Table of Contents

Chapter 1: What is Econometrics-What it is all about?

Definition, Scope and Division of Econometrics

1.1. Definition and Scope of Econometrics

Econometrics deals with the measurement of **economic relationships**. Econometrics is a combination of economic theory, mathematical economics, and statistics, but it is completely distinct from each of them.

Econometrics may be defined as the social science in which the tools of economic theory, mathematics, and statistical inference are applied to the **analysis of economic phenomena.**

Thus 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 e.g. elasticities, propensities, etc….

Econometric methods are statistical methods specifically adapted to the peculiarities of economic phenomenon. The most important characteristics of economic relationships are that they contain a **random** element. This **random** element is ignored by economic theory and mathematical economics that postulate **exact** relationships between various economic magnitudes. Econometrics has developed methods for dealing with the random component of economic relationships.

Example

Economic theory postulates that the demand for a commodity depends on its price, on the prices of other commodities, on consumer's income, and its tastes. This is an **exact** relationship because it implies that demand is completely determined by the above four factors. No other factor except that above influences demand.

- **In mathematical terms this is expressed as:**

$$Q = b_0 + b_1 P + b_2 P_O + b_3 Y + b_4 t$$

Where, **Q** = Quantity demanded of a particular commodity

P = Price of the commodity

Po = Prices of other commodities

Y = Consumers income

t = Tastes

b_0, b_1, b_2, b_3, b_4 = Coefficients of the demand function

The above equation states that it is only the four factors that affect the quantity demanded of a commodity. However, in economic life many factors may affect demand. For instance, expectations, the invention of a new product, a war, professional changes, changes in law, etc.

In econometrics the influence of these "**other**" factors is taken into account. A **random** variable with specific characteristics is introduced into the economic relationships. To our model in the above equation, we add **u** for random factors that affect quantity demanded.

$$Q = b_0 + b_1 P + b_2 P_O + b_3 Y + b_4 t + u$$

U = stands for random factors that affect quantity demanded.

Econometrics presupposes the existence of a body of economic theory. **Economic theory** should come **first** because it sets the hypotheses about economic behavior that should be tested with the application of econometric techniques.

In testing hypotheses:

a) Formulate the mathematical relation that constitutes the model of the maintained hypothesis.

b) Confront the model with observational data referring to the actual behavior of the economic units (i.e. consumers or producers).

- Test the validity of the hypotheses (by collecting and analyzing data using appropriate statistical techniques).

- This is done to establish whether the theory can explain the actual behavior of the economic units.

That is whether the theory is compatible with facts. If the theory is compatible with the actual data, we accept the theory as valid. If the theory is incompatible with the observed behavior: We either reject the theory, or in the light of the empirical evidence of the data, we may modify it. In the second case one needs additional new observations in order to test the revised version of the theory.

1.1.1. Econometrics and Mathematical Economics

Mathematical economics states economic theory in terms of mathematical symbols. There is no difference between mathematical economics and economic theory. Both state the same relationship. Economic theory makes **statements or hypotheses** that are mostly **qualitative** in nature (the theory uses verbal expression). E.g. the law of demand, the law does not provide any **numerical** measure of the relationship. This is the job of the econometrician. Both express the various economic relationships in an **exact** form. Neither economic theory nor mathematical

economics allows for **random elements.** These random elements might affect the relationship and make it stochastic. They do not provide **numerical values** for the coefficients of the relationships. Econometrics presupposes the expression of economic relationships in mathematical form. But it does not assume that economic relationships are exact. Econometric methods provide **numerical values** of the coefficients of economic phenomena.

1.1.2. Econometrics and Statistics

Econometrics differs both from mathematical statistics and economic statistics. An economic statistician gathers empirical data, records them, tabulates them or charts them, and attempts to describe the pattern in their development over time and perhaps detect some relationship between various economic magnitudes. It does **not go any further**. The one who does that is the econometrician. Mathematical (or inferential) statistics deals with the methods of measurement, which are developed on the basis of controlled experiments in the laboratories. Statistical methods of measurement are not appropriate for economic relationships which cannot be measured on the basis of evidence provided by controlled experiments, because such experiments cannot be designed for economic phenomenon.

For instance, in physics and some other sciences the researcher can hold all other conditions constant and change only one element in performing an experiment. **Example:** A plant scientist/Agronomist can measure the impact of **fertilizer** on **wheat productivity** by keeping the use of improved seeds constant. Or he can measure the impact of **improved wheat variety** on the **productivity of wheat** by keeping the utilization of fertilizers constant.

He can then record the results of such change and apply the classical methods to deduce the laws governing the phenomenon being investigated. However, in studying the economic behavior of human beings one cannot change only one factor while keeping all other factors constant. In the real world, all variables change continuously and simultaneously, so that controlled experiments are impossible.

We cannot change only incomes, keeping prices, tastes and other factors constant. Econometrics uses statistical methods after adopting them to the problems of economic life. These adopted statistical methods are called econometric methods. This means econometric methods are adjusted so that they become appropriate for the measurement of economic relationships that are stochastic, that is they include random elements. The adjustment consists primarily in specifying the stochastic (random) element that are supposed to operate in the real world and enter into the determination of the observed data, so that the latter can be interpreted as a (random) sample to which the methods of statistics can be applied.

1.2. Goals of Econometrics

There are about **three** main goals of econometrics.

a) First, econometrics is used for the **analysis of theory. Testing economic theory. Example:** Educated household heads are more likely to adopt new technologies than others. Males are less likely to repay their debts than females.

b) Second, econometrics is used for **policy making.** That is supplying numerical estimates of the coefficients of economic relationships, which may be used for decision-making.

Examples

- Price elasticity of demand for **Injera** is 0.5. What advise need to be given for producers/suppliers? Should the supplier increase or decrease the price of Injera?
- Price elasticity of demand for **Beer** is 1.5. What advise need to be given for producers/suppliers? Should the supplier increase or decrease the price of **Beer**?

c) Third, econometrics is used for **forecasting.**

1.2.1. Analysis: Testing Economic Theory

In earlier stages of the development of economic theory, economists formulated the basic principles of the functioning of economic systems using verbal exposition and applying deductive procedures. Earlier economic theories started from a set of observations concerning the behavior of individuals as consumers and producers. Some assumptions were set regarding the motivation of individual units.

For instance, in demand theory it was assumed that consumers maximize utility and in supply theory producers were assumed to maximize profit. By logical reasoning, economists were able to derive some logical conclusions (laws) concerning the working processes of the economic system. However, no attempt was made to test the theories against reality. Econometrics aims primarily at the **verification** of economic theories. This means obtaining empirical evidence to test the explanatory power of economic theories, to decide how well they explain the observed behavior of the economic units.

1.2.2. Policy Making (Decision Making)

Various econometric techniques are used to obtain reliable estimates of the individual **coefficients** of economic relationships. From these values one evaluate parameters of economic relationships such as elasticities, propensities, marginal values, multipliers, etc…

1.2.3. Forecasting the future value of economic magnitudes

Forecasting the values of economic magnitudes is essential for policy makers to judge whether it is necessary to take any measure in order to influence the relevant economic variable. For example, the government may be interested in knowing the level of unemployment after five years.

1.3. Methodology of Econometrics

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

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

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

1. **Statement of Theory or Hypothesis**

Keynes states that on average, consumers increase their consumption as their income increases, but not as much as the increase in their income (**MPC < 1**).

Formulation of the maintained hypothesis

It involves determining

a. Dependent and explanatory variables
b. The ***a priori*** theoretical expectations about the signs & the size of the parameters of the function to form the basis for the evaluation of the model
c. The mathematical form of the model: single vs. simultaneous equation; linear vs. nonlinear functional forms

2. **Specification of the Mathematical Model of Consumption (single-equation model)**

$Y = \beta_1 + \beta_2 X \qquad 0 < \beta_2 < 1 \qquad (1.3.1)$

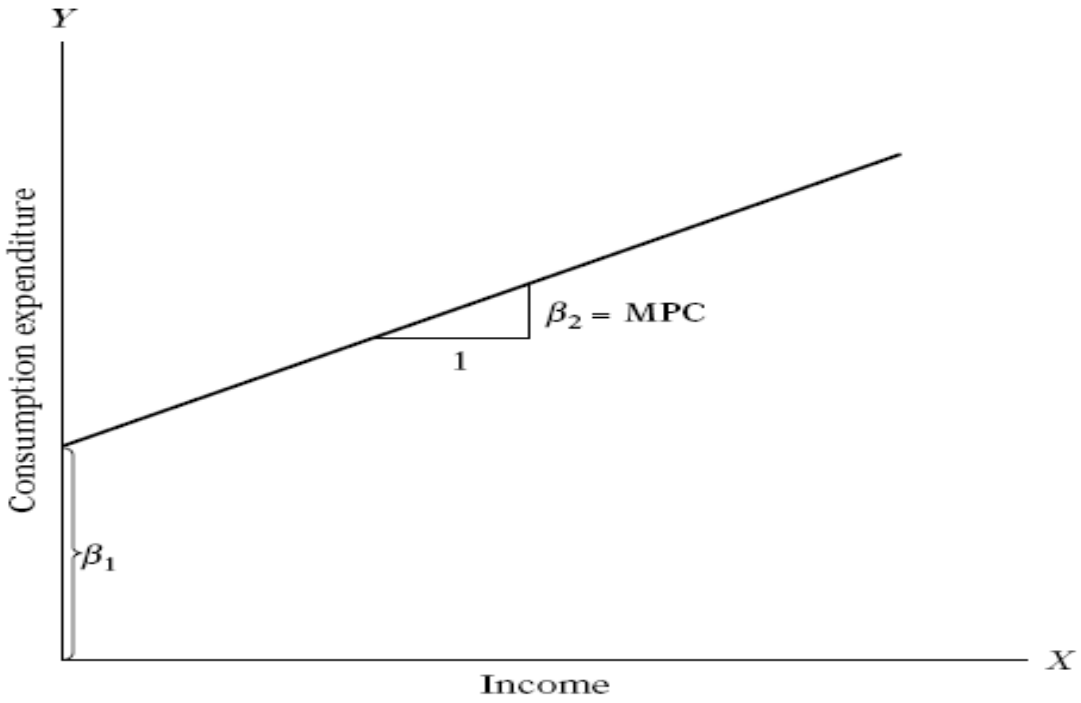
Y = consumption expenditure and (dependent variable)

X = income, (independent, or explanatory variable)

β_1 = the intercept

β_2 = the slope coefficient

- The slope coefficient β_2 measures the MPC.



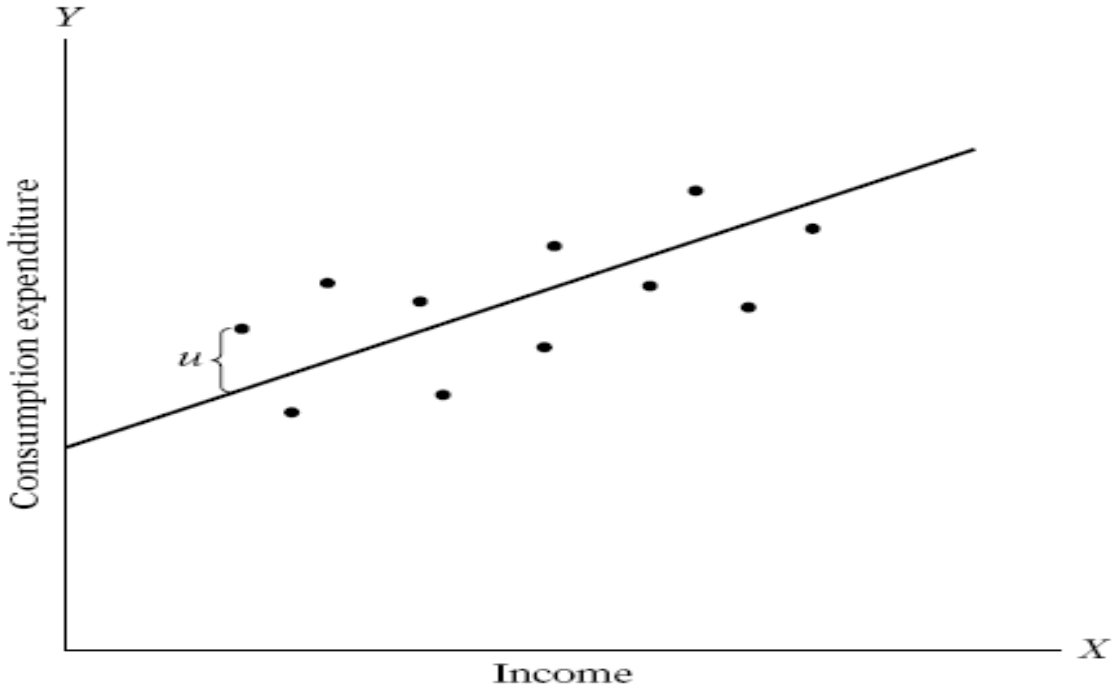
3. **Specification of the Econometric Model of Consumption**

The **relationships** between economic variables are generally **inexact.** In addition to **income**, other variables affect consumption expenditure. For example, **size of family, ages of the members in the family, family religion,** etc., are likely to exert some influence on consumption. To allow for the **inexact** relationships between economic variables, the above equation (I.3.1) is modified as follows:

- $Y = \beta_1 + \beta_2 X + u$ (1.3.2)

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

(I.3.2) is an example of a linear regression model, i.e., it hypothesizes that **Y is linearly related to X,** but that the relationship between the two **is not exact**. It is subject to individual variation. The econometric model of (I.3.2) can be depicted as shown in Figure I.2.



- In the above figure, it is observed that,
  - There are 11 household heads
  - A positive relationship between the two variables
  - **U** represents the difference between the actual expenditure which is represented by (**dots**) and the regression line (**predicted values**)

4. **Obtaining Data**

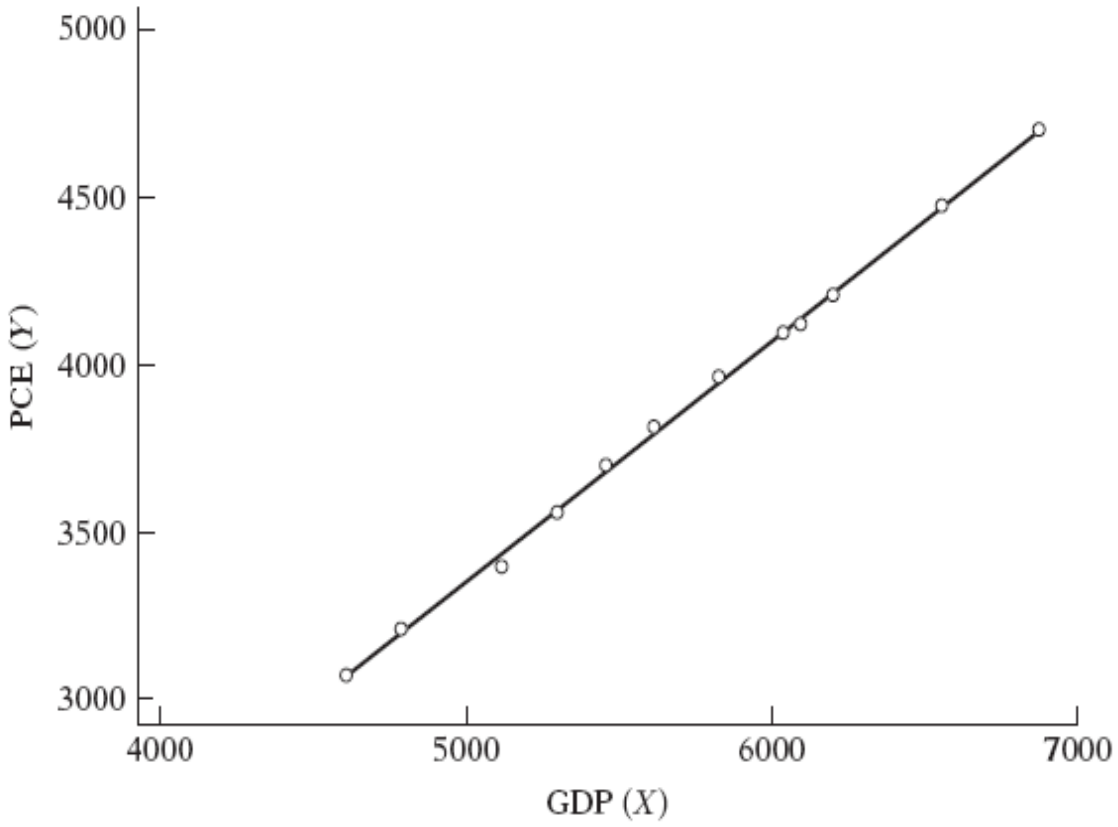
To obtain the numerical values of β_1 and β_2, we need data.

Look at Table 1, which relate to the **personal consumption expenditure (PCE**) and the gross domestic product (**GDP**).

- The data are in "real" terms.

Table 1

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



5. **Estimation of the Econometric Model**

Regression analysis is the main tool used to obtain the estimates. Using this technique and the data given in Table 1, we obtain the following estimates of

- $\beta_1 = -184.08$ **and**
- β_2, = **0.7064.**

Thus, the estimated consumption function is:

$$\hat{Y} = -184.08 + 0.7064X_i \qquad \text{(I.3.3)}$$

The estimated regression line is shown in Figure I.3. The **regression line fits** the data quite well.

The **slope** coefficient (i.e., the **MPC**) was about **0.70**. An increase in real income of 1 dollar led, on average, to an increase of about 70 cents in real consumption.

6. **Hypothesis Testing**

That is to find out whether the estimates obtained in, Eq. (I.3.3) are in accord **with the expectations of the theory that is being tested**. Keynes expected the **MPC to be positive but less than 1.** In our example we found the **MPC** to be about **0.70.** But before we accept this finding as confirmation of Keynesian consumption theory, we must enquire whether this estimate is sufficiently below unity. In other words, is **0.70 statistically less than 1?** If it is, it may support Keynes' theory. Such confirmation or refutation of economic theories on the basis of sample evidence is based on a branch of statistical theory known as **statistical inference (hypothesis testing).**

7. **Forecasting or Prediction**

To illustrate, suppose we want to predict the mean consumption expenditure for 1997. The GDP value for 1997 was **7269.8** billion dollars consumption would be:

$$\hat{Y}_{1997} = -184.0779 + 0.7064(7269.81) = 4951.30 \qquad \text{(I.3.4)}$$

The **actual value** of the consumption expenditure reported in **1997 was 4913.5** billion dollars. The estimated model (I.3.3) thus **over-predicted** the actual consumption expenditure by about **37.82** billion dollars. We could say the **forecast error** is about **37.8** billion dollars, which is about **0.76** percent of the actual GDP value for 1997.

Now suppose the government decides to propose a **reduction in the income tax**. What will be the effect of such a policy on income and thereby on consumption expenditure and ultimately on employment?

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

$$M = \frac{1}{1 - MPC} \qquad \text{(I.3.5)}$$

The multiplier is about M = 3.33.

That is, an increase (decrease) of a dollar in investment will eventually lead to more than a threefold increase (decrease) in income. Note that it takes time for the multiplier to work. The critical value in this computation is **MPC.**

Thus, a quantitative estimate **of MPC provides valuable information** for policy purposes. **Knowing MPC, one can predict the future course of income, consumption expenditure, and employment** following a change in the government's fiscal policies.

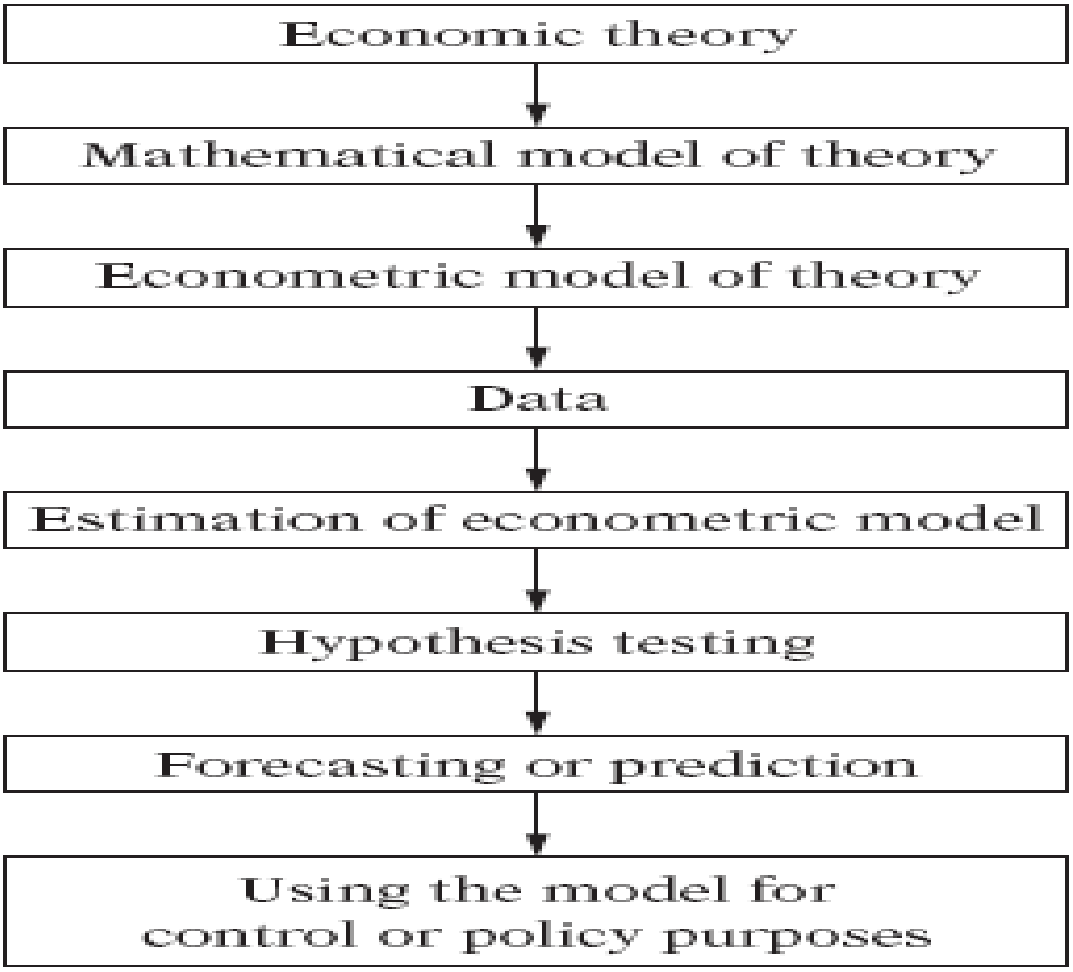
8. **Use of the Model for Control or Policy Purposes**

Suppose we have the estimated consumption function given in (I.3.3). Suppose further the government believes that consumer expenditure of about **4900** will keep the **unemployment rate at its current level of about 4.2%.** What level of income will guarantee the target amount of consumption expenditure? If the regression results given in (I.3.3) seem reasonable, simple arithmetic will show that:

$$\mathbf{4900 = -184.0779 + 0.7064X} \qquad \text{(I.3.6)}$$

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

Figure I.4 summarizes the anatomy of classical econometric modeling



1.4. Elements of Econometrics

- **Econometric inputs:**

✓ Economic Theory

✓ Mathematics

✓ Statistical Theory

✓ Data

✓ Computers (CPU power)

✓ Interpretation

- **Econometric outputs:**

✓ Estimation – Measurement

✓ Inference - Hypothesis testing

✓ Forecasting – Prediction

✓ Evaluation

1.5. Types of Econometrics

Econometrics may be classified in two branches as theoretical econometrics and applied econometrics.

a) Theoretical econometrics

It includes the development o appropriate methods for the measurement of economic relationships. These econometric methods may further be classified into two as single and simultaneous equation techniques.

