroskedasticity problem.

This problem can be overcome by using the following two step estimation procedure.

1. Estimate $y_i = \beta x_i + u_i$ using OLS

2. Compute $\hat{y}_i\left(1 - \hat{y}_i\right)$ and use weighted least squares, i.e,

$$w_i = \sqrt{\hat{y}_i\left(1 - \hat{y}_i\right)}$$

Then, regress $\frac{y_i}{w_i}$ *on* $\frac{x_i}{w_i}$

However, the problem with this procedure (the least squares or weighted least squares) is:

1. $\hat{y}_i\left(1 - \hat{y}_i\right)$ may be negative

2. u_i are not normally distributed and there is problem with the application of the usual tests of significance.

3. The conditional expectation $E(y_i / x_i)$ be interpreted as the probability that the event will occur. In many cases $E(y_i / x_i)$ can lie outside the limits $[0,1]$.

The Probit and Logit Models

An alternative approach is to assume the following regression model

$$y_i^* = x\beta + \varepsilon_i$$

Where y_i^* is not observed. It is commonly called a latent variable. What we observe is a dummy variable y_i defined by:

$$y_i = \begin{cases} 1 \ \ if \ \ \ y_i^* > 0 \\ 0 \ \ otherwise \end{cases}$$

The Probit and Logit models differ in the specification of the distribution of the error term ε_i.

For instance, if the observed dummy variable is whether or not a person is employed or not, y_i^* would be defined as 'propensity or ability to find employment.

Thus,

$$p_i = prob(y_i = 1) = prob(\varepsilon_i > -\beta x_i)$$

$$= 1 - F(-(\beta x_i))$$

Where, F is the cumulative density function of ε.

a. The Probit Model:

The cumulative standard density is given:

$$p(Y=1)=\int_{-\infty}^{t}\frac{1}{\sqrt{2\pi}}e^{-\frac{Z^2}{2}}dt=\Phi(Z)$$

Where, $Z=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_kX_k$

b. The Logit Model:

The cumulative logistic function for Logit model is based on the concept of an odds ratio.

Let the log odds that $Y=1$ be given by:

$$\ln\left(\frac{p}{1-p}\right)=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_kX_k$$

Solving for the probability that $Y=1$ we will get:

$$\frac{p}{1-p}=e^Z$$

$$\Rightarrow P=(1-P)e^Z=e^Z-pe^Z$$

$$\Rightarrow p+pe^Z=e^Z$$

$$\Rightarrow p(1+e^Z)=e^Z$$

$$\Rightarrow p=\frac{e^Z}{1+e^Z}=\frac{1}{e^{-Z}(1+e^{-Z})}=\frac{1}{e^{-z}+1}=\frac{1}{1+e^{-z}}$$

The above logistic probability is simply denoted as $\Lambda(Z)$.

Both Probit and Logit distributions are 'S' shaped, but differ in the relative thickness of the tails. Logit is relatively thicker than Probit. This difference would, however, disappear, as the sample size gets large.

The relationship between $Z \& p(Y=1)$ can be represented as a 'latent' underlying index that determines choices. The latent index function, *Z*, is determined in linear fashion by a set of independent variables *X*. In turn, the latent index *Z* determines *P(Y=1)*

The Bernoulli trail of Probit and Logit model conditional on Z is given by:

$$f(Y/Z) = P^{Y}(1-P)^{1-Y}$$

Plugging either the standard normal cumulative density function (for Probit) or the cumulative logistic function (for Logit) into the above function to have the appropriate probability function gives:

$$f(Y_i/Z) = \Phi(Z)^{Y_i}(1-\Phi)^{(1-Y_i)}$$

for Probit model

$$f(Y_i/Z) = \Lambda(Z)^{Y_i}(1-\Lambda(Z))^{1-Y_i}$$

for Logit model

The likelihood function for these models is given by:

$$L(\beta_k/Y_i, X_i) = \prod_{i=1}^{n} \Phi(z)^{Y_i}(1-\Phi)^{(1-Y_i)}$$

for Probit Model

$$L(\beta_k/Y_i, X_i) = \prod_{i=1}^{n} \Lambda(z)^{Y_i}(1-\Lambda(z))^{(1-Y_i)}$$

for Logit Model

The Log Likelihood function of these models is give as:

$$\ln L(\beta_k/Y_i, X_i) = \sum_{i=1}^{n}\left[Y_i \ln(\Phi(z)) + (1-Y_i)\ln(1-\Phi(z))\right]$$

for Probit

and $\ln L(\beta_k / Y_i, X_i) = \sum_{i=1}^{n}\left[Y_i \ln(\Lambda(z)) + (1-Y_i)\ln(1-\Lambda(z))\right]$

for Logit

These functions can be optimized using standard methods to get the parameter values.

In choosing between Probit and Logit models, there is no statistical theory for preferring one to the other. Thus, it makes no difference which one to choose. The two models are quite similar in large samples. But in small samples the two models differ significantly

However, choice between the two models can be made on convenience. It is much easier to compute Probit probabilities (table of z statistic). Logit is simpler mathematically.

The probability model in the form of a regression is:

$$E(Y/X) = 0\left[1-F(\beta' X)\right] + 1\left[F(\beta' X)\right]$$

$$= F(\beta' X)$$

Whatever distribution is used, the parameters of the model like those of any other nonlinear regression model, are not necessarily the marginal effects:

$$\frac{\partial E(Y/X)}{\partial X} = \left\{\frac{dF(\beta' X)}{d(\beta' X)}\right\}\beta$$

$$= f(\beta' X)\beta$$

Where, $f(.)$ is the density function that corresponds to the cumulative density distribution, $F(.)$.

a) For the normal distribution, this is:

$$\frac{\partial E[Y/X]}{\partial X} = \phi(\beta' X)\beta \qquad 6.22$$

where $\phi(.)$ is the standard normal density.

b) For logistic distribution

$$\frac{d\Lambda(\beta' X)}{d(\beta' X)} = \frac{e^{\beta' X}}{\left(1+e^{\beta' X}\right)^2} = \Lambda(\beta' X)\left[1-\Lambda(\beta' X)\right] \quad 6.23$$

$$\frac{\partial E[Y/X]}{\partial X} = \Lambda(\beta' X)\left[1-\Lambda(\beta' X)\right]\beta \quad 6.24$$

In interpreting the estimated model, in most cases the means of the regressors are used. In other instances, pertinent values are used based on the choice of the researcher.

For an independent variable, say *k*, that is binary the marginal effect can be computed as:

$$prob\left[Y=1/\overline{X_*}, K=1\right] - prob\left[Y=1/\overline{X_*}, K=0\right] \quad 6.25$$

Where, $\overline{X_*}$ denotes the mean of all other variables in the model.

Therefore, the marginal effects can be evaluated at the sample means of the data. Or the marginal effects can be evaluated at every observation and the average can be computed to represent the marginal effects.

More generally, the marginal effects are give as

$$\frac{\partial p_i}{\partial x_{ij}} = \begin{cases} \beta_j \text{ for linear probability model} \\ \beta_j p_i (1-p_i) \text{ for the logit model} \\ \beta_j \phi(z_i) \text{ for the probit model} \end{cases}$$

Estimation of Binary Choice Models

The log likelihood function for the two models is:

$$\log L = \sum \left\{ y_i \log F(\beta' X) + (1-y_i) \log\left(1 - F(\beta' X)\right) \right\}$$

The first order condition with respect to the parameters of the model is be given by:

$$\frac{\partial \log L}{\partial \beta} = \sum \left[\frac{y_i f_i}{F_i} + (1-y_i) \frac{-f_i}{(1-F_i)} \right] X_i = 0$$

Where f_i is the density $\frac{dF_i}{d(\beta' X)}$, here i indicates that the function has an argument $\beta' X_i$.