1) Single-equation techniques: are methods that are applied to one relationship at a time.

2) Simultaneous-equation techniques: are methods applied to all the relationships of the model simultaneously.

b) Applied econometrics

It includes the application of econometric methods to specific branches of economic theory.

It examines the problems encountered and the findings of applied research in the fields of: demand, supply, production, investment, consumption, and other sectors of the economy. Applied

econometrics involves the application of the tools of theoretical econometrics for the analysis of economic phenomena and forecasting economic behavior.

Chapter 2. Correlation Theory

Key concepts:

- Types of correlation
- Methods of studying correlation

a. Scatter diagram
b. Karl pearson's coefficient of correlation
c. Spearman's Rank correlation coefficient

2.1. Basic Concepts of Correlation

Correlation: The degree of relationship between the variables under consideration is measured through the correlation analysis. The measure of correlation is called the correlation coefficient. The degree (strength) of relationship is expressed by the coefficient which ranges from correlation ($-1 \leq r \leq +1$). The direction of change is indicated by a sign.

The correlation analysis enables us to have an idea about the degree & direction of the relationship between the two variables under study. Correlation is a statistical tool that helps to measure and analyse the degree of relationship between two variables. Correlation analysis deals with the association between two or more variables.The linear correlation coefficient *r* measures the strength of the linear relationship between paired *x*- and *y*- quantitative values in a sample.

What are correlation and causation and how are they different?

Two or more variables considered to be related, in a statistical context, if their values change so that as the value of one variable increases or decreases so does the value of the other variable (although it may be in the opposite direction). For example, for the two variables "hours worked" and "income earned" there is a relationship between the two if the increase in hours worked is associated with an increase in income earned. If we consider the two variables "price" and "purchasing power", as the price of goods increases a person's ability to buy these goods decreases (assuming a constant income).

Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables. A correlation between variables, however, does not automatically mean that the change in one variable is the cause of the change in the values of the other variable.
Causation indicates that one event is the result of the occurrence of the other event; i.e. there is a causal relationship between the two events. This is also referred to as cause and effect.

Theoretically, the difference between the two types of relationships are easy to identify-an action or occurrence can cause another (e.g. smoking causes an increase in the risk of developing lung cancer), or it can correlate with another (e.g. smoking is correlated with alcoholism, but it does not cause alcoholism). In practice, however, it remains difficult to clearly establish cause and effect, compared with establishing correlation.

Types of Correlation: Type I

- **Correlation (Positive and negative**)

Positive Correlation: The correlation is said to be positive correlation if the values of two variables changing with same direction. **Example:** Price and quantity supplied.

Negative Correlation: The correlation is said to be negative correlation when the values of variables change with opposite direction. **Example:** Price and quantity demanded.

Direction of the Correlation

Positive relationship: Variables change in the same direction. As **X** is increasing, **Y** is increasing. As **X** is decreasing, **Y** is decreasing. Example: As height increases, so does weight.
Negative relationship: Variables change in opposite directions. As **X** is increasing, **Y** is decreasing. As **X** is decreasing, **Y** is increasing. **Example:** As TV time increases, grades decrease. Indicated by sign; (+) or (-)
More examples. Positive relationships: Water consumption and temperature, study time and grades. **Negative relationships:** Alcohol consumption and driving ability.

Types of Correlation: Type II

- **Correlation (simple and multiple**)
- **Multiple (partial and total)**

a. **Simple correlation:** Under simple correlation problem only two variables are studied.

b. **Multiple correlations:** Under Multiple Correlation three or more than three variables are studied.

Example: Quantity demanded = f (Price, Price of other goods, expectations, taxes on and subsidies to consumers).

Partial correlation: analysis recognizes more than two variables but considers only two variables keeping the others constant. **Total correlation** is based on all the relevant variables, which is normally not feasible.

Types of Correlation: Type III

Correlation (linear and non-linear)

Linear correlation: Correlation is said to be linear when the amount of change in one variable tends to bear a constant ratio to the amount of change in the other. The graph of the variables having a linear relationship will form a straight line.

Example: X = 1, 2, 3, 4, 5, 6, 7, 8,

Y = 5, 7, 9, 11, 13, 15, 17, 19,

Y = 3 + 2x

Non Linear correlation: The correlation would be non-linear if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable.

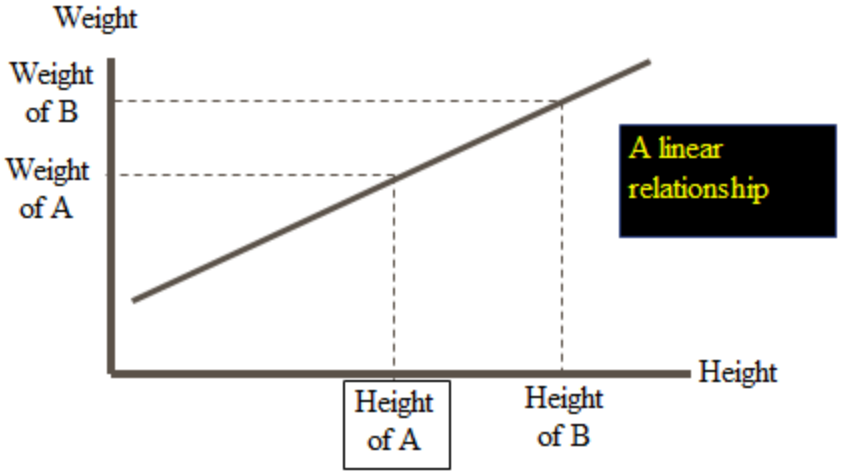
2.2. Coefficient of Linear Correlation

Methods of Studying Correlation: Scatter diagram method, graphic method, Karl Pearson's coefficient of correlation, and Spearman's Rank coefficient of correlation.

A. Scatter Diagram Method

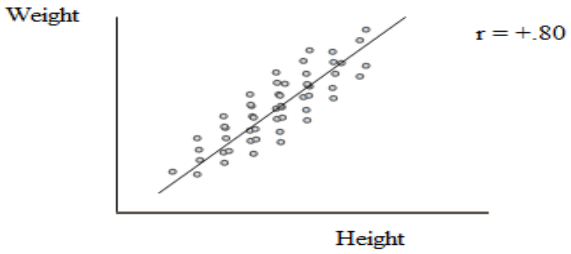
Scatter Diagram is a graph of observed plotted points where each point represents the values of X and **Y** as a coordinate. It portrays the relationship between these two variables graphically.

Degree of Correlation
A perfect positive correlation

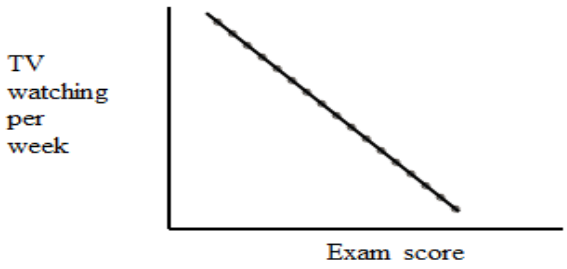


High Degree of positive correlation

- **Positive relationship**

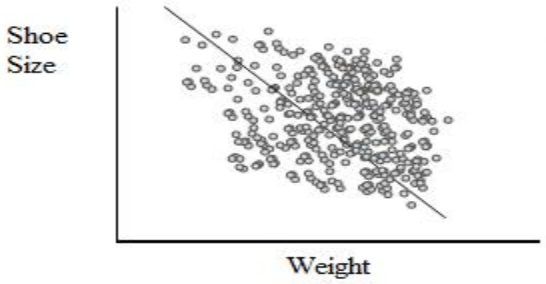


- **Perfect Negative Correlation**



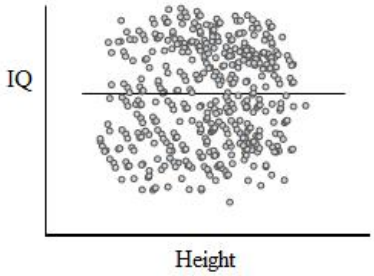
r = -1.0

- **Weak negative Correlation**



r = - 0.2

- **No Correlation (horizontal line)**



r = 0.0

Advantages of Scatter Diagram

It is simple and non-mathematical method, not influenced by the size of extreme item (outliers) and it is the first step in investing the relationship between two variables.

Disadvantage of scatter diagram: It cannot adopt an exact degree of correlation.

B. Karl Pearson's Coefficient of Correlation

Pearson's 'r' is the most common correlation coefficient. Karl Pearson's coefficient of correlation is denoted by 'r'. The coefficient of correlation measures the degree of linear relationship between two variables say X and Y. Karl Pearson's Coefficient of Correlation is denoted by '**r**' **-1 ≤ r ≤ +1.** Degree of correlation is expressed by the value of the coefficient. Direction of change is indicated by the **sign** (**- ve**) or (**+ ve**).

- **When deviation is taken from actual mean:**

$$r(x, y) = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

- **When deviation is taken from an assumed mean:**

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

Notation for the Linear Correlation Coefficient

n represents the number of pairs of data present.

$\sum$ denotes the addition of the items indicated.

$\sum x$ denotes the sum of all x-values.

$\sum x^2$ indicates that each x-value should be squared and then those squares added.

$(\sum x)^2$ indicates that the x-values should be added and the total then squared.

$\sum xy$ indicates that each x-value should be first multiplied by its corresponding y-value.

- After obtaining all such products, find their sum.

r represents linear correlation coefficient for a sample.

ρ represents linear correlation coefficient for a population.

Procedure for computing the correlation coefficient

- Calculate the mean of the two series 'x' &'y'
- Calculate the deviations 'x' &'y' in two series from their respective mean.
- Square each deviation of 'x' &'y' then obtain the sum of the squared deviation i.e.$\sum x^2$ and $\sum y^2$

- Multiply each deviation under x with each deviation under y and obtain the product of 'xy'.
- Then obtain the sum of the product of x , y i.e. $\sum xy$
- Substitute the value in the formula.

Requirements

1. The sample of paired (*x*, *y*) data is a random sample of independent quantitative data.
2. Visual examination of the scatter plot must confirm that the points approximate a straight-line pattern.
3. The outliers must be removed if they are known to be errors.

➢ The effects of any other outliers should be considered by calculating *r* with and without the outliers included.

Example: Calculating *r*

Using the simple random sample of data below, find the value of *r*.

Table 2.1.

	X	Y	XY	X^2	Y^2
	3	5	15	9	25
	1	8	8	1	64
	3	6	18	9	36
	5	4	20	25	16
Total	12	23	61	44	141
	↑	↑	↑	↑	↑
	$\sum X$	$\sum Y$	$\sum XY$	$\sum X^2$	$\sum Y^2$

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

$$r = \frac{4(61) - (12)(23)}{\sqrt{4(44) - (12)^2}\sqrt{4(141) - (23)^2}} = \frac{-32}{33.466} = -0.956$$

Given $r = -0.956$, if we use a 0.05 significance level,

- We conclude that there is a linear correlation between *x* and *y* since the absolute value of *r* exceeds the critical value of 0.950.
- However, if we use a 0.01 significance level,
- We do not conclude that there is a linear correlation because the absolute value of *r* does not exceed the critical value of 0.999.

Interpretation of Correlation Coefficient (r)

- The value of correlation coefficient '**r**' ranges from **-1** to **+1**
- If **r = +1**, then the correlation between the two variables is said to be **perfect** and **positive**
- If **r = -1**, then the correlation between the two variables is said to be **perfect** and **negative**
- If **r = 0**, then there exists **no correlation** between the variables

Assumptions of Pearson's Correlation Coefficient

- There is linear relationship between two variables,
- ✓ i.e. when the two variables are plotted on a scatter diagram a straight line will be formed by the points.
- Cause and effect relation exists between different forces operating on the item of the two variable series.

Advantages of Pearson's Coefficient

- It summarizes in one value, the degree of correlation & direction of correlation also.

Limitation of Pearson's Coefficient

- Always assume linear relationship
- Interpreting the value of r is difficult.
- Value of correlation coefficient is affected by the extreme values.
- Time consuming methods

Coefficient of Determination

The convenient way of interpreting the value of correlation coefficient is to use square of coefficient of correlation which is called **Coefficient of Determination**. The Coefficient of Determination = $\mathbf{r^2}$.

- **Suppose: r = 0.9, r^2 = 0.81.** This would mean that **81%** of the variation in the dependent variable has been explained by the independent variable.

The maximum value of $\mathbf{r^2}$ **is** 1 because it is possible to explain all of the variation in y but it is not possible to explain more than all of it.

$$Coefficient\ of\ determination = \frac{Explained\ variation}{Total\ variation}$$

Coefficient of Determination: An example

- **Suppose: r = 0.60**
  r = 0.30
- It does not mean that the first correlation is twice as strong as the second the '**r**' can be understood by computing the value of $\mathbf{r^2}$.

When **r = 0.60** **r^2 = 0.36** **……** (1)
r = 0.30 **r^2 = 0.09** **……** (2)

- This implies that in the first case **36%** of the total variation is explained whereas in second case **9%** of the total variation is explained .

C) **Spearman's Rank Coefficient of Correlation (R)**

The above correlation coefficient is used if the variables are quantitative; however, some variables may be qualitative and cannot be measured numerically. For example, sex, religion, educational level and profession are qualitative variables and it is impossible to compute correlation coefficient with the formula developed in the above section. In this case, we can use another formula, which is called rank correlation coefficient (Spearman's correlation coefficient).

In this method, we rank the observations in a specific sequence in order of size or importance or other thing in ascending or descending order and measure the relationship between their ranks instead of actual values.

- Spearman Rank correlation is used:
- ✓ When statistical series in which the variables under study are not capable of **quantitative** measurement,
- ✓ If it can be arranged in **serial order**,
- ➢ In such situations Pearson's correlation coefficient cannot be used

It is a non-parametric measure of correlation. This procedure makes use of the two sets of ranks that may be assigned to the sample values of **X** and **Y**. Spearman Rank correlation coefficient could be computed in the following cases:

- ➢ Both variables are qualitative.
- ➢ Both variables are qualitative ordinal.
- ➢ One variable is quantitative and the other is qualitative ordinal.

Procedures

1. Rank the values of **X** from **1** to **n** where **n** is the numbers of pairs of values of **X** and **Y** in the sample.
2. Rank the values of **Y** from **1** to **n**.
3. Compute the value of d_i for each pair of observation by subtracting the rank of $\mathbf{Y_i}$ from the rank of $\mathbf{X_i}$.

4. Square each d_i and compute $\sum \mathbf{di^2}$ which is the sum of the squared values.
5. Apply the following formula

$$r_s = 1 - \frac{6\sum (di)^2}{n\left(n^2 - 1\right)}$$

Where,
r = Rank correlation coefficient
D = Difference of rank between paired item in two series.
N = Total number of observation.

- The value of r_s denotes the magnitude and nature of association giving the same interpretation as simple r.

Interpretation of Rank Correlation Coefficient

The value of rank correlation coefficient, R ranges from -1 to +1. If r = +1, then there is complete agreement in the order of the ranks and the ranks are in the same direction. If r = -1, then there is complete agreement in the order of the ranks and the ranks are in the opposite direction. If r = 0, then there is no correlation.

Example 1: Suppose the following table shows how ten singers were ranked according to their performance by two judges. We want to find out whether there is agreement among the two judges in ranking the ten singers.

Singers	A	B	C	D	E	F	G	H	I	J
Judge 1	9	5	8	1	4	10	7	2	3	6
Judge 2	10	6	9	2	3	7	8	1	5	4

To decide whether there is an agreement or not among the two judges, we have to compute the rank correlation coefficient. In the above example the difference between the two rankings (**D**) made by the two judges is given as follows:

D	-1	-1	-1	-1	1	3	-1	1	-2	2
D^2	1	1	1	1	1	9	1	1	4	4

$$Rank\ correlation = 1 - \frac{6\sum (d_i)^2}{n\left(n^3 - 1\right)} = 1 - \left[\frac{6(24)}{10\left(10^2 - 1\right)}\right] = 0.855$$

- The high value of the rank correlation coefficient indicates that there is an agreement by the two judges in rating the singers.

Example 2: In a study of the relationship between level of education and income the following data was obtained. Find the relationship between them and comment.

Sample numbers	Level of education (X)	Level of income (Y)
A	Preparatory	25
B	Primary	10
C	University	8
D	Secondary	10
E	Secondary	15
F	Illiterate	50
G	University	60

Answer:

	X	Y	Rank (X)	Rank (Y)	di	di^2
A	Preparatory	25	5	3	2	4
B	Primary	10	6	5.5	0.5	0.25
C	University	8	1.5	7	-5.5	30.25
D	Secondary	10	3.5	5.5	-2	4
E	Secondary	15	3.5	4	-0.5	0.25
F	Illiterate	50	7	2	5	25
G	University	60	1.5	1	0.5	0.25

$\sum$ **di^2 = 64**

Comment: $r_s = 1 - \frac{6 \times 64}{7 \times 48} = -0.1$

- There is an indirect weak correlation between level of education and income.

b) Problems where Ranks are not given:

If the ranks are not given, then we need to assign ranks to the data series. The lowest value in the series can be assigned rank 1 or the highest value in the series can be assigned rank 1. We need to follow the same scheme of ranking for the other series. Then calculate the rank correlation coefficient in similar way as we do when the ranks are given.

Merits Spearman's Rank Correlation

This method is simpler to understand and easier to apply compared to karl pearson's correlation method. This method is useful where we can give the ranks and not the actual data (qualitative term). This method is used where the initial data is in the form of ranks.

Limitation Spearman's Correlation

It cannot be used for finding out correlation in a grouped frequency distribution. This method should be applied where **N** exceeds **30**.

Advantages of Correlation studies

- Show the amount (strength) of relationship present
- Can be used to make predictions about the variables under study.
- Can be used in many places, including natural settings, libraries, etc.
- Easier to collect co relational data

Disadvantages of correlation studies

- Cannot assume that a cause-effect relationship exists
- Little or no control (experimental manipulation) of the variables is possible
- Relationships may be accidental or due to a third, unmeasured factor common to the 2 variables that are measured

Chapter 3: Simple Linear Regression Models

3.1. Introduction (Basic Concepts and Assumptions)

Important points before we start a regression analysis:

Most important thing in deciding whether or not there is a relationship between **X** and **Y** is to have a **systematic model** that is based on logical reasons. Investigate the nature of the relationship between **X** and **Y.** Use **scatter diagram**, **covariance**, **coefficient of correlation**.

Remember that regression is not **an exact** or **deterministic** mathematical equation. It is a **behavioral relationship** that is subject to **randomness**. Remember that **X** is not the only thing that explains the behavior of **Y**. There are other factors that you may not have information about. All you are trying to do is to have an estimate of the relationship using the **best linear fit** possible.

The simplest economic relationship is represented through a two-variable model (also called the simple linear regression model) which is given by:

$$\mathbf{Y = a + bX.}$$

Where **a** and **b** are unknown parameters (also called regression coefficients) that we estimate using sample data.

- **a** is the **intercept** of the model
- **b** is the slope of the model
- **Y** is called the dependent variable
- **X** is called the independent (explanatory, exogenous) variable

Example: suppose the relationship between expenditure (**Y**) and income (**X**) of households is expressed as:

$$\mathbf{Y = 0.6X+30}$$

Here, on the basis of income, we can predict expenditure. For instance, if the income of a certain household is 150 Birr, then the estimated expenditure will be:

Mean expenditure = 0.6 (150) + 30 = 120 Birr

Note that since expenditure is estimated on the basis of income, expenditure is the dependent variable and income is the independent variable.

The error term

Consider the above model: **Y = 0.6X+30.** This functional relationship is **exact**, that is given income we can determine the exact expenditure of a household. But in reality this rarely happens, different households with the same income are not expected to spend equal amounts due to habit persistence, geographical and time variation, etc.

Thus, we should express the regression model as:

$$\mathbf{Y_i = \alpha + \beta X_i + \epsilon_i}$$

Where ϵ is the **random error term** (also called **disturbance term**)

Generally the reasons for including the error term include:

- **Omitted variables:** a model is a simplification of reality. It is not always possible to include all relevant variables in a functional form. For instance, we may construct a model

relating demand and price of a commodity. But demand is influenced not only by price: income, price of other goods, expectations, taste and several other variables also influence it. The omission of these variables from the model introduces an error.

- **Measurement error:** inaccuracy in collection and measurement of sample data.
- **Sampling error:** consider a model relating consumption (**Y**) with income (**X**) of households. The sample we randomly choose to examine the relationship may turn out to be predominately poor households. In such cases, our estimation of **α** and **β** from this sample may not be as good as from a balanced sample group.

Note that the size of the error is not fixed. **It is non-deterministic** or **stochastic**. This in turn implies that **Y** is also stochastic. On the other hand, the variable **X** is deterministic or non-stochastic.

In the regression model $\mathbf{Y_i = \alpha + \beta X_i + \epsilon_i}$, the values of the parameters **α** and **β** are not known. When they are estimated from a sample of size **n**, we obtain the **sample regression line** given by:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}\,\hat{X}_i \qquad i = 1, 2, ...n$$

Where **α and β** are estimated by $\hat{\alpha}$ *and* $\hat{\beta}$*, respectively.* $\hat{Y}$ *is the estimated value of* Y. Our objective is not only to estimate **α** and **β**, but also to draw inferences about their values. For this purpose, we need some assumptions.

Assumptions of the classical linear regression model (or OLS model)

1**.** The relationship between Y and X is linear, that is, $\mathbf{Y_i = \alpha + \beta X_i + \epsilon_i}$

Violations of this assumption may occur as a result of:

- Nonlinear relationship between regressors/predictors (**X_is**)
- Changing parameters (when the parameters are not constant).

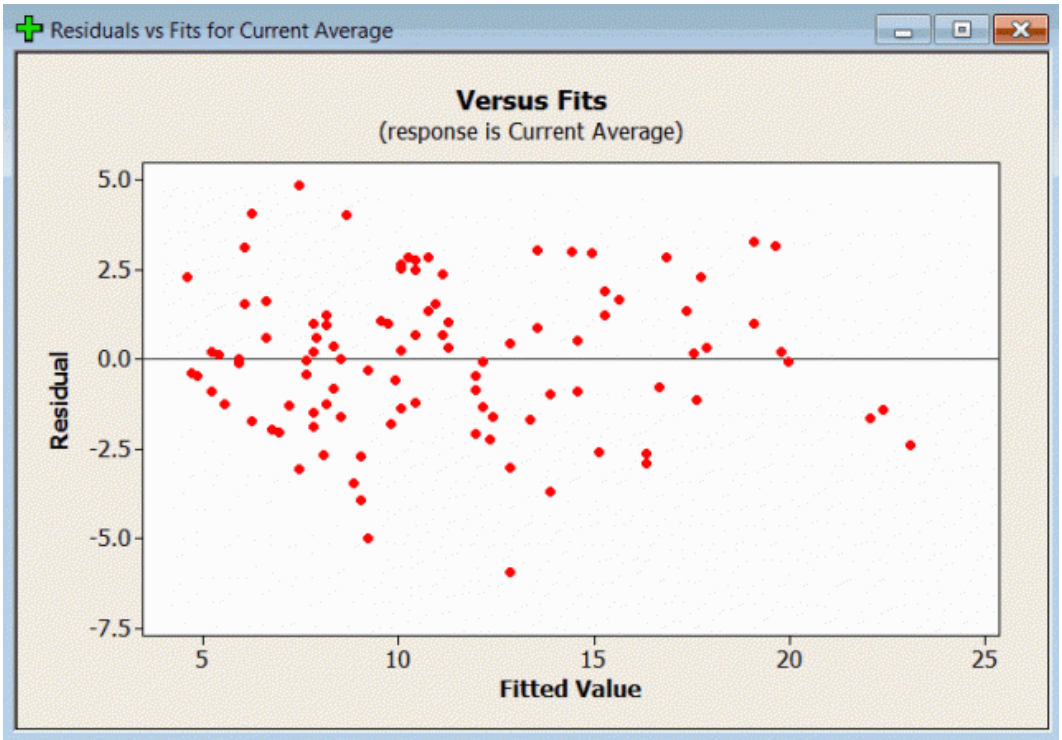
➢ A regression equation (or function) is linear when it is linear in the parameters (**b_is**)

2. The error has zero expected value, that is, $\mathbf{E(\varepsilon_i) = 0}$

3. The error terms have constant variance, that is $\mathbf{E}(\varepsilon_i^2) = \sigma^2$ for all i

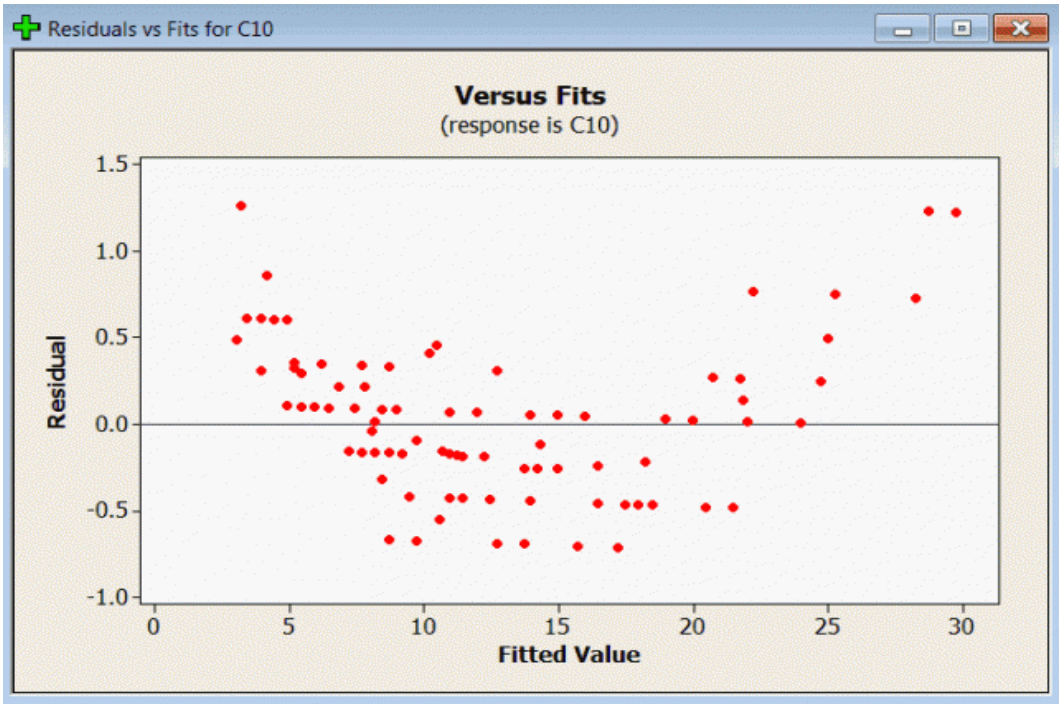
When you run a regression analysis, the variance of the error terms must be constant, and they must have a mean of zero. If this is not the case, your model may not be valid. To check these

assumptions, you should use residuals versus fitted values plot. Below is an example of the plot from the regression analysis. If, for example, the residuals increase or decrease with the fitted values in a pattern, the errors may not have constant variance.



The points on the plot above appear to be randomly scattered around zero, so assuming that the error terms have a mean of zero is reasonable. The vertical width of the scatter does not appear to increase or decrease across the fitted values, so we can assume that the variance in the error terms is constant.

But what if this was not the case? What if we did notice a pattern in the plot? Look at the following plot.



There is definitely a noticeable pattern here! The residuals (error terms) take on positive values with small or large fitted values, and negative values in the middle. The width of the scatter seems consistent, but the points are not randomly scattered around the zero line from left to right. This graph tells us we should not use the regression model that produced these results.

So what to do? There's no single answer, but there are several options. One approach is to adjust your model: adding a squared term to the model could solve the issue with the residuals plot. Alternatively, Minitab has a tool that can adjust the data so that the model is appropriate and will yield acceptable residual plots. It is called a Box-Cox transformation, and it is easy to use! First just open the General Regression dialog (**Stat** > **Regression** > **General Regression**). Then click the Box-Cox button.

4. The random variables ε_i are statistically independent of each other, that is, E ($\varepsilon_i \quad \varepsilon_j$) = 0 for i $\neq$ j.

5. The independent (explanatory) variable **X** is non-stochastic, that is its values are fixed. This implies that the values of **X** are not correlated with the error term E (X_i, ε_i) = 0 i = 1, 2, 3…n

6. The error term is normally distributed. $u_i \left(0, \sigma^2\right)$

The Gauss-Markov Theorem: given the above assumptions of the **CLRM**,

- The parameter estimators (**b_1**) and (**b_2**) are best (most efficient) Linear Unbiased estimators (**BLUE**) of the true population parameters (**B_1**) and (**B_2**).
    - **The BLUE criterion**
    - **B** for **Best** (Minimum error)
    - **L** for Linear (The form of the relationship)
    - **U** for Un-bias (does the parameter truly reflect the effect?)
    - **E** for Estimator

3.2. The Least Squares Criteria

There are many techniques in econometrics and statistics that use the least squares criterion. In regression techniques this criterion is of immense importance. Why should a criterion be used at all? The answer to this question is quite obvious: One has to have an objective measure for discrepancies between the **estimated values** (generated by the statistical model) and the (true) **observed values**. In fact we wish to create mathematical models of our surrounding world in order to be able to describe it, to draw conclusions from it, to forecast future behavior of some (economic) phenomena, and to explain why certain things happened in the past.
For obvious reasons these mathematical models are **not deterministic** but instead, **probabilistic or stochastic**. This is the reason why we have a need for a good criterion to decide whether our model does describe the real world as good as possible. Since we cannot hope for a model to describe a real phenomenon perfectly, the only thing we can do is to design a method for getting as close to the real behavior as possible. This can be achieved by **minimizing** the error of the mathematical model. The most obvious way to express the error made by a probabilistic model is to calculate the sum of the **deviations between the forecasted values and the real values**:

3.3. Normal Equations of OLS

3.3.1. The Meaning of Regression

Regression analysis is concerned with the study of the relationship between one variable called explained (dependent) variable and one or more of other variables called independent (explanatory) variables.
Example: The law of demand states that quantity demanded depends on the price of the commodity and various other factors.

- The objectives of regression analysis may be:

1) To estimate the mean value of the dependent variable given the values of the independent variables.

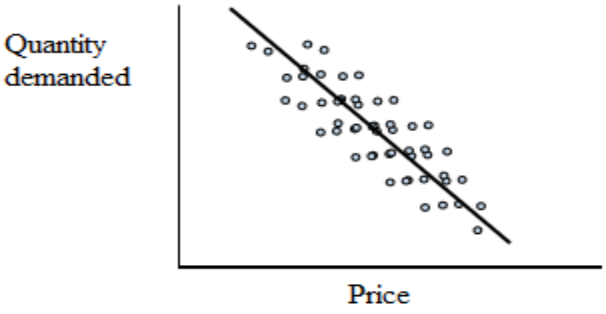
2) To test hypothesis about the nature of the dependence

3) To forecast the mean value of the dependent variable, given the values the independent variables.

The Population Regression Function (PRF)

Price (X)	Quantity Demanded (Y)	Number of consumers	Average Y demanded
1	45,46,47,48,49,50,51	7	48
2	44,45,46,47,48	5	46
3	40,42,44,46,48	5	44
4	35,38,42,44,46,47	6	42
5	36,39,40,42,43	5	40
6	32,35,37,38,39,42,43	7	38
7	32,34,36,38,40	5	
8	31,32,33,34,35,36,37	7	34
9	28,30,32,34,36	5	32
10	29,30,31	3	30

- From the table, at price of **1 Birr** there are seven consumers who purchase the good in quantities ranging from **45** to **51 units.**
- **These can be plotted on a scatter diagram**



From the scatter diagram above one can see that **Y** generally decreases as **X** increases and vice versa. When we draw a line through the mean values, the line obtained is called the **population regression line.** The population regression line gives the **average** or **mean** values of the dependent variable **Y** corresponding to each value of the independent variable, **X**.

- Mathematically, the **PRF** is given as:

$$E(Y / X_i) = B_1 + B_2 X_i \quad \text{.....} \quad 3.1$$

It represents the expected value of Y corresponding to or conditional up on a given value of X.

In the above table E (Y/X_i =2) = 46. E (Y/X_i) is a function of X. This means the dependence of Y on X can be defined simply as the mean of the distribution of Y values which has the given X. In other words the population regression line is a line that passes through the conditional means of Y.

- B_1 and B_2 are called the parameters or the regression coefficients.
- B_1 and B_2 are the intercept and slope of the function.
- In regression we are interested in examining the behavior of the dependent variable conditional upon given values of the independent variable (s), and our approach to regression analysis could be termed as conditional regression analysis.
- However, the conditional regression equation E (Y/X_i) is sometimes simply written as E (Y).

Stochastic Specification of the Population Regression Function

The average of Y corresponding to X = 1 is 48. However, if we peak a consumer at random from the seven consumers corresponding to this price, the quantity demanded by that consumer will not necessarily equal 48. The best way to explain this is that individuals demand is equal to the **average** for that group **plus** or **minus** some quantity.

$$Y_i = B_1 + B_2 X_i + u_i \quad \text{.....} \quad 3.2$$

Where, $\mathbf{u_i}$ is the stochastic or **random error term or** the **error term.**

The error term is a random variable, for its value cannot be controlled or known a priori. It can be characterized by its probability distribution.

From the above equation $B_1 + B_2 X_i$ is called the deterministic (systemic) component and u_i is the non-systemic (random) component determined by factors other than X_i.

- **The nature of the stochastic error term**

1) The error term may represent the influence of those variables that are not explicitly included in the model.
2) Some intrinsic randomness might occur because of human behavior.
3) It might represent measurement errors. E.g. data rounding.
4) The principle of Occam's razor: keep the regression model as simple as possible.

The combined influence of those variables not included might be small and non-systemic and could be incorporated in the error term.

- **The Sample Regression Function (SRF)**

To fit the **PRF** we need the entire population data at our disposal. Instead, we have a **sample** data. Then the task is to estimate the **PRF** on the basis of the sample information.

Suppose that we have not seen the table above about the population demand for the good and have the following sample information.

A random sample from the above population (sample 1)

Price (X)	1	2	3	4	5	6	7	8	9	10
Qt.D (Y)	49	45	44	39	38	37	34	33	30	29

A random sample from the above population (sample 2)

Price (X)	1	2	3	4	5	6	7	8	9	10
Qt.D (Y)	51	47	46	42	40	37	36	35	32	30

From the above two sample data we cannot **accurately** estimate the **PRF** because of sampling fluctuation (sampling error).

Analogous to PRF, we can develop the concept of sample regression function (**SRF**).

$$\hat{Y}_i = b_1 + b_2 X_i + e_i \quad \text{..... 3.3}$$

Where, $\hat{Y}_i$ is the estimator of E (Y/X_i), the estimator of the population conditional mean.

b_1 *is the estimator of* B_1

b_2 *is the estimator of* B_2

e_i *is the estimator of* u_i

- For a given X_i we have one sample observation Y_i.
- In terms of the observed Y_i it could be expressed as:

$$Y_i = \hat{Y}_i + e_i \quad \text{..... 3.4}$$

- In terms of the PRF it can be expressed as:

$$Y_i = E(Y / X_i) + u_i \quad \text{..... 3.5}$$

3.3.2. Estimation of Parameters: The Ordinary Least Squares (OLS) Method

Several methods of obtaining the SRF as an estimator of the true PRF exist. However, the most frequently used is that of OLS or sometimes called least squares (LS). OLS estimator has the desirable property that it makes the residuals as small as possible.

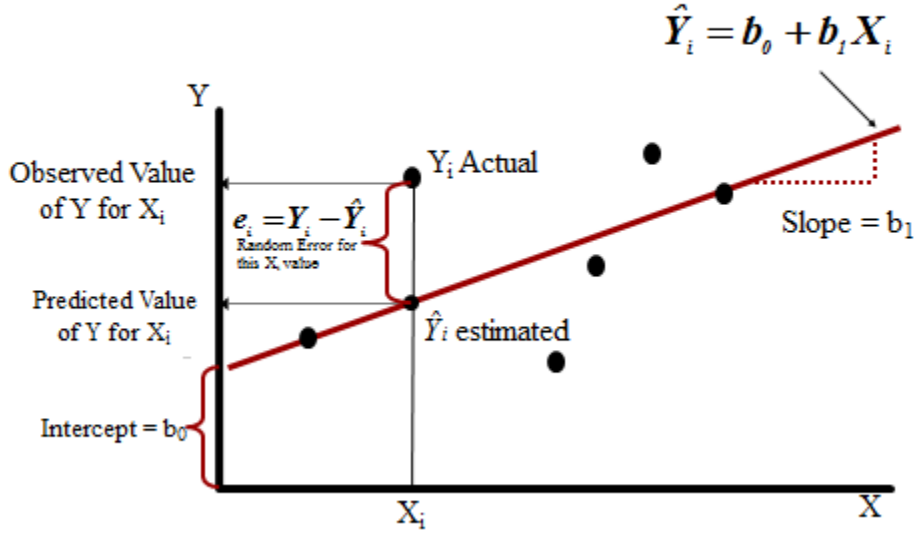
$Y - \hat{Y} = \varepsilon$ or the difference between the actual and predicted values of the dependent variable as small as possible or $Y_i - b_1 - b_2 X_i$ should be as small as possible.

- OLS states that b_1 and b_2 should be chosen in such a way that the residual sum of squares (RSS) or $\sum e_i^2$ is as small as possible.

$$\min \sum e_i^{\,2} = \min \sum \left(Y_i - \hat{Y} \right)^2 = \min \sum \left(Y_i - b_0 - b_1 X_i \right)^2$$

- The estimated values of b_0 and b_1 by OLS

➢ They are the only possible values of b_0 and b_1 that minimize the sum of the squared differences between Y and $\hat{Y}$.



Derivation of the Ordinary Least Squares Estimates

In deriving the parameter values that minimize the difference we can use the technique of differential calculus.

- **The goal here is to minimize**

$$\sum\left(Y_i - \hat{Y}_i\right)^2 = \sum e_i^2 \quad \text{..... 3.6}$$

$$= \sum\left(Y_i - b_1 - b_2 X_i\right)^2 \quad \text{..... 3.7}$$

If we take the derivatives with respect to each parameter, we could get the **normal equations**.

$$\frac{\partial}{\partial b_1}\sum\left(Y_i - b_1 - b_2 X_i\right)^2 = 2\sum\left(Y_i - b_1 - b_2 X_i\right)(-1) = 0 \quad \text{..... 3.8}$$

$$\frac{\partial}{\partial b_2}\sum\left(Y_i - b_1 - b_2 X_i\right)^2 = 2\sum\left(Y_i - b_1 - b_2 X_i\right)(-X_i) = 0 \quad \text{..... 3.9}$$

- **If we divide the equations by -2, we obtain:**

$$\sum \left(Y_i - b_1 - b_2 X_i\right) = 0 \quad \text{..... 3.10}$$

$$\sum X_i \left(Y_i - b_1 - b_2 X_i\right) = 0 \quad \text{..... 3.11}$$

$$\Rightarrow \sum Y_i = b_1 n + b_2 \sum X_i \quad \text{..... 3.12}$$

$$\Rightarrow \sum X_i Y_i = b_1 \sum X_i + b_2 \sum X_i^2 \quad \text{..... 3.13}$$

- Then we can solve for the parameters by multiplying equation 3.12 by $\sum X_i$ and equation 3.13 by **n:**

$$\sum X_i \sum Y_i = b_1 n \sum X_i + b_2 \left(\sum X_i\right)^2 \quad \text{..... 3.14}$$

$$n \sum X_i Y_i = b_1 n \sum X_i + b_2 n \sum X_i^2 \quad \text{..... 3.15}$$

- **Subtract equation 3.14 from equation 3.15.**

$$n \sum X_i Y_i - \sum X_i \sum Y_i = b_2 \left[n \sum X_i^2 - \left(\sum X_i\right)^2 \right] \quad \text{..... 3.16}$$

$$b_2 = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - \left(\sum X_i\right)^2} \quad \text{..... 3.17}$$

- **By using equation 3.19**

$$b_1 = \frac{\sum Y_i}{n} - b_2 \frac{\sum X_i}{n} = \bar{Y} - b_2 \bar{X} \quad \text{..... 3.18}$$

Example: The following results have been obtained from a sample of 11 observations on the values of sales (**Y**) of a firm and the corresponding prices (**X**).

$$\bar{X} = 519 \qquad \bar{Y} = 218$$

$$\sum X_i^2 = 3134543 \qquad \sum X_i Y_i = 1296836 \qquad \sum Y_i^2 = 539512 \qquad n = 11$$

- **Estimate the regression of sales on prices and interpret your results.**

$$b_2 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2}$$

But

$$\bar{X} = \frac{\sum X_i}{n} \Rightarrow \sum X_i = \bar{X}\, n, \text{ and } \left(\Sigma X_i\right)^2 = \left(n\bar{X}\right)^2$$

$$\bar{Y} = \frac{\sum Y_i}{n} \Rightarrow \sum Y_i = \bar{Y}\, n$$

$$b_2 = \frac{11(1296836) - (11\times 519)(11\times 218)}{11\times(3134543) - (11\times 519)^2} = 0.304$$

$$b_1 = \bar{Y} - b_2\,\bar{X} \Rightarrow b_1 = 218 - 0.304\,(519) = 60.22$$

$$\hat{Y} = 60.22 - 0.304X_i$$

Interpretation: If price increases by one unit, sales would increase by 0.3 *unit.*

- **Using deviations to derive the parameters of OLS**

Let $x_i = \left(X_i - \bar{X}\right)$

$$y_i = \left(Y_i - \bar{Y}\right) \quad \text{..... 3.19}$$

$$b_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum\left(X_i - \bar{X}_i\right)\left(Y_i - \bar{Y}\right)}{\sum\left(X_i - \bar{X}\right)^2} = \frac{\Sigma X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2} \quad \text{..... 3.20}$$

Because

$$\sum\left(Y_i-\bar{Y}\right)^2=\sum Y_i^{\,2}-n\bar{Y}^{\,2} \quad \text{.....} \quad 3.21$$

$$\sum\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)=\sum X_iY_i-n\bar{X}\bar{Y} \quad \text{.....} \quad 3.22$$

$$\sum\left(X_i-\bar{X}\right)^2=\sum X_i^{\,2}-n\bar{X}^{\,2} \quad \text{.....} \quad 3.23$$

And

$$b_1=\bar{Y}-b_2\bar{X} \quad \text{.....} \quad 3.24$$

3.4. Estimation of Elasticities from Regression Equations

- Note from our previous discussion that given the estimated function:

$\hat{Y}=b_1+b_2X_i$ the intercept is **b_1** and the slope is **b_2**.