i) For a normal distribution (Probit), the log likelihood is

$$\log L = \sum_{y_i=0} \log\left[1 - \Phi(\beta' X_i)\right] + \sum_{y_i=1} \log \Phi(\beta' X_i)$$

$$\frac{\partial \log L}{\partial \beta} = \sum_{y_i=0} \frac{-\phi_i}{1-\Phi_i} X_i + \sum_{y_i=1} \frac{\phi_i}{\Phi_i} X_i$$

ii) For a Logit model, the log likelihood is:

$$\ln L(\beta_k / Y_i, X_i) = \sum_{i=1}^{n} \left[Y_i \ln\left(\Lambda(z)\right) + (1-Y_i) \ln\left(1 - \Lambda(z)\right) \right]$$

$$\frac{\partial \log L}{\partial \beta} = \sum (y_i - \Lambda_i) X_i = 0$$

Measures of Goodness of fit

When the independent variable to be measure is dichotomous, there is a problem of using the conventional R^2 as a measure of goodness of fit.

1. Measures based on likelihood ratios

Let L_{UR} be the maximum likelihood function when maximized with respect to all the parameters and L_R be the maximum when maximized with restrictions $\beta_i = 0$.

$$R^2 = 1 - \left(\frac{L_R}{L_{UR}} \right)^{\frac{2}{n}}$$

Cragg and Uhler (1970) suggested a pseudo R^2 that lies between 0 and 1.

$$R^2 = \frac{L_{UR}^{2/n} - L_R^{2/n}}{\left(1 - LR^{2/n}\right) L_{UR}^{2/n}}$$

McFadden (1974) defined R^2 as

$$R^2 = 1 - \frac{\log L_{UR}}{\log L_R}$$

R^2 can also be linked in terms of the proportion of correct predictions. After computing $\hat{y}_i$, we can classify the i^{th} observation as belonging to group 1 if $\hat{y}_i < 0.5$ and group 2 if $\hat{y}_i > 0.5$. We can then count the number of correct predictions.

$$\hat{y}_i^* = \begin{cases} 1 \; if \;\; \hat{y}_i > 0.5 \\ 0 \;\; if \;\; \hat{y}_i < 0.5 \end{cases}$$

Count $$R^2 = \frac{No.\ of\ \ correct\ \ predictions}{total\ No.\ of\ \ observations}$$

Example: Regression results of a Probit model of house ownership and income is given below

$$y_i = 3.983 + 0.0485I + u_i$$
$$t \quad (69.53) \quad (19.56)$$

We want to measure the effect of a unit change in income on the probability of owning a house $(y = 1)$

$$\frac{dp_i}{dX_i} = f(\beta_1 + \beta_2 X_i)\beta_2$$

Where, $f(\beta_1 + \beta_2 X_i)$ is the standard normal probability density function evaluated at $\beta_1 + \beta_2 X_i$.

At value of $X = 6$, the normal density function at $f(-1.0166 + 0.04846) = f(-0.72548)$ that is equal to

$$f(-0.72548) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(-0.72548)^2}{2}\right) = 0.3066$$

Now multiplying this value by the slop coefficient of income, we get 0.01485.

Logit model of owning a house

$$L = -1.594\sqrt{w} + 0.07862X$$

This means for a unit increase in weighted income the weighted log of the odds in favour of owning a house goes up by 0.08 units.

Converting into odds ratio, we take the antilog

$$\frac{p_i}{1-p_i} = e^{-1.59\sqrt{w}+0.078X}$$

Maximum Likelihood Estimation

For Linear Regression Model, the MLE of a normal variable y_i conditional on x with mean $\alpha+\beta x$ and variance σ^2, the pdf for an observation is:

$$f\left(y_i/\alpha+\beta x,\sigma\right)=\frac{1}{\sigma\sqrt{2\Pi}}\exp^{\left(-\frac{1}{2}\frac{\left[y_i-\alpha-\beta x\right]^2}{\sigma^2}\right)}$$