- Using the concept of (partial) derivative **b_2** is given by:

$$b_2=\frac{\partial Y_i}{\partial X_i}$$

- This shows the rate of change in **Y** as **X** changes by small amount.
- If the estimated function is a linear demand function given by:

$$Q_d=b_1-b_2P$$

- The coefficient **b_2** is the component of the price elasticity of demand.
- Remember from your previous studies that the average price elasticity of demand is given by:

$$e_d=\frac{\partial Q_d}{\partial P}\times\frac{\bar{P}}{\bar{Q}}$$

- In the estimated linear demand function $b_2=\frac{\partial Q_d}{\partial P}$ and elasticity will be given as:

$$e_d = b_2 \times \frac{\bar{P}}{\bar{Q}}$$

Where, $\bar{P}$ is the average price in the sample and $\bar{Q}$ is the average value of quantity demanded.

Suppose the estimated demand function for a certain commodity is given by: **Q = 100-20P**. If the average **price** of the commodity = **Birr 4**. And the average **quantity** demanded is **100** units.

Compute the price elasticity of demand.

Solution: if the average price and average quantity demanded are given, using the coefficients of the estimated demand curve we can compute elasticity of demand.

$$e_d = b_2 \times \frac{\bar{P}}{\bar{Q}} = -20 \times \frac{4}{100} = -0.8$$

Since the absolute value, the price elasticity of demand is less than one, the demand for the commodity is **inelastic.**

3.5. Coefficient of Correlation and Determination

What is Correlation?

The term "**correlation**" refers to a measure of the strength of association between **two** variables. If the two variables increase or decrease together, they have a positive correlation. If, increases in one variable are associated with decreases in the other, they have a negative correlation.

For two quantitative variables X and Y, for which **n** pairs of measurements ($\mathbf{x_i}$, $\mathbf{y_i}$) are available Pearson's correlation coefficient (**r**) gives a measure of the **linear** association between X and Y.

The formula is given below for reference.

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

- If X and Y are perfectly positively correlated, r = 1
- If there is absolutely no association, r = 0
- If X and Y are perfectly positively correlated, r = -1
- Thus $-1 \le r \le 1$.

- The closer r is to +1 or -1, the greater is the strength of the association.

Coefficient of Determination

It is often difficult to interpret **r** without some familiarity with the expected values of **r**. A more appropriate measure to use when interest lies in the dependence of **Y** on **X**, is the **Coefficient of Determination, R^2**. It measures the **proportion of variation** in **Y** that is explained by **X**, and is often expressed as a percentage.

How good is the Model's prediction Power?

Total variation is made up of two parts:

SST = SSR + SSE

Where,

SST = Total sum of squares; SSR = Regression sum of squares; SSE = Error sum of squares

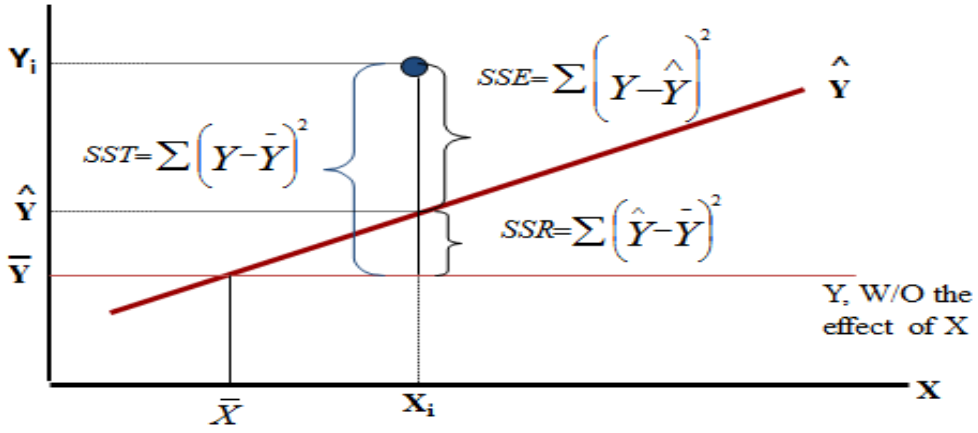
$$SST = \sum (Y_i - \bar{Y})^2 \qquad SSR = \sum (\hat{Y}_i - \bar{Y})^2 \qquad SSE = \sum (Y_i - \hat{Y}_i)^2$$

Where,

$\bar{Y}$ = Average value of the dependent variable

Y_i = Observed values of the dependent variable

$\hat{Y}$ = Predicted value of Y for the given $\mathbf{X_i}$ value



The coefficient of determination: It is the portion of the total variation in the dependent variable (**Y)** that is explained by variation in the independent variable (**X).** The coefficient of determination is also called **r-squared** and is denoted as **r^2.**

$$r^2 = \frac{SSR}{SST} = \frac{\text{regression} sum \text{ of squares}}{total \text{ sum of squares}}$$

$$0 \le r^2 \le 1$$

Using ANOVA to find R^2

ANOVA (for 11 rural female headed HHs) of coffee consumption (number of cups) per person per day and price of coffee (in Birr) is:

Source of variation	d.f.	Sum of squares	Mean square	F	Prob.
Regression	k-1 (1)	0.2935	0.2935	17.848	0.000
Residual	n-k (9)	0.148	0.0164		
Total	n-1 (10)	0.4415	0.2743		

$$R^2 = \frac{RSS}{TSS} = \frac{0.2935}{0.4415} = 0.665$$

Interpretation of R^2

From above, we can say that **66.5%** of the variability in the dependent variable is accounted for by the independent variable. Clearly there are many other factors that influence the dependent variable since about **33.5%** of the variability is left unexplained!

Benefits of R^2 and r

- **r** is useful as an initial exploratory tool when several variables are being considered.
- The **sign** of **r** gives the **direction** of the association.
- **R^2** is useful in regression studies to check how much of the variability in the key response can be explained.

- **R^2** is most valuable when there is more than one explanatory variable.
- High values of **R^2** are particularly useful when using the model for predictions!.

Limitations of r

Observe that seemingly high values of **r**, e.g. **r = 0.70**, explain only about **50%** $\left(0.7^2 = 50\%\right)$ of the variability in the response variable **Y**. So take care when interpreting correlation coefficients. A low value for **r** does not necessarily imply absence of a relationship-could be a curved relationship! So plotting the data is crucial! Tests exist for testing there is no association. But depending on the sample size, even low values of r, e.g. r=0.20 can give significant results – not a very useful finding!

Limitations of R^2

Note that R^2 is only a descriptive measure to give a quick assessment of the model. Other methods exist for assessing the goodness of fit of the model. Adding explanatory variables to the model always increases R^2. Hence in practice, it is more usual to look at the **adjusted R^2.**

- The **adjusted R^2** is calculated as $\bar{R}^2 = R^2 - \frac{k-1}{n-k}\left(1-R^2\right)$
- As with **R^2**, the adjusted **R^2** is often expressed as a percentage.

3.6. Hypothesis Testing

Research hypotheses attempt to explain, predict and explore the relationship between two or more variables. To this end, hypotheses can be thought of as the researcher's educated guess about how the study will turn out. Hypothesis testing is designed to detect significant differences: differences that did not occur by random chance. In the "one sample" case: we compare a random sample (from a large group) to a population. We compare a sample statistic to a population parameter to see if there is a significant difference.

- **There are two important points that should be kept in mind.**

1. All hypotheses must be falsifiable. That is, hypotheses must be capable of being refuted based on the results of the study. If a researcher's hypothesis cannot be refuted, then the researcher is not conducting a scientific investigation.

2. A hypothesis must be a prediction (usually, about the relationship between two or more variables).

Types of Hypotheses

There are two kinds of research hypotheses.

1. **The null hypothesis**

The null hypothesis always predicts that there will be no difference between the groups on the variable of interest being studied, or the independent variable has no effect on the dependent variable.

2. **The alternate (or experimental hypothesis)**

The alternate hypothesis predicts that there will be a difference between the groups, or that the independent variable determines the dependent variable.

Types of Alternate Hypotheses

Depending upon its **sign** a research hypotheses can be divided into two categories.

a) **Directional hypotheses**

Directional hypotheses stipulate the direction of the expected differences or relationships between variables. In other words, in directional hypotheses the researcher shows the sign of relationship between the dependent variable and explanatory variables. For example, a statement "credit has a positive effect on technology adoption" is a directional hypothesis.

b) **Non-directional hypotheses**

The statement "income determines technology adoption" is non-directional hypothesis. The decision regarding whether to use a directional hypothesis or a non-directional hypothesis depends on the **researcher's prior knowledge** of the relationship on the variables under consideration. If the researcher knows the relationship of variables, and the sign of relationship he or she may opt to assign a particular sign (plus or minus) on the hypothesized relationship.

In other words, if the researcher believes that the two groups differ but does not have a belief regarding how the groups differ i.e., in which direction they will differ, and then the researcher uses a non-directional hypothesis.

Steps Used in a Hypothesis Test

Regardless of the type of hypothesis being considered, the process of carrying out a significance test is the same and relies on four basic steps:

Step One: **State the null and alternative hypotheses**. Also think about the type 1 error (rejecting a true null) and type 2 error (declaring the plausibility of a false null) possibilities at this time and how serious each mistake would be in terms of the problem.

Step Two: Collect and summarize the data so that a **test statistic can be calculated**. A **test statistic** is a summary of the data that measures the difference between what is seen in the data and what would be expected if the null hypothesis were true. It is typically standardized so that a *p*-value can be obtained from a reference distribution like the normal curve.

Step Three: **Use the test statistic to find the *p*-value**. The ***p*-value** represents the likelihood of getting our test statistic or any test statistic **more extreme**, if in fact the null hypothesis is true.

For a one-sided "greater than" alternative hypothesis, the "more extreme" part of the interpretation refers to test statistic values larger than the test statistic given.

For a one-sided "less than" alternative hypothesis, the "more extreme" part of the interpretation refers to test statistic values smaller than the test statistic given.

For a two-sided "not equal to" alternative hypothesis, the "more extreme" part of the interpretation refers to test statistic values that are farther away from the null hypothesis than the test statistic given at either the upper end or lower end of the reference distribution (both "tails").

Step Four: **Interpret what the *p*-value is telling you** and **make a decision using the *p*-value**. Does the null hypothesis provide a reasonable explanation of the data or not? If not it is statistically significant and we have evidence favoring the alternative. State a conclusion in terms of the problem.

Common Decision Rules seen in the literature

If the *p*-value ≤ .05, we often see scientists declare their data to be "significant."

If the *p*-value ≤ .01, we often see scientists declare their data to be "highly significant".

If the *p*-value > .05, we often see scientists declare their data to be "not significant".

However, such cut-offs are arbitrary and we should not view data any differently when we see a *p*-value of 0.049 versus when we see a *p*-value of 0.051. There is no magic in the 0.05 value.

Statistical Inference in Simple Linear Regression Model

Estimation of Standard Error

To make statistical inferences about the true (population) regression coefficient, β we make use of the estimator $\hat{\beta}$ and its variance var ($\hat{\beta}$). We know that:

Var ($\hat{\beta}$) = $\frac{\sigma^2}{\sum x_i^2}$

Where, $x_i = X_i - \bar{X}$. This variance depends on the unknown parameter σ^2 .Thus, we have to estimate σ^2 . An unbiased estimator of σ^2 is given by:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{n=1}^{n}\left(Y_i - \hat{\alpha} - \hat{\beta} X_i\right)^2 = \frac{1}{n-2}\sum_{n=1}^{n}\hat{\varepsilon}^2$$

Then it follows that an unbiased estimator of var ($\hat{\beta}$) is given by:

$$\hat{\sigma}^2_{\hat{\beta}} = \frac{\hat{\sigma}^2}{\sum x_i^2} = \frac{\sum \hat{\varepsilon}_i^2}{(n-2)\sum x_i^2}$$

The square root of $\hat{\sigma}^2_{\hat{\beta}}$, that is $\hat{\sigma}_{\hat{\beta}}$ is called the standard error of $\hat{\beta}$.

Test of Model Adequacy

The variation in the dependent variable Y can be decomposed as:

$$\Sigma\left(Y_i - \bar{Y}\right)^2 = \Sigma\left(Y_i - \hat{Y}\right)^2 = \Sigma\left(\hat{Y}_i - \bar{Y}\right)^2$$

Total Sum of Squares (TSS) | Error Sum of Squares (ESS) | Regression Sum of Squares (RSS)

The coefficient of determination is defined as:

$$R^2 = \frac{RSS}{TSS} = \frac{\sum\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum\left(Y_i - \bar{Y}\right)^2} = 1 - \frac{\sum \hat{\varepsilon}^2}{\sum\left(Y - \bar{Y}\right)^2}$$

R^2 measures the proportion of variation in the dependent variable, **Y** that is explained by the explanatory variables (or by the linear regression model). It is a **goodness-of-fit statistic**.

To test for the significance of **R^2** (i.e. the adequacy of the model), we calculate the **F-ratio**.

$$F_{Cal} = \frac{RSS/(k-1)}{ESS/(n-k)} = \frac{\hat{\beta}^2 \sum\left(X - \bar{X}\right)^2 /(k-1)}{\sum \hat{\varepsilon}^2 /(n-k)}$$

Where **k** is the number of explanatory variables (or number of parameters estimated from the sample data) and **n** is the sample size. The model is said to be adequate if:

$F_{Cal} > F_{\alpha}(k-1, n-k)$

Where $F_{\alpha}(k-1, n-k)$ is the value of F-distribution with (k-1) and (n-k) degrees of freedom in the numerator and denominator, respectively, for a level of significance α (usually α = 0.01 and

α = 0.05). Model adequacy test results are often presented in analysis of variance (ANOVA) tables (shown below).

ANOVA TABLE

Source of variation	Sum of squares	d.f	Mean square	Variance ratio
Regression	RSS	k-1	$\frac{RSS}{k-1}$	$F_{Cal} = \frac{RSS/(k-1)}{ESS/(n-k)}$
Residual	ESS	n-k	$\frac{ESS}{n-k}$	
Total	TSS	n-1		

Note: The **F test** is designed to test the significance of all variables or a set of variables in a regression model. In the two-variable model, however, it is used to test the explanatory power of a single variable (X), and at the same time, is equivalent to the **test of significance of R^2**.

Test of Significance of Regression Coefficients

Consider the simple linear regression model:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

If there is no relationship between **X** and **Y**, then this is equivalent to saying β = 0 (β is not significantly different from zero). Thus, the null hypothesis of no relationship between **X** and **Y** is expressed as:

H_0: β = 0

The alternative hypothesis is that there is a significant relationship between **X** and **Y**, that is:

H_1: β ≠ 0

In order to reject the null hypothesis, we calculate the **test statistic** given by:

$$t = \frac{\hat{\beta} - \beta_0}{\hat{\sigma}_\beta} \qquad t = \frac{\hat{\beta} - 0}{\hat{\sigma}_\beta} = \frac{\hat{\beta}}{\hat{\sigma}_\beta}$$

If $|t| > t_{\alpha/2}(n-2)$ then we reject the null hypothesis and conclude that there is a significant relationship between **X** and **Y** where $t_{\alpha/2}(n-2)$ is the value from the student's t distribution with (n-2) degrees of freedom for a given significance level α.

Confidence interval for β.

Confidence interval provides a range of values which are likely to contain the true regression parameter. A (1-α) 100% confidence interval for **β** is given by: $\hat{\beta} \pm t_{\alpha/2}(n-2)\hat{\sigma}_{\hat{\beta}}$

Example 1: Consider the following data where the dependent variable (Y) is coffee consumption (number of cups) per person per day and independent variable (X) is price of coffee (in Rupees).

Year	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	**Total**	**Mean**
X	0.77	0.74	0.72	0.73	0.76	0.75	1.08	1.80	1.39	1.20	1.17	**11.11**	**1.01**
Y	2.57	2.50	2.35	2.30	2.25	2.20	2.11	1.94	1.97	2.06	2.02	**24.27**	**2.21**

Summary statistics

$$\sum x_i y_i = \sum \left(X_i - \bar{X} \right) \left(Y_i - \bar{Y} \right) = \sum X_i Y_i - n \bar{X} \bar{Y} = -0.6083$$

$$\sum x_i^2 = \sum \left(X_i - \bar{X} \right)^2 = \sum X^2 - n \bar{X}^2 = 1.2582$$

$$\sum y_i^2 = \sum \left(Y_i - \bar{Y} \right)^2 = \sum Y^2 - n \bar{Y}^2 = 0.442$$

The **OLS** estimator of **β** is: $\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{-0.6083}{1.2582} = -0.483$

The **OLS** estimator of α is: $\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X} = 2.21 - (-0.483)(1.01) = 2.6978$

Thus, the estimated regression line (the consumption function for coffee) is: $\hat{Y}_i = -0.483 X_i + 2.6978$.This means that the estimated slope coefficient of the consumption function for coffee is -0.483 and its standard deviation (standard error) is 0.1143. This is a measure of the variability of **β** from sample to sample.

The estimated errors (residuals) are obtained as:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i + 0.483 \hat{X}_i - 2.6978$$

Year	X	Y	$\hat{Y}_i = -0.483 + 2.6978$	$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$	$\hat{\varepsilon}^2$
1990	0.77	2.57	2.326	0.244	0.060
1991	0.74	2.50	2.340	0.160	0.025
1992	0.72	2.35	2.350	0.000	0.000
1993	0.73	2.30	2.345	-0.045	0.002
1994	0.76	2.25	2.331	-0.081	0.007
1995	0.75	2.20	2.336	-0.136	0.018
1996	1.08	2.11	2.176	-0.066	0.004
1997	1.80	1.94	1.828	0.112	0.012
1998	1.39	1.97	2.026	-0.056	0.003
1999	1.20	2.06	2.118	-0.058	0.003
2000	1.17	2.02	2.133	-0.113	0.013
					0.148

Thus, the error sum of squares (ESS) equals 0.148. We then calculate the regression sum of squares (RSS) and the total sum of squares (TSS):

$$RSS = \hat{\beta}^2 \sum x_i^2 = \left(-0.148\right)^2 \left(1.2582\right) = 0.2935$$

$$TSS = RSS + ESS = \hat{\beta}^2 \sum X_i^2 + \sum \hat{\varepsilon}^2 = 0.2935 + 0.148 = 0.4415$$

The ANOVA table is shown below:

ANOVA TABLE

Source of variation	Sum of squares	d.f	Mean square	F
Regression	0.2935	2-1=1	0.2935	17.848
Residual	0.148	11-2=9	0.0164	
Total	0.4415	11-1=10		

The coefficient of determination is:

$$R^2 = \frac{RSS}{TSS} = \frac{0.2935}{0.4415} = 0.665$$

This figure tells us that 66.5% of the variation in coffee consumption is due to changes in price of coffee, whereas 33.5% of the variation is due to other factors (variables) not included in our model.

To test the significance of R^2 (or model adequacy), the test statistic is 17.848 (shown in the ANOVA table). For level of significance α = 0.01, the value from the F-distribution with degrees of freedom 1 and 9 is: $F_{0.01}(1, 9) = 10.56$.

Decision: Since the test statistic is greater than the tabulated value, we reject the null hypothesis and conclude that the linear model is adequate to explain the relationship between price of coffee (X) and coffee consumption (Y).

The variance of $\hat{\beta}$ is estimated as:

$$\hat{\sigma}_{\beta}^{\ 2} = \frac{\sum \hat{\varepsilon}_i^{\ 2}}{(n-2)\sum x_i^{\ 2}} = \frac{0.148}{(11-2)(1.2582)} = 0.01307$$

The standard error of $\hat{\beta}$ is: $\Rightarrow \hat{\sigma}_{\beta} = \sqrt{0.01307} = 0.1143$

Test of hypothesis

Is there a significant relationship between coffee consumption and price of coffee? The hypothesis to be tested is:

H_0: $\beta = 0$

H_1: $\beta \neq 0$

The test statistic is calculated as: $t = \frac{\hat{\beta}}{\hat{\sigma}_{\beta}} = \frac{-0.483}{0.1143} = -4.226$

For significance level α = 0.01 = 1% (that is 99% level of confidence):

$t_{0.005}(11-2) = t_{0.005}(9) = 3.25$. If the computed (calculated) absolute t value | t |, exceeds the critical (tabulated) t-value at the chosen level of significance, we reject the null hypothesis.

Decision: Since $| t | > t_{\alpha/2}(n-2)$, we **reject the null hypothesis** and conclude that there is a significant relationship between coffee consumption and price of coffee.

Example 2. Left Handed Artists

About 10% of the human population is left-handed. A researcher at Penn State speculates that students in the College of Arts and Architecture are more likely to be left-handed than people in the general population. A random sample of 100 students in the College of Arts and Architecture is obtained and 18 of these students were found to be left-handed.

Research Question: Are artists more likely to be left-handed than people in the general population?

Step 1: State Null and Alternative Hypotheses

- **Null Hypothesis**: Population proportion of left-handed students in the College of Art and Architecture = 0.10 ($p = 0.10$).
- **Alternative Hypothesis**: Population proportion of left-handed students in the College of Art and Architecture > 0.10 ($p > 0.10$).

Now that you know the null and alternative hypothesis, did you think about what the type 1 and type 2 errors are? It is important to note that Step 1 is before we even collect data. Identifying these errors helps to improve the design of your research study. Let's write them out:

Type 1 error: Claim artists are more likely to be left-handed than people in the general population, when in truth they are not more likely (reject the null hypothesis when it is true).

Type 2 error: Fail to claim artists are more likely to be left-handed than people in the general population, when they are in fact more likely (accept the null hypothesis when it is false).

In this case, the consequences of these two errors are fairly similar (e.g. installing more or fewer left handed desks in classrooms than are needed).

Step 2: Collect and summarize the data so that a test statistic can be calculated.

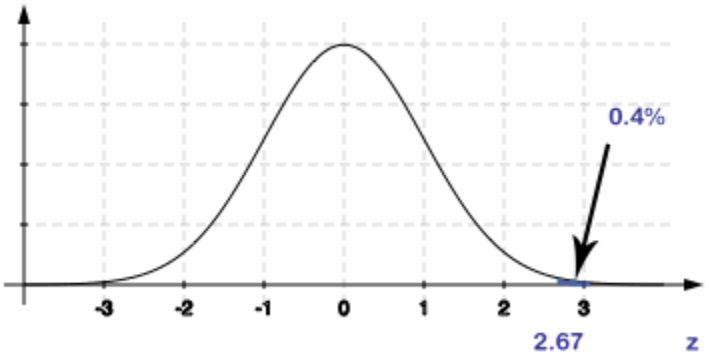
In the sample of 100 students listed above, the sample proportion is 18/100 = 0.18. The hypothesis test will determine whether or not the null hypothesis that $p = 0.1$ provides a plausible explanation for the data. If not we will see this as evidence that the proportion of left-handed Art & Architecture students is greater than 0.10.

If the null hypothesis is true then the standard error of the sample proportion would be

$\sqrt{\frac{0.1(1-0.1)}{100}} = 0.03$ and the sample proportion would follow the normal curve. Thus, we can use the standard score $z = (0.18\text{-}0.10) / 0.03 = 2.67$ as our test statistic.

Step 3: Use the test statistic to find the *p*-value.

Using the normal curve table for the *Z*-value of 2.67 we find the *p*-value to be about 0.004. Notice that the one-sided alternative hypothesis says to watch out for large values so we look at the percentage of the normal curve above 2.67 to get the *p*-value.



Interpretation of the *p*-value. The likelihood of getting our test statistic of 2.67 or any higher value, if in fact the null hypothesis is true, is 0.004.

Step 4: Make a decision using the *p*-value.

Since the *p*-value of 0.004 is so small, the null hypothesis provides a very poor explanation of the data. We find good evidence that the population proportion of left-handed students in the College of Art and Architecture exceeds 0.10.

Now that we have made our decision, we are only at risk of making a type 1 error. It is not possible at this point to make a type 2 error because we rejected the null hypothesis.

Inferential statistics

- Inferential statistics attempt to generalize the results of descriptive statistics to a larger population of interest.
- Interval estimates
    - Estimates the margin of error in sample statistics compared to population values
    - The range around the sample statistics where the true population value is likely to be found
- Example: 95% confidence interval for mean
- Statistical theory tells us 95% of the distribution will be within +/– two standard errors of the mean

Standard error of the mean

$$S_M = \frac{S}{\sqrt{n}}$$

Where s = standard deviation; n = sample size

- If mean from the sample is m, 95% confidence interval for the population mean

$m - 2S_m \leq x \leq m + 2S_m$

Standard Error Calculation

Step 1. Calculate the mean (total of all samples divided by the number of samples).

Step 2. Calculate each measurement's deviation from the mean (mean minus the individual measurement).

Step 3. Square each deviation from the mean. Squared negatives become positive.

Step 4. Sum the squared deviations (Add up the numbers from step 3).

Step 5. Divide that sum from step 4 by one less than the sample size (n-1, that is the number of measurements minus one).

Step 6. Take the square root of the number in step 5.

- That gives you the "standard deviation (S.D.)".

Step 7. Divide the standard deviation by the square root of the sample size (n). That gives you the "standard error"

Step 8. Subtract the standard error from the mean and record that number.

- Then add the standard error to the mean and record that number.
- You have plotted mean ±1 standard error (S.E.), then the distance from 1 standard error below the mean to 1 standard error above the mean.

Tests of Statistical Significance between groups

- You probably have observed differences between groups.
- You may want to find out if these differences are likely to be due to:
  - ✓ Chance, or

- ✓ If they are real (statistically significant) differences.

- In order to determine this, you can perform two types of tests.
- These are:
    - The t-test, and
    - The chi-square test (it will not be discussed under this chapter).
- The **t-test** is used for numerical data, when comparing the means of two groups.
- The **chi-square test** is used for categorical data, when comparing proportions of events occurring in two or more groups.

T-Test

- When the sample size is small (approximately < 100) then the Student's *t* distribution should be used.
- The t-test, also referred to as **Student's t-test,** is used for
    - ✓ Numerical data to determine whether an observed difference between the means of two groups can be considered statistically different.

Example:

- It has been observed that in a certain area the proportion of women who are delivered through Caesarean section is very high.
- A study is therefore conducted to discover why this is the case.
- Small height is known to be one of the risk factors related to difficult deliveries.
- The researcher wants to find out if there is a difference between the **mean height** of women in this area who had **normal deliveries** and of those who had **Caesarean sections.**
- The **null hypothesis** would be that there is **no difference** between the **mean heights** of the two groups of women.
- Suppose the following results were found:

Mean heights of women with normal deliveries and of women with Caesarean sections

Type of delivery	No. of women included in the study	Mean height in cm	Standard deviation
Normal delivery	60	156	3.1
Caesarean section	52	154	2.8

- A **t-test** would be the appropriate way to determine whether the observed difference of 2 cm can be considered statistically significant.
- **To actually perform a t-test you have to complete 3 steps:**
    1. Calculate the t-value

2. Choose the level of significance and use a t-table
3. Interpret the results

Step 1. Calculating the t-value

To calculate the t-value you need to complete the following tasks:

Calculate the difference between the means. In the above example the difference is 156-154 = 2 cm

a) Calculate the standard deviation for each of the study groups.
b) Calculate the standard error of the difference between the two means.

- **The standard error of the difference is given by the following formula:**

$$\sqrt{\frac{SD_1^{\,2}}{n_1}+\frac{SD_2^{\,2}}{n_2}}$$

Where: SD_1 is the standard deviation of the first sample

SD_2 is the standard deviation of the second sample

n_1 is the sample size of the first sample

n_2 is the sample size of the second sample.

- For our data if we take the women with normal deliveries as sample 1, and
- Those with Caesarean sections as sample 2 the standard error of the difference is:

$$\sqrt{\frac{3.1^2}{60}+\frac{2.8^2}{52}}=0.56$$

d) Finally, divide the difference between the means by the standard error of the difference. The value now obtained is called t-value.

- In the above example:

$$t=\frac{2}{0.56}=3.6$$

- Expressed in one single formula:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}}$$

Where X_1 is the mean value of the first sample, and X_2 is the mean value of the second sample)

Step 2. Using a t-table

- Once the t-value has been calculated, you will have to refer to a t-table, from which you can determine whether the null hypothesis is rejected or not.

1. First, decide which significance level (a-value or alpha value) you want to use.

- Remember that the chosen significance level (a-value) is an expression of the likelihood of finding a difference by chance when there is no real difference.
  - Usually we choose a significance level of 0.05.

2. Second, determine the number of degrees of freedom for the test being performed.

- Degree of freedom is a measure derived from the sample size, which has to be taken into account when performing a t-test.

3. Third, the t-value belonging to the a-value (the significance level we choose) and the degrees of freedom are located in the table.

If the calculated t-value is equal to or larger than the value derived from the table, we then reject the null hypothesis and conclude that there is a statistically significant difference between the two means. If the calculated t-value is smaller than the value derived from the table, we then accept the null hypothesis and conclude that the observed difference is not statistically significant.

- The way the number of degrees of freedom is calculated differs from one statistical test to the other.
- For student's t-test the number of degrees of freedom is calculated as the sum of the two sample sizes minus 2.

Thus, for the above example, comparing the heights of women with and without Caesarean sections, the number of degrees of freedom is:

- d.f. = 60+52-2 = 110

Note:

This is an approximate way of determining degrees of freedom. For the exact method, refer to a statistics textbook.

In our example we look up the t-value belonging to a = 0.05 and d.f. = 120 and we find it is 1.98.

Step 3. Interpreting the result

- We now compare the absolute value of the t-value calculated in Step 1 (i.e., the t-value, ignoring the sign) with the t-value derived from the table in Step 2.
- In our example the calculated t-value (3.6) is larger than the tabulated t-value (1.98).
    - Thus the p-value is smaller than 0.05, and we therefore reject the null hypothesis and
    - Conclude that the observed difference of 2 cm between the mean heights of women with normal deliveries and women with Caesarean sections is a **statistically significant difference**.

- We can express this conclusion in different ways:

We can say that the probability that the observed difference of 2 cm of height between the two groups of women is due to chance is less than 5%.

We can also say that the difference between the two groups is 3.6 times the standard error.

Chapter 4: Multiple Linear Regression Analysis

4.1. Model with Two Independent Variables

In the previous chapter we considered a regression model with a single independent variable. But in practice, economic models generally contain one dependent variable and two or more independent variables. Such models are called **multiple regression models**.

Suppose $Y = f(X_1, X_2, X_3 \ldots\ldots X_k)$, the general population regression function of multiple regression model is given as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots\ldots\ldots\ldots + \beta_k X_{ki} + \varepsilon_i$$

However, the simplest form of multiple linear regression model (i.e. a model with two explanatory variables) is given by:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Taking the expected value of the above model, we obtain,

$$E(Y_i / X_{1i}, X_{2i}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

Where: $E(Y_i / X_{1i}, X_{2i})$ represents the conditional mean of Y_i given fixed values of X_{1i} and X_{2i}

: β_0 is the average value of Y_i when $X_{1i} = X_{2i} = 0$.

: β_1 is obtained by taking the partial derivatives of Y_i with respect to X_{1i} keeping X_{2i} constant. That is,

$\beta_1 = \frac{\partial Y_i}{\partial X_{1i}}, keeping\ X_{2i}\ constant$ which represents the change in the mean value of Y_i with respect to X_{1i} keeping X_{2i} constant. Similarly, β_2 is obtained by taking the partial derivatives of Y_i with respect to X_{2i} keeping X_{1i} constant, i.e.

$\beta_2 = \frac{\partial Y_i}{\partial X_{2i}}, keeping\ X_{1i}\ constant$ which represents the change in the mean value of Y_i with respect to X_{2i} keeping X_{1i} constant.

In the multiple regression model, we add one more assumption to the assumptions of classical linear regression model that we have not discussed in the simple linear regression model in the previous unit. This additional assumption says there is no exact collinearity between the explanatory variables (no perfect multicollinearity), i.e. no exact linear relationship between the explanatory variables. Hence, none of the explanatory variables can be written as a linear combination of the remaining explanatory variables. Lack of multicollinearity implies that each independent variable does have some information content not contained in the other independent variable.

Example

Assume the dependent variable is television sales (Y_i) per month in a certain company and the explanatory (independent) variables are price of television (X_1) and the amount spent in advertising (X_2). The multiple regression model is given as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X2_i + \varepsilon_i$$

β_1 is obtained by taking the partial derivatives of Y_i with respect to X_{1i} keeping X_{2i} constant, which represents the change in the mean value of TV sales due to change in price keeping advertising constant.

Similarly, β_2 is obtained by taking the partial derivatives of Y_i with respect to X_{2i} keeping X_{1i} constant which represents the change in the mean value of TV sales due to change in advertising spending keeping price constant.

β_0 represents the mean value of TV sales when price and advertising spending are zero. This may not have an economic meaning. In a multiple regression model the coefficients measure the change in the mean value of the dependent variable due to the change in one of the variable while keeping the other variable constant.

Multiple regression model is an extension of simple linear regression model with more than one explanatory variable. Let us examine a multiple regression model with two explanatory variables. From the theory of demand, quantity demanded for a given commodity (Q) depends on its price (P) and consumer income (Y) and given as follows:

$$Q = \beta_0 + \beta_1 P + \beta_2 Y$$

Since the theory of demand does not specify the mathematical form, assume the relationship between Q, P and Y is linear.

The above mathematical form is an exact relationship which indicates that the variation in the quantity demanded is fully explained by changes in price and consumers income. If this relationship were true, then any observation on Q, P and Y would represent a point which lies on a plane. However, if we gather observations on these variables and plot them on a diagram, we will observe that all of them will not lie on a plane. Some points will lie on the plane while others will lie above or below the plane. This scatter is due to various factors omitted from the function. The influences of such factors are taken into account by introducing a random variable. As a result, the above function is given as:

$$Q = \beta_0 + \beta_1 P + \beta_2 Y + \varepsilon_i$$

Where: $\beta_0 + \beta_1 P + {}_2Y$ is the systematic component and ε_i is a random component

From the theory of demand, we will expect that the coefficient β_1 to have a negative sign because of the law of demand while β_2 is expected to be positive for normal commodities or negative for inferior commodities.

Similar to what we did in unit three, the next step in order to complete the specification of the model is, adding assumptions regarding to the random variable.

4.2. Assumptions of the Multiple Regression Model

The assumptions of the classical linear regression model are:

Assumption 1: The model is linear

i.e. $Y_i = \beta_1 + \beta_2 X_{2_i} + \beta_3 X_{3_i} + ... + \beta_K X_{K_i} + \varepsilon_i$

Assumption 2: Randomness of ε
The variable ε is a random variable, i.e. its value is unpredictable and hence depends on chance.

Assumption 3: Zero mean of the random variable ε.
The random variable ε has a zero mean for each X_i

i.e. $E(\varepsilon_i) = 0$

Assumption 4: Homoscedasticity
The variance of each ε_i is the same for all the X_i values

i.e. $\text{var}(\varepsilon_i) = E[\varepsilon_i - E(\varepsilon_i)]^2 = E(\varepsilon_i^2) = \sigma^2$, constant, since $E(\varepsilon_i) = 0$

Assumption 5: Normality of U
The values of each ε_i are normally distributed, i.e.

$U_i \approx N(0, \sigma_u^2)$

Assumption 6: Non autocorrelation or serial independence of the U's
The values of ε_i corresponding to X_i are independent from the values of any other U_j corresponding to X_j,

i.e. $E(\varepsilon_i \varepsilon_j) = 0$, for $i \neq j$

Assumption 7: Independence of ε_i and X_i

Every disturbance term ε_i is independent of the explanatory variables,

i.e. $E(\varepsilon_i X_i) = 0$

Assumption 8: No errors of measurements in the X's
The explanatory variables are measured without error (each of the explanatory variables is non-stochastic).
Assumption 9: No perfect multicollinear X's
The explanatory variables are not perfectly linearly correlated.
Assumption 10: Correct specification of the model
The model has no specification error, i.e. all the important explanatory variables appear explicitly in the function and the mathematical form is correctly specified.

4.3. Estimation of Partial Regression Coefficients

After specification of the model, the next step is using sample observations on Y, X_{1i} and X_{2i} and obtain estimates of the population parameters β_0, β_1 and β_2.

These estimates b_1, b_2 *and* b_3 of the population parameters β_1, β_2 *and* β_3 respectively will be obtained by minimizing the sum of squared residuals.
The population regression function is given as:

$$Y_i = \beta_1 + \beta_2 X_{2_i} + \beta_3 X_{3i} + \varepsilon_i \quad \text{..... 4.1}$$

and the counterpart sample regression function is given as:

$$\hat{Y}_i = b_1 + b_2 X_{2_i} + b_3 X_{3i},$$

Note that, the true value of Y for given X values is computed as follows:

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

The difference between the actual and the estimated values is given by:

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

To obtain the OLS estimates of the parameters choose β_0, β_1, *and* β_2 such that the sum of square of the residuals is as small as possible, i.e.

$$Minimize \sum \varepsilon_i^{\,2} = Minimize\ \Sigma \left(Y_i - b_1 - b_2 X_{2i} + b_3 X_{3i} \right)^2 \quad \text{.....} \quad 4.2$$

To minimize this function, take the partial derivative of equation (4.2) with respect to $b_0, b_1, and\ b_2$ and set to zero and solve for, $b_0, b_1, and\ b_2$.

In this regard the partial derivative of β_i^2 with respect to $\hat{\beta}_0$ gives us the following result.

$$\frac{\partial \varepsilon_i^{\,2}}{\partial b_1} = -2\Sigma \left(Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i} \right) = 0 \quad \text{.....} \quad 4.3$$

$$\frac{\partial \varepsilon_i^{\,2}}{\partial \beta_2} = -2\ \Sigma X_{2i}\ \left(Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i} \right) = 0 \quad \text{.....} \quad 4.4$$

$$\frac{\partial \varepsilon_i^{\,2}}{\partial \beta_3} = -2\ \Sigma X_{3i}\ \left(Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i} \right) = 0 \quad \text{.....} \quad 4.5$$

Expressing Y_i, X_{2i} and X_{3i} in deviation form, we have

$$y_i = b_2 x_{2i} + b_3 x_{3_i} + \varepsilon_i \quad \text{.....} \quad 4.6$$

The error sum of squares (ESS) is:

$$ESS = \varepsilon_i^{\,2} = \Sigma\left(y_i - b_2 X_{2_i} - b_3 X_{3i}\right)^2 \quad \text{.....} \quad 4.7$$

Thus we derive the normal equations:

$$\frac{\partial \varepsilon_i^{\,2}}{\partial b_2} = -2\ \Sigma x_{2i}\ \left(y_i - b_2 x_{2i} - b_3 x_{3i}\right) = 0 \quad \text{.....} \quad 4.8$$

$$\Rightarrow \Sigma x_{2i}\, y_i = b_2 \Sigma x_{2i}^{\,2} + b_3 \Sigma x_{2i}\, x_{3i} \quad \text{.....} \quad 4.9$$

$$\frac{\partial \varepsilon_i^{\,2}}{\partial b_3} = -2\ \Sigma x_{3i}\ \left(y_i - b_2 x_{2i} - b_3 x_{3i}\right) = 0 \quad \text{.....} \quad 4.10$$

$$\Rightarrow \Sigma x_{3i}\, y_i = b_2 \Sigma x_{2i} x_{3i} \ + b_3 \Sigma x_{3i}^{\,2} \quad \text{.....} \quad 4.11$$

Multiply equation 4.9 by $\Sigma x_{3i}^{\,2}$ and multiply equation 4.11 by $\Sigma x_{2i}\, x_{3i}$ and subtracting equation 4.11 from equation 4.9:

$$\Sigma x_{2i}\, y_i \Sigma x_{3i}^{\,2} \ - \Sigma x_{3i}\, y_i \Sigma x_{2i}\, x_{3i} = b_2 \left[\left(\Sigma x_{2i}^{\,2}\right)\left(\Sigma x_{3i}^{\,2}\right) - \left(\Sigma x_{2i}\, x_{3i}\right)^2\right] \quad \text{.....} \quad 4.12$$

$$b_2 = \frac{\left(\Sigma x_{2i}\, y_i\right)\left(\Sigma x_{3i}^{\,2}\right) - \left(\Sigma x_{3i}\, y_i\right)\left(\Sigma x_{2i}\, x_{3i}\right)}{\left(\Sigma x_{2i}^{\,2}\right)\left(\Sigma x_{3i}^{\,2}\right) - \left(\Sigma x_{2i} x_{3i}\right)^2} \quad \text{.....} \quad 4.13$$

Similarly,

$$b_3 = \frac{\left(\sum x_{3i} y_i\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{2i} y_i\right)\left(\sum x_{2i} x_{3i}\right)}{\left(\sum x_{2i}^2\right)\left(\sum x_{3i}^2\right) - \left(\sum x_{2i} x_{3i}\right)^2} \quad \text{..... 4.14}$$

And

$$b_1 = \bar{Y} - b_2 \bar{X}_2 - b_3 X_3 \quad \text{.....4.15}$$

Example: Consider the following data on per capita food consumption (**Y**), price of food ($\mathbf{X_2}$) and per capita income ($\mathbf{X_3}$) for the years **1927-1941** in a certain country. Retail price of food and per capita disposable income are deflated by dividing the Consumer Price Index.

Year	Y	X_2	X_3	Year	Y	X_2	X_3
1927	88.9	91.7	57.7	1935	85.4	88.1	52.1
1928	88.9	92	59.3	1936	88.5	88	58
1929	89.1	93.1	62	1937	88.4	88.4	59.8
1930	88.7	90.9	56.3	1938	88.6	83.5	55.9
1931	88	82.3	52.7	1939	91.7	82.4	60.3
1932	85.9	76.3	44.4	1940	93.3	83	64.1
1933	86	78.3	43.8	1941	95.1	86.2	73.7
1934	87.1	84.3	47.8				

We want to fit a multiple linear regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i \qquad i=1, 2, 3, ..., 15$$

Summary statistics:

$$\bar{Y} = 88.90667, \quad \bar{X}_2 = 852.9, \quad \bar{X}_3 = 56.52667$$

The sums in deviation forms are:

$$\sum x_{2i} y_i = 27.63, \quad \sum x_{3i} y_i = 257.397, \quad \sum x_{2i} x_{3i} = 275.9, \quad \sum x_{2i}^2 = 355.14, \quad \sum x^2_{3i} = 838.289$$

$$\sum y_i^2 = 99.929$$

Estimates of the regression coefficient are:

$$b_2 = \frac{\left(\sum x_{2i} y_i\right)\left(\sum x_{3i}^2\right) - \left(\sum x_{3i} y_i\right)\left(\sum x_{2i} x_{3i}\right)}{\left(\sum x_{2i}^2\right)\left(\sum x_{3i}^2\right) - \left(\sum x_{2i} x_{3i}\right)^2} = -0.21596$$

$$b_3 = \frac{\left(\sum x_{3i} y_i\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{2i} y_i\right)\left(\sum x_{2i} x_{3i}\right)}{\left(\sum x_{2i}^2\right)\left(\sum x_{3i}^2\right) - \left(\sum x_{2i} x_{3i}\right)^2} = 0.378127$$

$$b_1 = \bar{Y} - b_2 \bar{X}_2 - b_3 X_3 = 86.08318$$

Hence, the estimated model is:

$$\hat{Y}_i = 86.08318 - 0.21596 X_{2i} + 0.378127 X_{3i}$$

4.4. The Partial Correlation Coefficient

In the multiple regressions, the partial correlation coefficient measures the correlation between any two variables when all the other variables are held constant. In a partial correlation coefficient, we measure the correlation between two variables by removing the influence of other variables. For example $ryx_1.x_2$ measures the correlation coefficient between Y and X_1 when the influence of X_2

has been removed. In this type of correlation, X_2 may be considered as constant when studying the relationship between Y and X_1.