The pdf of a normal variable with mean μ and σ^2 is often expressed in terms of the pdf standardized normal variable ϕ with mean 0 and variance of 1.

$$\phi(z)=\frac{1}{\sqrt{2\Pi}}\exp^{\left(-\frac{z^2}{2}\right)}$$

$$f\left(y_i/\alpha+\beta x,\sigma\right)=\frac{1}{\sigma}\left[\frac{1}{\sqrt{2\Pi}}\exp^{\left(-\frac{\left(\frac{y_i-\left[\alpha+\beta x_i\right]^2}{\sigma}\right)}{2}\right)}\right]$$

Thus,

$$=\frac{1}{\sigma}\phi\left(\frac{y_i-\left[\alpha+\beta x\right]}{\sigma}\right)$$

The Likelihood can be written as

$$L(\alpha,\beta,\sigma/y,x)=\prod_{t=1}^{n}\frac{1}{\sigma}\phi\left(\frac{y_i-\alpha-\beta x}{\sigma}\right)$$

Limited Dependent Variables

The density function of a normally distributed variable with mean μ and variance of σ^2 is given by:

$$f(y)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp^{\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma^2}\right)^2\right\}}$$

Where $y\sim N(\mu,\sigma^2)$

For a standard normal distribution,

$$\left(\frac{y-\mu}{\sigma}\right)\square\ N(0,1)$$

The density of a standard normal variable is

$$\phi(y)=\frac{1}{\sqrt{2\pi}}\exp^{\left\{-\frac{1}{2}y^2\right\}}$$

The cumulative density function of a normal distribution is

$$\Phi\left(\frac{y-\mu}{\sigma}\right)=\int_{-\infty}^{\left(\frac{y-\mu}{\sigma}\right)}\phi(t)\,dt$$

Due to symmetry, $\Phi(y)=1-\Phi(-y)$

In limited variable models we may encounter some form of truncation.

If y has density $f(y),$ the distribution of y truncated from below at a given c $(y\geq c)$ is given by:

$$f(y/y\geq c)=\frac{f(y)}{p(y\geq c)}$$ if $y\geq c$ and 0 otherwise

If y is a standard normal variable, the truncated distribution of $y \geq c$ has the probability:

$$p(y/y \geq c) = \lambda_1(c), \text{ where } \lambda_1 = \frac{\phi(c)}{1-\Phi(c)}$$

If the distribution is truncated from above $(y \leq c)$

$$p(y/y \leq c) = \lambda_2(c), \text{ where } \lambda_1 = \frac{-\phi(c)}{\Phi(c)}$$

If y has a normal distribution with mean μ and variance σ^2, the truncated distribution $y \geq c$ has mean

$$E(y/y \geq c) = \mu + \sigma\lambda_1(c^*), \text{ where } (c^*) \geq \mu$$

Where, $c^* = \left(\frac{c-\mu}{\sigma}\right)$

And $E(y/y \leq c) = \mu + \sigma\lambda_2(c^*), \text{ where } (c^*) \text{is} \leq \mu$

Tobit (Censored Regression) Model

In certain applications, the dependent variable is continuous, but its range may be constrained. Most commonly this occurs when the dependent variable is zero for a substantial part of the population but positive for the rest of the population.

$$y_i = y_i^* = x\beta + \varepsilon, \text{ if } \quad y_i^* > 0$$
$$= 0, \qquad \text{if } \quad y_i^* \leq 0$$

Where, $\varepsilon_i \square N(0, \sigma^2)$

In this model all negative values are mapped to zeros. i.e. observations are censored (from below) at zero

The model describes two things:

1. The possibility that $y_i = 0$ given x_i

$$p(y_i = 0) = p(y_i^* \le 0) = p(\varepsilon_i \le -x\beta)$$

$$= p\left\{\frac{\varepsilon_i}{\sigma} \le \frac{-x\beta}{\sigma}\right\} = \Phi\left\{\frac{-x\beta}{\sigma}\right\} = 1 - \Phi\left(\frac{x\beta}{\sigma}\right)$$