This partial correlation coefficient will be given as:

$$ryx_1.x_2 = \frac{ryx_1 - ryx_2 rx_1x_2}{\sqrt{(1-ryx_2^{\,2})\,(1-rx_1x_2^{\,2})}}$$

Similarly, the partial correlation between Y and X_2 when X_1 is kept constant is obtained as:

$$ryx_2.x_1 = \frac{ryx_2 - ryx_1 rx_1x_2}{\sqrt{(1-ryx_1^{\,2})\,(1-rx_1x_2^{\,2})}}$$

The example below explains the concept discussed above

Example

Suppose the following computation results were obtained from a sample of **12** firms on their output (**Q**), labor input (**L**) and capital input (**K**) measured in millions of Birr.

$\sum Q_i = 753$	$\sum Q_i^2 = 48139$,	$\sum L_iQ_i = 40830$	$\sum K_iQ_i = 6796$
$\sum L_i = 643$	$\sum L_i^2 = 34843$	$\sum K_i = 106$	$\sum K_i^2 = 976$
$\sum L_iK_i = 5779$			

Assuming linearity we can estimate the model: $Q_i = \beta_1 + \beta_2 L_i + \beta_3 K_i + Ui$ as follows

The estimates of the regression are given as:

$$b_2 = \frac{\Sigma L_i Q_i \Sigma K_i^2 - \Sigma K_i Q_i \Sigma L_i K_i}{\sum L_i^2 \Sigma K_i^2 - (\Sigma L_i K_i)^2} = \frac{(40830)(976) - (6796)(5779)}{(34843)(976) - (5779)^2} = 0.944$$

$$b_3 = \frac{\Sigma K_i Q_i \Sigma L_i^2 - \Sigma L_i Q_i \Sigma L_i K_i}{\sum L_i^2 \Sigma K_i^2 - (\Sigma L_i K_i)^2} = \frac{(6796)(34843) - (40830)(5779)}{(34843)(976) - (5779)^2} = 1.37$$

$$\bar{Q} = \frac{\Sigma Q_i}{n} = \frac{753}{12} = 62.75,\ \bar{L} = \frac{\Sigma L_i}{n} = \frac{643}{12} = 53.58,\quad \bar{K} = \frac{\Sigma K_i}{n} = \frac{106}{12} = 8.83$$

$$b_1 = \bar{Q} - b_2 \bar{L}_2 - b_3 \bar{K}_3 = 62.75 - (0.944)(53.58) - (1.37)(8.83) = 0.073$$

Hence the estimated regression model is given as:

$$\hat{Q} = 0.073 + 0.944 L_i + 1.37 K_i$$

Note that $b_2 = 0.944$ represents the amount in which the level of output increases when the level of labor input increases by one unit while the level of capital remain the same. Similarly, $b_3 = 1.37$ represents the amount in which the level of output increases when the level of capital input increases by one unit while the level of labor remain the same. On the other hand $b_1 = 0.073$ represents the level of output, when the amount of labor and capital inputs are zero.

Based on the result obtained we can calculate the partial correlation coefficient rQL.K. That is, the partial correlation coefficient between output and labor keeping capital constant.

$$rQL.K = \frac{rQL - rQK rLK}{\sqrt{(1 - rQL^2)\,(1 - rLK^2)}}$$

In order to compute this partial correlation coefficient, first we have to compute the simple correlation coefficient among the variable as:

$$rQL = \frac{\Sigma (Q_i - \bar{Q})\,(L_i - \bar{L})}{\sqrt{\Sigma \left(Q_i - \bar{Q} \right)^2}\sqrt{\Sigma \left(L_i - \bar{L} \right)^2}} = \frac{n\Sigma L_i Q_i - \Sigma L_i \Sigma Q_i}{\sqrt{\left(n\Sigma L_i{}^2 - (\Sigma L_i)^2 (n\Sigma Q_i{}^2 - (\Sigma Q_i)^2 \right)}}$$

$$rQL = \frac{12\ (40830) - (643)(753)}{\sqrt{(12)(48139) - (753)^2)(12)(34843) - (643)^2}} = \frac{5781}{\sqrt{(4667)(10659)}}$$

$$= \frac{5781}{7053.053} = 0.82$$

$$rQK = \frac{\Sigma (Q_i - \bar{Q})\,(K_i - \bar{K})}{\sqrt{\Sigma \left(Q_i - \bar{Q} \right)^2}\sqrt{\Sigma \left(K_i - \bar{K} \right)^2}} = \frac{n\Sigma K_i Q_i - \Sigma K_i \Sigma Q_i}{\sqrt{\left(n\Sigma K_i{}^2 - (\Sigma K_i)^2 (n\Sigma Q_i{}^2 - (\Sigma Q_i)^2 \right)}}$$

$$rQK = \frac{12(6796) - (106)(753)}{\sqrt{(12)(976) - (106)^2 (12)(48139) - (753)^2}} = \frac{1734}{\sqrt{(476)(10659)}} = \frac{1734}{2252.484}$$

$$rQK = 0.77$$

Similarly, we compute the simple correlation between **L** and **K** as:

$$rLK = \frac{\Sigma (L_i - \bar{L})\,(K_i - \bar{K})}{\sqrt{\Sigma \left(L_i - \bar{L} \right)^2}\sqrt{\Sigma \left(K_i - \bar{K} \right)^2}} = \frac{n\Sigma L_i K_i - \Sigma L_i \Sigma K_i}{\sqrt{\left(n\Sigma L_i{}^2 - (\Sigma L_i)^2 (n\Sigma K_i{}^2 - (\Sigma K_i)^2 \right)}}$$

$$rLK = \frac{(12)(5779)-(106)(643)}{\sqrt{((12)(976)-(106)^2((12)(34843)-(643)^2}} = \frac{1190}{\sqrt{(476)(4667)}} = \frac{1190}{1490.467}$$

$$rLK = 0.80$$

Now the partial correlation coefficient rQL.K is given as:

$$rQL.K = \frac{rQL - rQKrLK}{\sqrt{(1-rQL^2)(1-rLK^2)}} = \frac{0.82-(0.77)(0.8)}{\sqrt{(1-(0.82)^2)(1-(0.8)^2)}} = \frac{0.204}{\sqrt{(0.3276)(0.36)}}$$

$$= \frac{0.204}{\sqrt{0.1179}} = \frac{0.204}{0.343} = 0.59$$

It gives us the correlation between output and labor input when the influence of capital input is kept constant.

The partial correlation coefficient rQK.L which measures the correlation between output and capital input when the influence of labor input is kept constant is given as:

$$rQK.L = \frac{rQK - rQLrLK}{\sqrt{(1-rQK^2)(1-rLK^2)}} = \frac{0.77-(0.82)(0.8)}{\sqrt{(1-(0.77)^2)(1-(0.8)^2)}} = \frac{0.114}{\sqrt{(0.4071)(0.36)}}$$

$$= \frac{0.114}{0.383} = 0.30$$

It gives us the correlation between output and capital input when the influence of labor input is kept constant. Notice that we can compute the multiple correlation coefficients $\mathbf{R^2}$ and the adjusted $\mathbf{R^2}$ as follows:

$$R^2 = \frac{\hat{b}_2 \Sigma (Q_i - \bar{Q})(L_i - \bar{L}) + \hat{b}_3 \Sigma (Q_i - \bar{Q})(K_i - \bar{K})}{\Sigma (Q_i - \bar{Q})^2}$$

$$R^2 = \frac{\hat{b}_2 (n\Sigma Q_i L_i - \Sigma Q_i \Sigma L_i) + \hat{b}_3 (n\Sigma Q_i K_i - \Sigma Q_i \Sigma K_i)}{n\Sigma Q_i^{\,2} - (\Sigma Q_i)^2}$$

$$R^2 = \frac{(0.944)(12)(40830) - (753)(643) + (1.37)(12)(6796) - (753)(106)}{(12)(48139) - (753)^2}$$

$$R^2 = \frac{(0.944)(5781) + (1.37)(1734)}{10659} = \frac{7832.844}{10659} = 0.735$$

It represents the total variation in the total output that is explained by the variation in to two explanatory variables, i.e. by the variation in labor input and capital input.

The adjusted coefficient of multiple determinations is given as:

$$\bar{R}^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-k}\right) \qquad \bar{R}^2 = 1 - (1 - 0.735)\left(\frac{12-1}{12-3}\right)$$

$$\bar{R}^2 = 1 - (0.265)\ (1.222) = 0.676$$

4.5. Analysis of Variance

Analysis of variance for regression: the procedure to compute the F-ratio which is used to test the overall significance of the regression coefficients.

Goodness of fit: The coefficient of determination

The coefficient of determination (**R^2**) can be calculated as usual as:

$$R^2 = \frac{RSS}{TSS} = \frac{\Sigma\left(\hat{Y}_i - \bar{Y}\right)^2}{\Sigma\left(Y_i - \bar{Y}\right)} = 1 - \frac{\Sigma \hat{\varepsilon}_i^2}{\Sigma\left(Y_i - \bar{Y}\right)^2} = 1 - \frac{ESS}{TSS}$$

- **For a three variable model:**

$$RSS = b_2 \Sigma x_{2i} y_1 + b_3 \Sigma x_{3i} y_i$$

Thus,

$$ESS = \Sigma y_i^2 - b_2 \Sigma x_{2i} y_1 - b_3 \Sigma x_{3i} y_i$$

$$R^2 = \frac{b_2 \Sigma x_{2i} y_i + b_3 \Sigma x_{3i} y_i}{\Sigma y_i^2}$$

R^2 measures the proportion of variation in the dependent variable **Y** that is explained by the explanatory variables (or by the multiple linear regression model). It is a **goodness-of-fit statistic**. To test the significance of **R^2**, we calculate the **F-ratio**:

$$F_{cal} = \frac{RSS/(k-1)}{ESS/(n-k)}$$

Where **k** is the number of parameters estimated from the sample data and **n** is the sample size. We say the linear model is adequate in explaining the relationship between the dependent variable and one or more of the independent variables if:

$$F_{cal} > F_{\alpha}\left(k-1, n-k\right)$$

4.6. Hypothesis Testing

Hypothesis testing in a multiple regression can be tests on individual coefficients or tests on overall significance.

For example, testing the following hypothesis:

H_0: $\beta_i = 0$ against the alternative hypothesis

Ha: $\beta_i \neq 0$ i = 0, 1, 2, this is a test on individual coefficients.

This is significance test for individual parameter. This is similar to our unit three discussion. In multiple regression models the econometrician can perform various tests other than individual test of significance.

In general, let $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \ldots\ldots\ldots\ldots + \beta_k X_{ki} + U_i$ represent the more general formulation of the multiple regression model.

To test the hypothesis

H_0: $\beta_2 = \beta_3 = \beta_4 = \ldots\ldots\ldots = \beta_k = 0$ against the alternative hypothesis

Ha: Not all slope coefficients are simultaneously zero.

This test aims at finding out whether the explanatory variables do have any significance influence on the dependent variable. If the null hypothesis is true, then there is no linear relationship between the dependent variable and the explanatory variables.

To test this hypothesis we use the following test statistic:

$$F = \frac{\frac{RSS}{k-1}}{\frac{ESS}{n-k}} \quad F(k-1, n-k)$$

Where **RSS** represents the regression (explained) sum of squares, and **ESS** represents the error (residual) (unexplained) sum of squares.

The decision rule will be:

Compare the computed (calculated) **F-value** with the critical (table) value at the chosen level of significance, (k-1) for numerator and (n-k) for denominator degrees of freedom which is obtained from the F-distribution table. And decide based on the following procedures:

- ❖ If the computed F-value is greater than the critical value, reject the null hypothesis and accept that the regression is significant and not all coefficients are zero.
- ❖ On the other hand, if the computed F-value is less than the critical value obtained from the F-distribution table, then accept the null hypothesis, i.e. accept that the regression is not significant and all coefficients are zero.

We can illustrate the above discussion by taking the following simple multiple regression model suppose that $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + U_i$,

Note that K= 3. To test H_0: $\beta_2 = \beta_3 = 0$ against the alternative hypothesis which says at least one coefficient is non-zero, we can use F-test. The calculated F value is given by:

$$F = \frac{\frac{RSS}{2}}{\frac{ESS}{n-3}} \quad F(2, n-3)$$

The decision rule will be:

- If the calculated F > the critical F-value (table), then reject the null hypothesis.
- If the calculated F < the critical F-value (table), then do not reject the null hypothesis.

In other words, the test of overall significance may be conducted by following the procedure stated below.

- Compute the sum of squared deviations of the dependent variable $\sum (Y_i - \bar{Y})^2$. This is the total sum of squares (TSS).
- Compute the sum of squared deviations explained by all the explanatory variables $\sum (\hat{Y}_i - \bar{Y})^2$. This represents the explained sum of squares (RSS).

Compute the sum of squared of residual deviations, $\sum \hat{U}_i^2$ which is the error (residual) sum of squares (ESS). Remember from unit three discussions that total sum of squares (TSS) is the sum of explained sum of squares (RSS) and residual sum of squares (ESS).

$$TSS = RSS + ESS$$

$$\Sigma (Y_i - \bar{Y})^2 = \Sigma (\hat{Y}_i - \bar{Y})^2 + \Sigma (Y_i - \hat{Y}_i)^2$$

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

Now let us determine the degrees of freedom for each term. The degrees of freedom for the total sum of squares is (n-1), the degrees of freedom for explained sum of squares is (k-1) and the degrees of freedom for residual sum of squares is (n-k), where n is the total number of observations and k is the total number of parameters to be estimated including the constant term. To test the above kind of hypothesis, we follow the procedure below:

Compute the F-ratio using the following formula

$$F = \frac{\frac{\Sigma \hat{y}_i^2}{k-1}}{\frac{\Sigma \hat{U}_i^2}{n-k}} = \frac{(n-k)}{(k-1)} \frac{\Sigma \hat{y}_i^2}{\Sigma \hat{U}_i^2}$$

Dividing both sides by $\sum y_i^2$ we obtain

$$F = \frac{(n-k)}{(k-1)} \frac{\sum \hat{y}_i^2}{\Sigma \hat{U}_i^2} \cdot \frac{\sum y_i^2}{\sum y_i^2} = \frac{(n-k)}{(k-1)} \frac{\frac{\sum \hat{y}_i^2}{\Sigma y_i^2}}{\frac{\sum \hat{U}_i^2}{\sum y_i^2}}$$

Reject H_0 if the calculated F exceeds the table value

Relationship between F and R^2

The discussion in the preceding sub section point out that there is a relationship between F and R^2. That is, there is a relationship between the coefficient of determination and F-test. Recall that we said for (k-1) independent variables under the null hypothesis

$$H_0 : \beta_2 = \beta_3 = \beta_4 = ... = \beta_k = 0$$

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

Re-writing the above expression allows us to obtain the following result

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

If we divide both the numerator and denominator by TSS (total sum of squares), we get

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

But recall from unit three discussion that $R^2 = \frac{RSS}{TSS}$ *and* $\frac{ESS}{TSS} = \left(1 - \frac{RSS}{TSS}\right)$

Thus, substituting this relationship in the above function we obtain,

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

From this result, the following relationship between F and R^2 can be established:

- If $\mathbf{R^2}$ is equal to zero, then F will be zero and in this case don not reject the null hypothesis.
- If $\mathbf{R^2}$ is equal to one, then F will be infinite and in this case reject the null hypothesis.
- If $\mathbf{R^2}$ is higher, then F will be higher and in this case reject the null hypothesis.

Therefore testing the hypothesis

$$H_0 : \beta_2 = \beta_3 = \beta_4 = \ldots\ldots\ldots \beta_k = 0$$

$$Against\ the\ alternative\ hypothesis$$

$$H_a : Not\ all\ slope\ coefficients\ are\ simultaneously\ zero$$

is similar to testing the hypothesis

H_0: $\mathbf{R^2} = 0$ against the alternative hypothesis

Ha: $\mathbf{R^2} \neq 0$

The following example illustrates the discussion about individual and joint hypothesis testing

Example

Consider the following regression models;

$$\hat{Y}_i = -0.709 + 0.15X_1 + 0.08X_2$$
$$(2.683) \quad (0.03) \quad (0.1586)$$
$$R^2 = 0.935 \quad n = 10 \quad and \quad \hat{\sigma} = 0.8407$$

Note that the figures in the bracket measure the standard error of each estimate. Using the above information, we can test the overall goodness of fit at 5% level of significance. That is we can assess the joint significance of the model. The hypothesis to be tested in this case is

H_0: $\beta_1 = \beta_2 = 0$ against the alternative hypothesis

Ha: Not all slope coefficients are simultaneously zero.

To test this hypothesis, we use the following test statistic:

$$F = \frac{\frac{R^2}{k-1}}{\frac{(1-R^2)}{n-k}} \qquad F = \frac{\frac{0.935}{3-1}}{\frac{(1-0.935)}{10-3}} = \frac{0.4675}{0.00929} = 50.32$$

The (table) critical value at 0.05 level of significance and 2 and 7 degrees of freedom for numerator and denominator from the F- distribution table, i.e. $F_{0.05}(2.7) = 4.74$

The decision rule will be since F calculated value is greater than the F critical (k-1, n-k), then reject the null hypothesis. That is, since the computed F-value 50.32 is greater than the critical value 4.74, we reject the null hypothesis. Therefore, we reject the hypothesis that all slope coefficients are simultaneously zero.

Based on the result given, it is also possible to perform individual significance test. That is we can test

H_0: $\beta_1 = 0$ against the alternative hypothesis
Ha: $\beta_1 \neq 0$

And

H_0: $\beta_2 = 0$ against the alternative hypothesis
Ha: $\beta_2 \neq 0$

Note that this is test of significance of individual parameters. Recall that in this case, we use the following test statistic:

$$t = \frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}_{\beta k}} \quad t_{\alpha/2}\ (n-k)\ \text{degrees of freedom}\ \ k = 0, 1\ \text{and}\ 2$$

An estimator of the error variance σ^2 is:

$$\sigma^2 = \frac{\Sigma \varepsilon_i^{\,2}}{n-k} = \frac{\Sigma\left(Y_i - \hat{Y}\right)^2}{n-3}$$

Where,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_{1_i} + \hat{\beta}_2 X_{2_i}$$

k is the number of parameters to be estimated which in this case is 3.

Thus:

$$\sigma^2 = \frac{\Sigma \varepsilon_i^{\,2}}{n-3}$$

In order to estimate the variances of estimated regression coefficients, the coefficient of correlation between $\mathbf{X_1}$ and $\mathbf{X_1}$ must be estimated.

It can be shown that:

$$Var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\Sigma x_{1i}^{\,2}\left(1-r^2_{12}\right)}$$

$$Var\left(\hat{\beta}_2\right) = \frac{\sigma^2}{\Sigma x_{2i}^{\,2}\left(1-r^2_{12}\right)}$$

Where $\mathbf{r_{12}}$ is the coefficient of correlation between $\mathbf{X_1}$ and $\mathbf{X_2}$ that is:

$$r_{12} = \frac{\sum x_{1i}x_{2i}}{\sqrt{\left(\sum x_{1i}^{\ 2}\right)\left(\sum x_{2i}^{\ 2}\right)}}$$

Taking the square roots, we obtain the **standard errors**: $se\left(\hat{\beta}_1\right)$ and $se\left(\hat{\beta}_2\right)$.

The test for β_1 therefore, will be

$$t = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\beta_1}} = \frac{0.15 - 0}{0.03} = 5$$

The critical value at 0.05 level of significance and 7 degrees of freedom from the t-distribution table, i.e. $t_{0.025}(7) = 2.365$

The decision is since the computed t-value which equals 5 is greater than the critical value of 2.365, we reject the null hypothesis. Therefore, we accept the alternative hypothesis.

Similarly, we can test the significance of β_2. We use the following test statistic:

$$t = \frac{\hat{\beta}_2 - \beta_2}{\hat{\sigma}_{\beta_2}} = \frac{0.08 - 0}{0.1586} = 0.504$$

Note that, the critical value at 0.05 level of significance and 7 degrees of freedom from the t-distribution table, i.e. $t_{0.025}(7) = 2.365$

The decision, therefore, is to accept the null hypothesis, since the computed t-value 0.5 is less than the critical value 2.365.

4.7. Other Functional Forms (Linear Regression Model and the Non-linear Relationship)

For many economic theories, the assumption of linear relationship between the dependent variable and independent variable may not hold, rather the relationship may be non-linear forms. For example, cost functions are usually non-linear. Similarly, production functions exhibit nonlinear pattern. Other economic functions like demand, supply, income-consumption curves, etc. can also be non-linear.