2. The distribution of y_i given that it is positive.

This is truncated normal distribution with expectation

$$E(y_i / y_i > 0) = x_i\beta + E(\varepsilon_i > -x\beta)$$

$$= x\beta + \sigma \frac{\phi\left(\frac{x\beta}{\sigma}\right)}{\Phi\left(\frac{x\beta}{\sigma}\right)}$$

The last term shows the conditional expectation of a mean zero normal variable given that it is no larger than $-x\beta.$ The conditional expectation of y_i no longer equals $x\beta$ but depends non linearly on x_i through $\frac{\phi(.)}{\Phi(.)}$.

Marginal effects of the Tobit Model

1. The probability of a zero outcome is:

$$p(y_i = 0) = 1 - \Phi\left(\frac{x\beta}{\sigma}\right)$$

$$\frac{\partial p(y_i = 0)}{\partial x_k} = -\phi\left(\frac{x\beta}{\sigma}\right)\frac{\beta_k}{\sigma}$$

2. The expected value of y_i (positive values) is

$$E(y_i) = x\beta\Phi\left(\frac{x\beta}{\sigma}\right) + \sigma\phi\left(\frac{x\beta}{\sigma}\right)$$

Thus the marginal effect on the expected value of y_i of a change in x_k is given by

$$\frac{\partial E(y_i)}{\partial x_k} = \beta_k\Phi\left(\frac{x\beta}{\sigma}\right)$$

This means the marginal effect of a change in x_k upon the expected outcome y_i is given by the model's coefficient multiplied by the possibility of having a positive outcome.

3. The marginal effect up on the latent variable is

$$\frac{\partial E(y_i^*)}{\partial x_k} = \beta_k$$

Maximum Likelihood Estimation of the Tobit Model

The contribution of an observation either equals the probability mass (at the observed point $y_i = 0$) or the conditional density of y_i, given that it is positive times the probability mass of observing $y_i > 0$.

$$\log L(\beta, \sigma^2) = \sum \log p(y = 0) + \sum \log f(y_i / y_i > 0) + \log p(y_i > 0)$$

$$= \sum \log p(y = 0) + \sum \log f(y_i)$$

Using appropriate expression for normal distribution we can obtain:

$$\log L\left(\beta,\sigma^2\right)=\sum\log\left[1-\Phi\left(\frac{x\beta}{\sigma}\right)\right]+\sum\log\left[\frac{1}{\sqrt{2\Pi\sigma^2}}\exp^{\frac{-1}{2}\left(\frac{y_i-x\beta}{\sigma}\right)^2}\right]$$

Maximizing this function with respect to the parameters will give the maximum likelihood estimate

Sample Selection

Tobit model imposes a structure that is often restrictive: exactly the same variables affecting the probability of nonzero observation determine the level of positive observation and more over with the same sign.

This implies, for example, that those who are more likely to spend a positive amount are, on average, also those that spend more on durable goods.

For example, we might be interested in explaining wages. Obviously wages are observed for people that are actually working, but we might be interested in (potential) wages not conditional on this selection.

For example, a change in some variable x may lower someone's wage such that he decides to stop working. Consequently, his wage would not be observed and the effect of this variable could be underestimated from the available data.

Because, a sample of workers may not be a random sample of the population (of potential workers), one may expect that people with lower (potential) wages are more likely to be unemployed - This problem is often referred to as sample selection.

Consider the following sample selection model of wage:

$$W_i^*=\chi_1\beta_1+\varepsilon_1$$

Where, χ_1 denotes vector of exogenous characteristics of the person, W_i^* denotes the persons wage.

The wage w_i^* is not observable for people that are not working.

Thus to describe whether a person is working or not, a second equation is specified, which is binary choice type:

$$h_i^* = x_2\beta_2 + \varepsilon_2$$

Where,

$$w_i = w_i^*, h_i = 1, \text{ if } h_i^* > 0$$

$$w_i \text{ is not observed}, h_i = 0 \text{ if } h_i^* \le 0$$

The binary variable h_i indicates working or not working. The error terms of the two equations have mean of zero with variances of σ_1^2, σ_2^2, respectively and covariance of σ_{12}.

One usually sets the restriction, $\sigma_2^2 = 1$ for normalization restriction of the Probit model. The conditional expected wage given that a person is working is given by:

$$E\left[w_i / h_i = 1\right] = x_1\beta_1 + E\left[\varepsilon_1 / h_i = 1\right]$$

$$= x_1\beta_1 + E\left[\varepsilon_1 / \varepsilon_2 > -x_2\beta_2\right]$$

$$= x_1\beta_1 + \frac{\sigma_{12}}{\sigma_2^2} E\left[\varepsilon_1 / \varepsilon_2 > -x_2\beta_2\right]$$

$$= x_1\beta_1 + \frac{\sigma_{12}}{\sigma_2^2} E\left[\varepsilon_1 / \varepsilon_2 > -x_2\beta_2\right]$$

$$= x_1\beta_1 + \sigma_{12} \frac{\phi\left(x_2\beta_2\right)}{\Phi\left(x_2\beta_2\right)}$$

The conditional expected wage equals $x_1\beta_1$ only if $\sigma_{12} = 0$. So if the error terms of the two equations are not correlated the wage equation can be consistently estimated by OLS.

A sample selection bias of OLS arises if $\sigma_{12} \neq 0.$

The term $\frac{\phi\left(x_2\beta_2\right)}{\Phi\left(x_2\beta_2\right)}$ is known as the inverse Mill's ratio and is denoted by $\lambda\left(x_2\beta_2\right)$ by Heckman (1979) and is referred as Heckman's model.

Time Series analysis

A time series data set consists of observations on a variable or several variables over time. In economics examples of time series data include stock prices, money supply, consumer price index, gross domestic product, exchange rates, exports, etc. In such time series data set, time is an important dimension because past as well as current events influence future events (that is, lags do matter in time series analysis and time series data. Unlike the arrangement of cross-sectional data, the chronological ordering of observations(variables) in a time series expresses potentially important information. Thus, a key feature of time series data that makes it more difficult to analyze is the fact that economic observations can rarely be assumed to be independent across time. Therefore, in general a time series data is a sequence of numerical data in which each variable is associated with a particular instant in time. Univariate time-series analysis- analysis of single sequence of data describing the behavior of one variable in terms of its own past values.

Example: Autoregressive models:

$u_t = \rho u_{t-1} + \varepsilon_t$ first order autoregressive or

$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \varepsilon_t$ second order autoregressive

Analysis of several sets of data(variables) for the same sequence of time periods is called multivariate time-series analysis. Examples, analysis of the relationships among price level, money supply and GDP on the basis of say quarterly or annual collected data). The main purpose of **time-series analysis** is to study the **dynamics** or temporal structure of the data.

Stationary and Non-stationary stochastic processes

From theoretical point of view, the collection of random variable yt ordered in time is called a stochastic process or random process. There are two different classes of the stochastic process.

- ✓ Stationary stochastic process-gives rise to stationary time series.
- ✓ Nonstationary stochastic process- give rise to nonstationary time series.

Stationary Stochastic Processes

Stochastic process is said to be stationary if its mean and variance are constant over time (do not depend on time or do not change as time changes). Moreover, the value of the covariance between the two time periods depends only on the lag between the two time periods and not on the actual time. In the time series, a stochastic process that satisfies such conditions is known as weakly stationary, or covariance stationary.