For example, the traditional theory of cost curves may be approximated by a polynomial of third degree in output.

$$C = \beta_0 + \beta_1 Q - \beta_2 Q^2 + \beta_3 Q^3 + U$$

Where: C is cost
Q is output

A demand function represented by the following formula is also another example of regression model with non-linear relationship

$$Q = \beta_0 P^{\beta_1} Y^{\beta_2} U$$

Where: Q is the demand for a commodity
P is price of the commodity
Y is consumers' income

Such kind of non-linear relationship may be estimated by fitting non-linear functions and can be estimated by the method of ordinary least squares (OLS). However, there are other forms of relationship in which the relationship are non-linear in the parameters not in variables. Estimating such kind of relationship is more complex. In this section, we will focus on estimating a function which is non-linear in the variables but linear in the parameters, by making transformations of the data before the estimation of the parameters.

For example, to estimate the function $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + U$, we may set $Z = X^2$ and $W = X^3$ and transform the function into $Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 W + U$ and apply ordinary least squares to linear function to estimate the parameters.

Example

To illustrate the above discussion consider the following estimated cubic cost function,

$$\hat{C} = 124.2 + 50.4Q - 10.36Q^2 + 0.54Q^3$$

(3.4) (4.2) (0.24) (0.05)

Cov (β_2, β_3) = -0.02 and R^2 = 0.96 n = 10

Where C is total cost and Q is output.

Suppose we want to test the hypothesis that the coefficients of Q_2 and Q_3 are the same, i.e.

H_0: $\beta_2 = \beta_3$ against the alternative hypothesis
H_a: $\beta_2 \neq \beta_3$ at = 5% level of significance

To test this hypothesis, first we have to compute the following test statistic

$$t = \frac{(\hat{\beta}_2 - \hat{\beta}_3) - (\beta_2 - \beta_3)}{\hat{\sigma}_{\hat{\beta}_2 - \hat{\beta}_3}}$$

$$Where, \quad \hat{\sigma}_{\hat{\beta}_2 - \hat{\beta}_3} = \sqrt{Var(\hat{\beta}_2) + Var(\hat{\beta}_3) - 2Cov(\hat{\beta}_2, \hat{\beta}_3)}$$

$$t = \frac{-10.36 - 0.54}{\sqrt{(0.24)^2 + (0.05)^2 - 2(-0.02)}} = 34.49$$

The next step is finding the critical value at 0.05 level of significance, with 6 degrees of freedom from the t-distribution table for two tailed test, which is 2.447. Then by comparing the computed t-value with the critical value, we can give the conclusion. Since, the computed t-value 34.49 is greater than the critical value 2.447; we can reject the hypothesis that the coefficients of Q^2 and Q^3 in the cubic cost function are identical.

Suppose we want to estimate the function

$$Y=\beta_0 X_1^{\beta_1} X_2^{\beta_2} e^U$$

We may take logarithm to the base **e** on both sides of the given function and obtain the following function.

$$\log_e Y=\log_e \beta_0 + \beta_1 \log_e X_1 + \beta_2 \log_e X_2 + U$$

This function will retain the usual assumptions so that

$$E(U) = 0, \qquad Var(U_i) = E(U^2) = \sigma^2_u$$

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

$$E(UX_i) = 0$$

Now set $\log_e Y = Y^*$, $\log_e \beta_0 = \beta_0^*$, $\log_e X_1 = X_1^*$ and $\log_e X_2 = X_2^*$. This will transform the function into:

$$Y^* = \beta_0^* + \beta_1 X_1^* + \beta_2 X_2^* + U$$

Which is a linear function and we can apply ordinary least squares to obtain estimates of the parameters (after transforming **X** and **Y** into logarithms). Note that to get back the β_0 we use the backward transformation: $\beta_0 = 10^{\beta_0^*}$

The above transformed model is called **log-log, double-log or log-linear model**. One attractive feature of the log-log model, which has made it popular in applied work, is that the slope coefficient β_1 and β_2 measures the elasticity of Y with respect to X_1 and X_2.

Example

Suppose the estimated value $\beta_1 = 0.5$. This implies that a one percent increase in X_1 will result a 0.5% increase in Y assuming that X_2 is held constant. Similarly if $\beta_2 = 0.75$, it implies that a one percent increase in X_2 will result in a 0.75% increase in Y assuming that X_1 is held constant.

Consider the Cobb-Douglas production function which is given as follows:

$$Q=\beta_0 L^{\beta_1} K^{\beta_2} e^U$$

Where **Q** represents output: **L** represents labor input; and **K** represents capital input

To interpret the coefficients $\beta_1$and β_2, let's take logarithms to the base **e** on both sides. In this case, we will obtain the following.

$$\mathbf{ln\ (Q) = ln\ (}\beta_0\mathbf{) + }\beta_1\mathbf{ln(L) + }\beta_2\mathbf{ln(K) + U}$$

In this relationship, β_1 represents the percentage change in the level of output as a result of a one percentage change in the labor input while the effect of capital input is held constant. It is the elasticity of the factor input labor. Similarly, β_2 represents the percentage change in the level of output as a result of a one percentage change in the capital input while the effect of labor input is held constant. It is the elasticity of the factor input capital.

Note that, the marginal product of labor ($\mathbf{MP_L}$) is obtained by taking the partial derivatives of ln (**Q**) with respect to ln (**L**). It is given as:

$$MPL = \frac{\partial \ln(Q)}{\partial \ln(L)} = \beta_1$$

Similarly, the marginal product of capital ($\mathbf{MP_K}$) is obtained by taking the partial derivatives of ln (**Q**) with respect to ln (**K**). It is given as:

$$MPK = \frac{\partial \ln(Q)}{\partial \ln(K)} = \beta_2$$

Since the values of β_1 and β_2 are constant, the marginal product functions of the two factors are constant.

Chapter 5: Dummy Variable Regression Analysis

5.1. Definitions of Dummy Variables

In a regression analysis, the dependent variable is influenced not only by quantitative variables like income, output, price, cost, temperature…etc., but also by variables that are qualitative in nature like sex, race, color, religion, marital status, job category, region, season etc. These qualitative variables indicate the presence or absence of a "quality" or an "attribute" such as male or female, black or white…etc. such variables are called **dummy variables**.

We quantify such variables by artificially assigning the values of **0** and **1**, where **0** indicates one category and **1** indicates another category (like male = 0, female = 1), and use them in the regression equation together with the other independent variables.

5.2. ANOVA Models

Dummy variables can be used in regression model just as quantitative variables. A regression model may contain **only** qualitative explanatory variables. Such models are called **analysis of variance (ANOVA) models.**

Consider the following model:

$$Y_i = \alpha_1 + \alpha_2 D_i + U_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

$$Where\ Y_i\ represents\ the\ annual\ salary\ of\ a\ lecturer$$

$$D_i = 1\ if\ the\ lecturer\ is\ male$$

$$= 0, otherwise\ (i.e.\ if\ the\ lecturer\ is\ female)$$

This is a two-variable regression model, which enables us to check whether sex makes any difference in salary of a lecturer, if all variables which affect salary are held constant. By assuming the disturbance term satisfy the assumptions of classical linear regression model,

Average salary of male lecturer = $E(Y_i/D_i = 1) = \alpha_1 + \alpha_2 (1) = \alpha_1 + \alpha_2$

Average salary of female lecturer = $E(Y_i/D_i = 0) = \alpha_1 + \alpha_2 (0) = \alpha_1$

The intercept term $\boldsymbol{\alpha_1}$ gives the average salary of female lecturer and $\alpha_1 + \alpha_2$ gives the average salary of male lecturer. The slope $\boldsymbol{\alpha_2}$ gives the amount by which the average salary of male lecturer differs from the average salary of female lecturer. If $\boldsymbol{\alpha_2}$ is equal to zero, then there is no sex discrimination, i.e. if all other variables which affect salary are held constant, then the sex of a person cannot bring a change in the salary of a lecturer. A test of the hypothesis that there is no sex discrimination can be made by running regression using the method of ordinary least squares and use **t-test** to check whether the estimated coefficient $\boldsymbol{\alpha_2}$ is statistically significant or not. If the t-test shows that $\boldsymbol{\alpha_2}$ is statistically significant, then we reject the null hypothesis that the average salary of male lecturer is the same as the average salary of female lecturer ($\mathbf{H_0: \alpha_2 = 0}$).

Example

Consider the following estimated model:

$$Y = 12000 + 500 D_i$$

$$t = (50.4) \quad (10.1) \qquad n = 10$$

Where $\mathbf{D_i = 1}$, if the lecturer is male

= **0**, otherwise

Y_i is the salary of the lecturer

As this result shows, the estimated average salary of female lecturer is **12,000**. This holds when $\mathbf{D_i = 0}$. For male lecturer, the estimated average salary is 12,500. This is because $\mathbf{D_i = 1.}$ Since the coefficient of $\mathbf{D_i}$ is statistically significant (using t-test), the average salaries of the two categories are different and the average salary of female lecturer is lower than the average salary of male lecturer.

5.3. ANCOVA Models

Regression models in most economic research involve quantitative explanatory variables in addition to dummy variables. Such models are known as **analysis of covariance** (**ANCOVA**) models.

Take the previous example, which shows the relationship between average salaries of lecturers and sex and include another quantitative explanatory variable like years of teaching experience and re-write as:

$$Y_i = \alpha_1 + \alpha_2 D_i + \alpha_3 X_i + U_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

Where Y_i *represents the annual salary of a lecturer*

X_i *represents the teaching* exp*eriences of a lecturer*

$D_i = 1$ *if the lecturer is male*

$= 0$, *otherwise* (*i.e. if the lecturer is female*)

This model contains one qualitative variable (sex) and one quantitative variable (years of teaching experience). By assuming the disturbance term satisfy the assumptions of classical linear regression model,

Average salary of male lecturer = $E(Y_i/X_i, D_i = 1) = \alpha_1 + \alpha_2 + \alpha_3 X_i$
Average salary of female lecturer = $E(Y_i/X_i, D_i = 0) = \alpha_1 + \alpha_3 X_i$

You can observe that both male and female lecturers' salary functions in relation to the years of teaching experience have the same slope ($\boldsymbol{\alpha_3}$) but different intercepts ($\boldsymbol{\alpha_1}$ for female lecturers and $\boldsymbol{\alpha_1 + \alpha_2}$ for male lecturers). In other words, the level of the male lecturers' average salary is different from female lecturers' average salary by $\boldsymbol{\alpha_2}$ but the rate of change in the average salary due to a change in years of experience (slope) is the same for both sexes.

Under the assumption of common slope (α_3), a test of hypothesis that the two regression models for male and female lecturers have the same intercept, i.e. there is no sex discrimination can be made by running regression using the method of ordinary least squares on equation number **2** and using t-test we can check whether the estimated α_2 is statistically significant or not. If the **t-test** shows that α_2 is statistically significant, then we reject the null hypothesis that the male and female lecturers' level of average salary is the same ($\mathbf{H_0: \alpha_2 = 0}$)

So far, we include a qualitative variable with two categories like male versus female; however, qualitative variables may have more than two categories. For example, suppose we are interested in analyzing beer consumption. Factors affecting consumption are income, age, gender, season, race, religion, education and marital status and so on. In this example, variables such as: income, expenditure and age are quantitative variables whereas variables such as: gender, religion, race, marital status and season are qualitative variables. Education may be qualitative or quantitative depending on the value taken. For example, if one takes the number of years of schooling (year 1, 2, 3…) as a value, then education becomes a quantitative variable. Otherwise, if one take level of education like primary, secondary and tertiary as a value then education becomes a qualitative variable.

Some qualitative variables may take more than two categories. In this case, in order to avoid the dummy variable trap, we have to include **m-1** dummy variables if the qualitative variable has **m** categories. Otherwise, a perfect multicollinearity problem may arise and the method of ordinary least squares estimation is not possible.

Let's consider three mutually exclusive levels of education: less than high school, high school and college. Since we have three categories, we should introduce two dummies to take care of the three levels of education as:

$D_1 = 1$, if high school education
$= 0$, otherwise

$D_2 = 1$, if college education
$= 0$, otherwise

Consider the following model:

$$Y_i = \alpha_1 + \alpha_2 D_1 + \alpha_3 D_2 + \alpha_4 X_i + U_i \ldots\ldots\ldots\ldots\ldots \quad (3)$$

$Where\ Y_i\ represents\ the\ annual\ salary\ of\ a\ lecturer$

$X_i\ represents\ the\ teaching\ experiences\ of\ a\ lecturer$

$D_i = 1,\ if\ high\ school\ education$

$= 0,\ otherwise$

$D_2 = 1,\ if\ college\ education$

$= 0,\ otherwise$

In this model, the less than high school education category is the base category which is represented by **α_1**. The other coefficients **α_2** and **α_3** tell by how much the intercepts of the other two categories differ from the intercept of the base category. By assuming the disturbance term satisfies the assumptions of classical linear regression model.

Average salary for less than high school = $E(Y_i/D_1 = 0, D_2 = 0, X_i) = \alpha_1 + \alpha_4 X_i$

Average salary for high school complete = $E(Y_i/D_1 = 1, D_2 = 0, Xi) = (\alpha_1 + \alpha_2) + \alpha_4 X_i$

Average salary for college complete = $E(Y_i/D_1 = 0, D_2 = 1, Xi) = (\alpha_1 + \alpha_3) + \alpha_4 X_i$

Under the assumption of common slope (**α_4**), a test of hypothesis that the three regression models for less than high school, for high school and for college complete teachers have the same intercept can be performed. That is, employing the method of ordinary least squares on equation **3** and using t-test we can check whether the estimated **α_2** and **α_3** is statistically significant or not. Note that, this tests whether they are different from the base category or not. If the **t-test** shows that the differential intercepts **α_2** and **α_3** are individually statistically significant, then we reject the null hypothesis that there is no difference from the base category. Using **F-test**, it is also possible to test the hypothesis that **$\alpha_2 = \alpha_3 = 0$.**

Chapter 6: Econometric Problems

In this Chapter, we will see the method of assessing the reliability of the estimates of the parameters from econometric criteria point of view. Recall that, in the previous Chapters we said that after the estimation of the parameters with the method of ordinary least squares, we should assess the reliability of the estimates of the parameters based on three types of criteria before using the estimates for forecasting purpose. These are:

- A priori economic criteria which are determined by economic theory and related to the sign and magnitude of the parameters.
- Statistical criteria which are determined by the statistical theory.
- Econometric criteria which are determined by the econometric theory.

Further recall that, the statistical criteria are the coefficient of determination, the standard errors of the estimates and the related **t** and **F**-statistics. These tests are valid only if the assumptions of the linear regression model are satisfied. Thus, if the assumptions of an econometric model are violated, then the estimates obtained using the method of Ordinary Least Squares do not possess some or all of their optimal properties discussed in the earlier Chapters. Therefore, their standard error becomes unreliable criteria.

Econometric criteria provide evidence about the validity or the violation of the assumptions of the linear regression model. In this Chapter, therefore, we will see the violation of the assumptions, the sources of violation, the consequences of the violation of the assumptions on the parameters and on their standard error, the test available for each assumption and the solution that have been suggested as "remedies" of the situation created by the violation of the assumption.

6.1. Non-normality

Building a linear regression model is only half of the work. In order to actually be usable in practice, the model should conform to the assumptions of linear regression.

- Classical normal linear regression (CNLR) assumes that each U_i is distributed normally

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

Mean = $E(U_i) = 0$

Variance = $E(U_i^2) = \sigma^2$

Cov $(U_i, U_j) = E(U_i, U_j) = 0 \quad (i \# j)$

Note: For two normally distributed variables, the zero covariance or correlation means independence of them, so U_i and U_j are not only uncorrelated but also independently distributed. Therefore $U_i \sim NID(0, \sigma^2)$ is Normal and independently distributed.

Violations of normality create problems for determining whether model coefficients are significantly different from zero and for calculating confidence intervals for forecasts. Sometimes the error distribution is "skewed" by the presence of a few large outliers. Since parameter estimation is based on the minimization of squared error, a few extreme observations can exert a disproportionate influence on parameter estimates. Calculation of confidence intervals and various significance tests for coefficients are all based on the assumptions of normally distributed errors.

If the error distribution is significantly non-normal, confidence intervals may be too wide or too narrow.

Technically, the normal distribution assumption is not necessary if you are willing to assume the model equation is correct and your only goal is to estimate its coefficients and generate predictions in such a way as to minimize mean squared error. The formulas for estimating coefficients require no more than that, and some references on regression analysis do not list normally distributed errors among the key assumptions. But generally we are interested in making inferences about the model and/or estimating the probability that a given forecast error will exceed some threshold in a particular direction, in which case distributional assumptions are important. Also, a significant violation of the normal distribution assumption is often a "red flag" indicating that there is some other problem with the model assumptions and/or that there are a few unusual data points that should be studied closely and/or that a better model is still waiting out there somewhere.

6.2. Multicollinearity

6.2.1. The nature and causes of the Multicollinearity Problem

Multicollinearity is a term that is used to denote the presence of linear relationship among explanatory variables. If the explanatory variables are perfectly linearly correlated, i.e. if the correlation coefficient is one, then the parameters become indeterminate. In this situation, it is impossible to obtain numerical values for each parameter separately and the method of ordinary least squares does not hold.

If, on the other hand, the correlation coefficient for explanatory variables is equal to zero, then the variables are called **Orthogonal**, and there are no problems concerning the estimates of the coefficients. Orthogonal variables are the variables whose covariance is zero.

In the case of orthogonal variables, there is no need to perform a multiple regression analysis. Each parameter can be estimated by simple regression of Y on the corresponding regressor. In practice neither of the two extreme cases i.e. **Orthogonal** or perfect collinearity exist, rather there is some degree of inter correlation among the explanatory variables. Thus, the correlation coefficient for each pair of explanatory variables will have a value between zero and one.

Multicollinearity is not a condition that either exists or does not exist in economic variables but rather inherent in most economic relationships due to the interdependence of many economic variables. However, there is no consensus on the degrees of collinearity that affect the parameter estimates. In other words, Multicollinearity is a question of degree and not of its existence.

When any two explanatory variables are changing in the same way, it becomes difficult to measure the influence of each variable on the dependent variable.

Multicollinearity may arise for various reasons. These are

- There is a tendency of economic variables to move together over time. Economic variables may be influenced by the same factors and show the same pattern of behavior over time. For example, consumption, income, saving, investment and employment tend to rise in periods of economic expansion (boom) and decrease in periods of recessions. Hence growth and trend factors in time series are the most serious causes of multicollinearity.
- The use of lagged values of some explanatory variables as separate independent variables in the model. For example, in consumption function, past as well as the present values of income are included as explanatory variables. We know that the successive values of a variable are intercorrelated. Thus multicollinearity is almost certain to exist in distributed lag models.
- When the model has more explanatory variables than the number of observations (an over determined model), there may be a problem of multicollinearity.

Multicollinearity tends to be more serious problem in time series. However, it is quiet frequent in cross section data as well.

6.2.2. Consequences of Multicollinearity

If the correlation between the explanatory variables is perfect (**r = 1**), then the estimates of the coefficients are indeterminate and the standard errors of these estimates become infinitely large. If, on the other hand, the explanatory variables are not perfectly collinear but are correlated to a certain degree, then the effect of collinearity is uncertain.

Even if multicollinearity is strong, the estimates of the coefficients are unbiased i.e. the unbiasedness of the OLS estimates is not affected by correlation of explanatory variables. However, the instability of the estimates may be serious and even cause a change in the sign of the parameter estimates as the degree of collinearity increases, depending on the importance of each explanatory variable measured by correlation coefficient of dependent and independent variables.

When certain explanatory variables are more important than others and correlated with the dependent variable, the seriousness of the problem is greater. In general, multicollinearity is not necessarily a problem unless it is high relative to the overall degree of multiple correlations among all variables simultaneously, i.e.

Collinearity is harmful if $\mathbf{r^2\ X_i\ X_j \geq R^2y.X_1,X_2\ldots,X_k}$ where $\mathbf{r^2x_ix_j}$ is the simple correlation coefficient between any two explanatory variables X_i and X_j and $R^2y.x_1,x_2\ldots,x_k$ is the overall (multiple) correlation coefficient of the relationship.

In general, although there may be exceptions, increasing standard errors appear when we include correlated variables as explanatory variables in the function.

With multicollinearity, we may face the problem of mis-specification because we may reject a variable whose standard error appears high although this variable is an important determinant of the variations of the dependent variable. Therefore, multicollinearity results in the wrong decision and in the wrong specification of the model.

The concept of multicollinearity is best elaborated with the following example. Consider the following estimated three variable regression model:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

The effect on the regression coefficients of X_1 and X_2 are perfectly collinear is explained as follows. Note that $\hat{\beta}_1$ gives the rate of change in the average value of Y as X_1 changes by a unit holding X_2 constant. However, if X_1 and X_2 are perfectly collinear, then there is no way X_2 can be kept constant. So that when X_1 changes, X_2 also changes by a certain constant factor, i.e. there is no way to distinguish the separate influences of X_1 and X_2 on the dependent variable. As a result, there is no way to estimate the coefficients of X_1 and X_2 uniquely and we cannot get a unique solution for the individual regression coefficients.

Example:

Suppose **Y** = consumption, $\mathbf{X_1}$ = wealth, $\mathbf{X_2}$ = income

Consumption (Y)	**Income (X_1)**	**Wealth (X_2)**
90	100	500
140	150	750
190	200	1000
240	250	1250
280	300	1500

The relationship $\mathbf{X_2 = 5X_1}$ means that the wealth level for each individual is five times the level of his/her income. Suppose our aim is to see the consumption pattern of individuals at varying levels of income keeping wealth constant (that is, with the same wealth). This task needs data on individuals with the same wealth but different income. But this is not the case here since whenever income changes (varies) so does wealth (by five times). Thus, the task cannot be settled.

Question: Can we estimate $\mathbf{B_1}$ and $\mathbf{B_2}$?