Thus, weakly stationary or covariance stationary process is a process that satisfies the following conditions for a given stochastic time series Yt;

$$Mean = E(Y_t) = u$$
$$Variance = \text{var}(Y_t) = E(Y_t - u)^2 = \sigma^2$$
$$\text{Covariance} = \gamma k = E[(Y_t - u)(Y_{t-k} - u)]$$

where γk, is the covariance (at lag k) between the values of Yt and Yt-k. If k = 0, we obtain γ0, which is simply the variance of Y (= σ2); if k = 1, γ1 is the covariance between two adjacent values of Y. Even if we shift the origin of Y from Yt to Yt+m, time will not affect the mean, variance, and covariance. So, at any point we measure them, they are time invariant.

If a time series is not stationary as defined above, it is called a non-stationary time series. In other words, a non-stationary time series will have a time varying mean or a time-varying variance or both. Time series stochastic processes that are stationary (stationary time series) are important to make generalization for other time periods and thus to conduct reliable forecasts.

If a time series is non-stationary, for example, we can study its behavior only for the time period under consideration. Which means it is not possible to generalize the analysis to other time periods. As such non-stationary time series are not useful for forecasting.

Non-stationary Stochastic Processes

In practical research one often encounters non-stationary time series. The classic example is the Random Walk Model (RWM). According to RWM, it is often said that stock prices follow a random walk; that is, they are non-stationary.

We distinguish two types of random walks:

- ✓ random walk model without drift (with no intercept term)
- ✓ random walk model with drift (constant term is present).

Random Walk without Drift

The series process, Yt is said to be a random walk without drift if; $Y_t = Y_{t-1} + u_t$ where ut is a white noise error term (error term with mean 0 and variance σ2).

This model says that the value of Y at time period t (i.e., Yt)is equal to its value at time (t−1) plus a random shock(ut) and it is an AR(1) model, because it is regressed on itself lagged one period.

We can write the above model as;

$$Y_1 = Y_0 + u_1$$
$$Y_2 = Y_1 + u_2 = Y_0 + u_1 + u_2$$
$$Y_3 = Y_2 + u_3 = Y_0 + u_1 + u_2 + u_3$$

An interesting feature of RWM is the persistence of random shocks (i.e., random errors).

In the model above:

Yt is the sum of initial Y0 plus the sum of random shocks. As a result, the impact of a particular shock does not die away.

For example, if u2 = 2 rather than u2 = 0, then all Yt's from Y2 onward will be 2 units higher and the effect of this shock never dies out.

In general, if the process started at some time 0 with a value of Y0, we have;

$$Y_t = Y_0 + \sum u_t$$
$$E(Y_t) = E\left(Y_0 + \sum u_t\right) = Y_0 \quad \text{(Why?)}$$

But the variance of Yt is not constant and given by:

$$\mathbf{var}(Y_t) = t\sigma^2$$

The expressions say that the mean value of Y is equal to its initial, or starting, value, which is constant, but as t increases, its variance increases indefinitely, thus violating a condition of stationarity. In short, the RWM without drift is a nonstationary stochastic process. First differencing of the above RWM gives as the following interesting feature.

$$(Y_t - Y_{t-1}) = \Delta Y_t = u_t$$

where Δ is the first difference.

From this it is easy to show that, while the initial expression, that is, Yt, is nonstationary, its first difference is stationary. Therefore, the first differences of a random walk time series are stationary

Random Walk with Drift (with intercept)

Let us modify the above RWM as follows:

$$Y_t = \delta + Y_{t-1} + u_t$$

where δ is known as the drift parameter

Why we call it drift? because if we write the preceding equation as;

$$Y_t - Y_{t-1} = \Delta Y_t = \delta + u_t$$

The model will show that Yt drifts upward or downward, depending on whether δ being positive or negative. Note that RWM with drift is also an AR(1) model. Following similar procedure discussed above for random walk without drift, it can be shown that for the random walk with drift model;

$$E(Y_t) = Y_0 + t \cdot \delta$$

$$\text{var}(Y_t) = t\sigma^2$$

From this expression we can see that for the RWM with drift the mean and the variance increase over time (that is, they are not constant), violating the conditions for weak stationarity. Therefore, in general we can conclude that the Random Walk Model (with or without drift) is non-stationary stochastic process.