We have seen earlier that the OLS estimator of $\mathbf{B_1}$ is:

$$\hat{\beta}_1 = \frac{\left(\Sigma x_{1i} y_i\right)\left(\Sigma x_{2i}^2\right) - \left(\Sigma x_{2i} y_i\right)\left(\Sigma x_{1i} x_{2i}\right)}{\left(\Sigma x_{1i}^2\right)\left(\Sigma x_{2i}^2\right) - \left(\Sigma x_{1i} x_{2i}\right)^2}$$

Since we have $\mathbf{X_2 = 5X_1}$ we can replace $\mathbf{X_2}$ by $\mathbf{5X_1}$**:**

$$\hat{\beta}_1 = \frac{\left(\Sigma x_{1i} y_i\right)\left(\Sigma (5x_{1i})^2\right) - \left(\Sigma 5x_{1i} y_i\right)\left(\Sigma x_{1i} 5x_{1i}\right)}{\left(\Sigma x_{1i}^2\right)\left(\Sigma (5x_{1i})^2\right) - \left(\Sigma x_{1i} 5x_{1i}\right)^2}$$

$$\hat{\beta}_1 = \frac{25\left(\Sigma x_{1i} y_i\right)\left(\Sigma x_{1i}^2\right) - 25\left(\Sigma x_{1i} y_i\right)\left(\Sigma x_{1i}^2\right)}{\left(\Sigma x_{1i}^2\right)\left(25\Sigma x_{1i}^2\right) - 25\left(\Sigma x_{1i}^2\right)^2} = \frac{0}{0}$$

Meaning $\hat{\beta}$ is indeterminate. Therefore, in the presence of perfect multicollinearity, the regression coefficients cannot be estimated.

Less than perfect multicollinearity (moderate to strong MC)

Consider the case when there is a high degree but not perfect MC. What happens to the parameter estimates when there is a high degree of MC?

The estimated coefficients are still unbiased, that is $E(\hat{\beta}_j) = \beta_j \qquad j=1, 2, 3$

We have seen earlier that the variance of $\hat{\beta}_2$ is given by:

$$Var(\hat{\beta}_2) = \frac{\sigma^2}{\Sigma x_{2i}^2 (1 - r_{23}^2)}$$

Where $\mathbf{r_{23}}$ is the coefficient of correlation between $\mathbf{x_2}$ and $\mathbf{x_3}$, and the variables $\mathbf{x_2}$ and $\mathbf{x_3}$ are expressed in deviation from their mean. It can clearly be seen that as the correlation between $\mathbf{x_2}$ and $\mathbf{x_3}$increases, that is, as $\mathbf{r_{23}}$tends towards one:

$$\Rightarrow r_{23}^2 \; approaches \; one$$

$$\Rightarrow 1 - r_{23}^2 \; approaches \; to \; zero$$

$$\Rightarrow \Sigma x_{23}^2 (1 - r_{23}^2) \; approaches \; to \; zero$$

$$\Rightarrow Var(\hat{\beta}_2) \; becomes \; very \; large$$

Particularly, if $\mathbf{r_{23}} = 1$, then the variances become infinite.
Recall that to test whether each of the coefficients is significant or not, i.e,
H_0: $\beta_j = 0$
H_A: $\beta_j \neq 0$
The test statistic is:

$$t_j = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

Where $se(\hat{\beta}_j) = \sqrt{Var(\hat{\beta}_j)}$

Thus, under a high degree of MC, the test statistic will be very small number. This often leads to accepting the null hypothesis when in fact the parameter is significantly different from zero.

Major implications of a high degree of multicollinearity

1. **OLS** coefficient estimates are still unbiased.
2. **OLS** coefficient estimates will have large variances.
3. There is a high probability of accepting the null hypothesis of zero coefficient (using the **t-test**) when in fact the coefficient is significantly different from zero.

4. The regression model may do well, that is $\mathbf{R^2}$ may be quite high.
5. The **OLS** estimates and their standard errors may be quite sensitive to small changes in the data.

6.2.3. Tests for detecting multicollinearity

Detection of multicollinearity

Multicollinearity almost always exists in most applications. So the question is not whether it is present or not; it is a question of degree. Multicollinearity is not a statistical problem; it is a data (sample) problem. Therefore, we do not "test for MC"; but measure its degree in any particular sample (using some rules of thumb).

The seriousness of the effects of multicollinearity depends on the degree of correlation between variables and on the overall correlation coefficient. Thus, the standard errors, the partial correlation coefficients and the overall correlation coefficients may be used for testing multicollinearity.

Let's examine some of these rules of detecting or measuring the degree of multicollinearity as follows:

- High $\mathbf{R^2}$ but few significant **t-ratios**. If $\mathbf{R^2}$ is high, then the **F-test** will reject the hypothesis that the partial slope coefficients are simultaneously equal to zero, but the individual **t-tests** may show none or very few of the partial slope coefficients are statistically different from zero. In this case the problem of multicollinearity is serious.
- High pair wise correlations among regressors. If the pair wise correlation coefficient between two regressors is high (in excess of 0.8), then multicollinearity is a serious problem.
- Auxiliary regressions: Variance inflation factor (VIF)

The VIF for each estimated regression coefficient is defined as:

$$VIF(\hat{\beta}_j) = \frac{1}{1-R_j^2}$$

Where $\mathbf{R_j^2}$ is the coefficients of determination obtained when the j^{th} variable is regressed on the remaining X variables (called auxiliary regression).

a) If $VIF(\hat{\beta}_j)$ exceeds 10, then $\hat{\beta}_j$ is poorly estimated because of MC. Or the j^{th} regressor variable (X_j) is responsible for MC.
b) **Klien's rule**: MC is troublesome if any of the $\mathbf{R_j^2}$ exceeds the overall $\mathbf{R^2}$.

6.2.4. Solutions for multicollinearity

The solutions for multicollinearity depend on the severity of multicollinearity, on availability and sources of data, on the importance of factors which are multicollinear and on the purpose for which the model is being estimated.

If multicollinearity affects some of the less important factors (variables), one may exclude these factors from the model. If, on the other hand, multicollinearity has serious effects on the coefficient estimates of important factors, then take one of the following corrective solutions.

1. Increase the size of the sample

Multicollinearity may be avoided or reduced if we increase the size of the sample by gathering more observations. By increasing the sample size, the covariances among estimated parameters resulting from multicollinearity can be reduced. This is because covariances are inversely proportional to sample size. This is true only if multicollinearity is due to errors of measurement and multicollinearity exists only in the sample but not in the population. If the populations of the variables are multicollinear, then an increase in the size of the sample will not help in reduction of multicollinear relations among the variables.

2. Substitution of lagged variables for other explanatory variable (distributed lag models)

Some economic variables may be determined not only by the current values of the explanatory variables but also by past values of these variables. The successive values of any explanatory variable are highly correlated. In this case, multicollinearity may be avoided by adopting Koyck's suggestion of substitution of the lagged value of X for single lagged values of dependent variable. Instead of having all the lagged values of the explanatory variables, we use lagged values of the dependent variable which are expected to be less correlated than the lagged values of explanatory variables.

3. Introduction of additional equations in the model.

Multicollinearity may be overcome by introducing additional equations into our model to express fully the relationships between the multicollinear explanatory variables. By explicitly formulating these relationships, we can form a simultaneous equation technique. This will reduce the problem of multicollinearity.

4. Dropping a variable(s) from the model

When there is a multicollinearity problem, the simplest solution may be to drop one or more of the collinear variables. However, dropping variables from the model may lead to model specification error. Therefore, in reducing the severity of the collinearity problem by dropping variable(s), we may obtain biased estimates of the coefficients retained in the model.

5. Transformation of variables

When there is a multicollinearity problem, transformation of variables included in the model can minimize the problem of collinearity.

6.3. Heteroscedasticity

6.3.1. Definition and Sources of Heteroscedasticity

An important assumption of the classical linear regression model is that the population disturbances term, U_i are homoscedastic i.e. they all have the same variance. Which means "equal scatter" (of the error terms u_i around their mean, 0). Equivalently, this means that the dispersion of the observed values of Y around the regression line is the same across all observations.

This holds since $E(U_i) = 0$. That is,

$$Var(U_i) = E[U_i - E(U_i)]^2 = E(U_i^2) = \sigma^2,$$

This is the assumption of homoscedasticity. It suggests that the conditional variance of the dependent variable conditional upon the given value of the explanatory variable remains the same regardless of the values taken by the variable X. On the other hand, when the conditional variance of the dependent variable increases as the value of the explanatory variable increases, there is heteroscedasticity, i.e.

$$Var(U_i) \neq Var(U_j)$$

The problem of heteroscedasticity is likely to be more common in cross sectional than time series data. In cross- sectional data, members of a population like individual consumer, firms; industries etc are considered at a given time. These members may be of different sizes such as small, medium or large sizes. In time series data, the variables tend to be similar and collect the data for the same entity over a period of time.

Consider the case of data on income and expenditure of individual families. Here the assumption of homoscedasticity is not very reasonable since we expect less variation in consumption for low income families than for high income families. At low levels of income, the average level of consumption is low and the variation around this level is restricted: consumption cannot fall too

far below the average level because this might mean starvation, and it cannot rise too far above the average level because the asset does not allow it. These constraints are likely to be less binding at higher income levels.

6.3.2. Consequences of Heteroscedasticity

The following are the consequences of the presence of heteroscedasticity in a regression model.

- Even in the presence of heteroscedasticity the estimators are linear and unbiased. Thus, in a repeated sampling, on average, the value of the estimator will be equal to the true population parameter (consistency).

Consider the following model (in deviation form).

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i(\beta x_i + \varepsilon_i)}{\sum x_i^2} = \beta \frac{\sum x_i^2}{\sum x_i^2} + \frac{\sum x_i \varepsilon_i}{\sum x_i^2} = \beta + \frac{\sum x_i \varepsilon_i}{\sum x_i^2}$$

$$\Rightarrow E(\hat{\beta}) = \beta + \frac{\sum x_i \, E(\varepsilon_i)}{\sum x_i^2} = \beta \qquad (\text{since } E(\varepsilon_i) = 0)$$

Thus, $\hat{\beta}$ is unbiased in the presence of heteroscedasticity.

If β_1 is OLS estimator and β_1* is weighted least square estimator in the presence of heteroscedasticity, then β_1* is efficient i.e. it has the smallest variance than β_1 (OLS estimator allowing for heteroscedasticity).

- We cannot use OLS estimators to establish confidence interval and test hypothesis using the usual t and **F** statistics. This is because the estimated variances of the OLS estimators are biased (higher than the variances of other methods of estimation) (such as weighted least squares-WLS). This means that the confidence intervals based on OLS will be unnecessarily larger. As a result **t** and **F** tests are likely to give inaccurate results because the variance is large and the test will give statistically insignificant coefficients (**t**-test is smaller).

Therefore, the formula for the estimates and their variances, tests and confidence intervals using it are invalid or inappropriate.

The prediction of the dependent variable for a given new observation on the X's is inefficient using the method of ordinary least squares estimators and standard errors.

6.3.3. Detection of Heteroscedasticity

In this section, we try to get answer for the question how does one know whether heteroscedasticity is present or not? There are no hard and fast rules for detecting heteroscedasticity because σ_i^2 can be known only if we have the entire population of the dependent variable corresponding to the chosen X's. However, with this limitation, let us examine some of the formal and informal methods of detecting heteroscedasticity.

A. Informal method

Under the informal method of detection of heteroscedasticity, let us examine the graphical method. This method involves plotting of estimated residuals obtained by applying OLS against the explanatory variables. Plot U_i^2 against each explanatory variable or against the estimated value of Y_i ($\hat{Y}$) and check whether the estimated mean value of Y is systematically related to the squared residual.

B. Formal Methods

1. White's test

The test is based on the regression of $\hat{\varepsilon}_i^2$ on all the explanatory variables (X_i), their squares (X_i^2), and all their cross products. E.g., when the model contains p = 2 explanatory variables, the test is based on an estimation of the model: $\hat{\varepsilon}_i^2 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + u_i$ and calculating the coefficient of determination R^2_W. The test statistic is:

$\chi^2_{Cal} = nR^2_W$, where **n** is the number of observations. Consider the model $Y_i = \beta_1 + \beta_2 X_i$. First we have to regress Y_i on X_i and obtain the residuals. Then square the residuals and regress $\hat{\varepsilon}_i^2$ on X_i and X_i^2 to obtain R^2_W.

Decision rule: reject the null hypothesis H_0: $B_1 = B_2 = B_3 = B_4 = B_5 = 0$, the hypothesis of homoscedasticity, if the above test statistic exceeds the value from the Chi-square distribution with **p** degrees of freedom for a given level of significance **α**. If our model has only one independent

variable X_i, then p = 2: X_i, X_i^2. If we have two independent variables X_{1i} and X_{2i}, then p = 5: X_{1i}, X_{2i}, $X_{1i}X_{2i}$, X_{1i}^2, X_{2i}^2. And so on.

2. Goldfeld - Quandt test:

This method is applicable if one assumes that heteroscedastic variance σ_i^2 is positively related to one of the explanatory variables in the regression model. Moreover, it is applicable for large samples. In other words, the observation must be at least twice as many as the parameters to be estimated. The test assumes normality and serially independent disturbance term.

Here we are going to test the following hypothesis.

H_0: Var (U_i) = σ^2 (homoscedastic) against the alternative hypothesis

Ha: Var (U_i) = σ_i^2 (heteroscedastic)

Suppose we have a model with one explanatory variable X.

To test the hypothesis, Goldfeld and Quandt suggest the following steps.

Step 1: Order or rank the observations according to the value of X_i beginning with the lowest X-value.

Step 2: Divide the observations into three parts: $\mathbf{n_1}$ observations in the first part, **p** observations in the middle part, and $\mathbf{n_2}$ observations in the second part ($\mathbf{n_1 + n_2 + p = n}$). Usually **p** is taken to be **one-sixth** of **n**.

Step 3: Run a regression on the first $\mathbf{n_1}$ observations, obtain the residuals, $\hat{\varepsilon}_i$ and calculate the residual variance $s_1^2 = \frac{\sum \hat{\varepsilon}_i^2}{n_1 - 2}$. Similarly, run a regression on the second $\mathbf{n_2}$ observations, obtain the residuals $\hat{\varepsilon}_i$, and calculate the residual variance $s_2^2 = \frac{\sum \hat{\varepsilon}_i^2}{n_2 - 2}$

Or obtain the respective residuals sum of squares $\mathbf{ESS_1}$ and $\mathbf{ESS_2}$, respectively. $\mathbf{ESS_1}$ represents the residual sum of squares corresponding to the smaller X_i values (small variance group) and $\mathbf{ESS_2}$ represents the residual sum of squares corresponding to the larger X_i values (larger variance group).

Step 4: Calculate the test statistic: $F_{cal} = \frac{s_2^{\,2}}{s_1^{\,2}} = \frac{\frac{ESS_2}{df}}{\frac{ESS_1}{df}} = \frac{ESS_2}{ESS_1}$

Where,

- **df** represents the degrees of freedom
- **ESS** refers to the residual sum of squares.

Reject the null hypothesis $H_0: \sigma_1^{\,2} = \sigma_2^{\,2}, \, if:$

$$F_{cal} > F_{\alpha}(n_1 - 2, \, n_2 - 2)$$

Note that rejecting the null means the errors are heteroscedastic.

The omitted **C** **c**entral observations are observations to sharpen the difference between small variance and large variance groups. The ability of this test to do successfully depends on how **C** is chosen.

If there are more than one variable in the model, the ranking of observations can be done according to any one of them. If we are not a prior sure which variable is appropriate, then we can conduct the test on each of the explanatory variables.

Consider the following example. Suppose that we have data on expenditure on durable goods in relation to monthly income for 30 individuals. Suppose expenditure is linearly related to income but we suspected the presence of heteroscedasticity in the data. Suppose further that the middle **8** observations are dropped after the necessary reordering of the data. Suppose we obtain the following result after we perform a separate regression based on the two **11** observations.

$$F = \frac{\frac{126}{9}}{\frac{10}{9}} = 12.6$$

Note from the F-table that the critical F value for 9 numerator and 9 denominator df at the 5% level is 3.18. Since the estimated F value exceeds the critical value, we may conclude that there is heteroscedasticity in the error variance.

C. Breusch-Pagan test

This involves applying **OLS** to:

$\frac{\hat{\varepsilon}_i^2}{\hat{\sigma}^2} = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + ... + \gamma_k X_{ki} + u_i$ and calculating the regression sum of squares (RSS). The test statistic is: $\chi^2_{cal} = \frac{RSS}{2}$

Decision rule: reject the null hypothesis $\gamma_1 = \gamma_2 = ... = \gamma_k = 0$ if the test statistic exceeds the value from the Chi-square with k degrees of freedom for a given value of **α**.

Example

Consider the following data on consumption expenditure (Y) and income (X) for 20 households (both in thousands of Dollars):

Household	Income	Expenditure	Household	Income	Expenditure
1	22.3	19.9	11	8.1	8
2	32.3	31.2	12	34.5	33.1
3	36.6	31.8	13	38	33.5
4	12.1	12.1	14	14.1	13.1
5	42.3	40.7	15	16.4	14.8
6	6.2	6.1	16	24.1	21.6
7	44.7	38.6	17	30.1	29.3
8	26.1	25.5	18	28.3	25
9	10.3	10.3	19	18.2	17.9
10	40.2	38.8	20	20.1	19.8

Applying OLS we get the following results:

$Y_i = 0.847 + 0.899X_i$

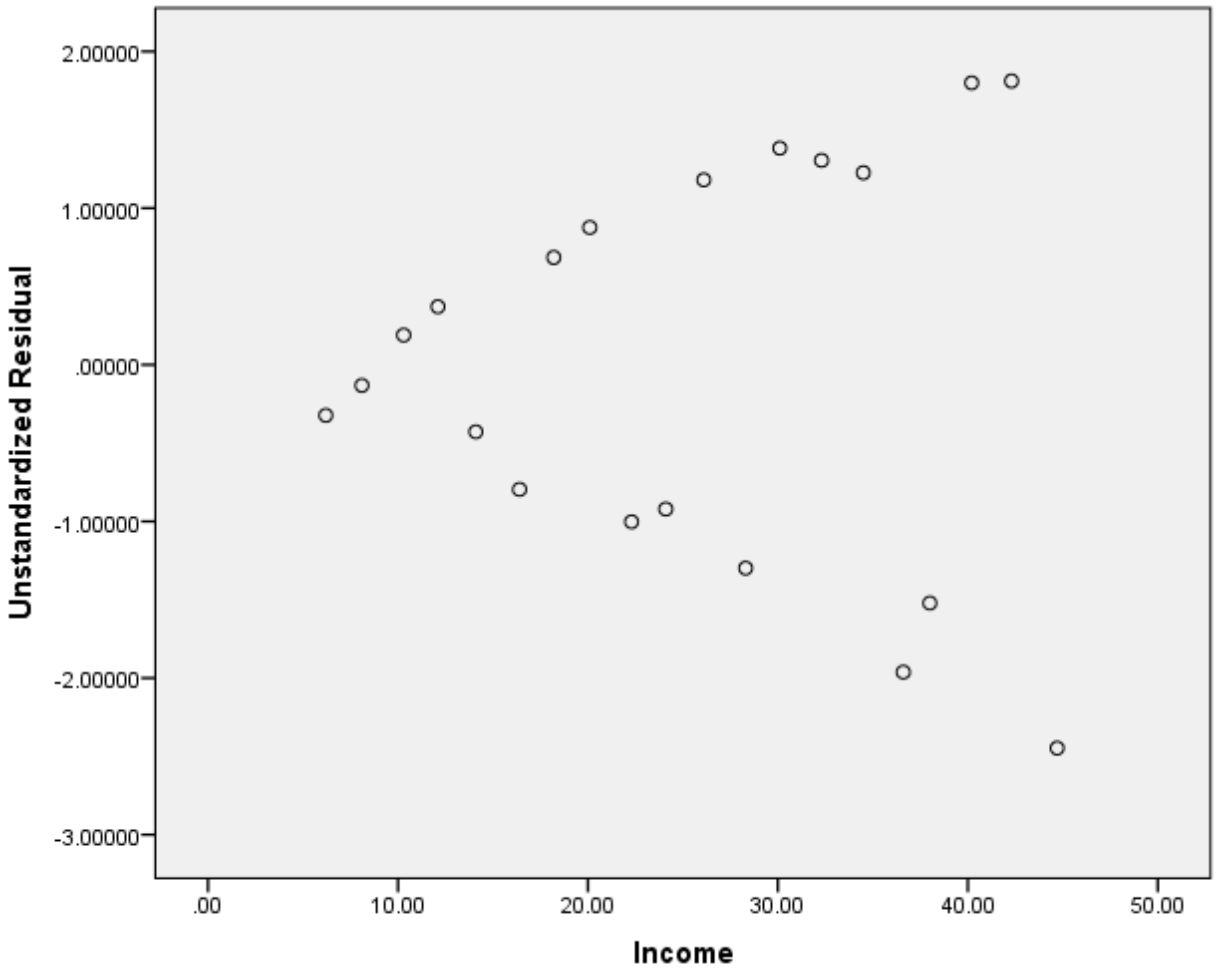
(1.204) (35.534) the figures in parenthesis represent t-values.

$R^2 = 0.986$ $\hat{\sigma}^2 = 1.726$

Error (residual) sum of squares = 31.074, degrees of freedom = 20-2 = 18

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}_i^2}{n-2} = \frac{31.074}{18} = 1.726$$

A plot of residuals $\hat{\varepsilon}_i$ against the values of the explanatory variable X_i is shown below.



It can clearly be seen that the scatter of the residuals (i.e., the variance of the residuals) increases with X_i). This is an indication of heteroscedasticity problem. However, we should not come to a conclusion until we apply formal tests of the hypothesis of homoscedasticity.

1. Goldfeld-quandt test

In order to apply this test, we should first order the observations based on the absolute magnitude of the explanatory variable X. We then divide the data into three parts: $n_1 = 8$, $p = 4$ and $n_2 = 8$. Note that the variances of the last several disturbances in the first part are likely to be similar to

those of the first several disturbances in the second part. To increase the power of the test, it is recommended that the two parts be some distance apart. Thus, we drop the middle p = 4 residuals together. We then run a separate regression on the first and the second parts, and calculate the residual variance for each of the two parts. The results are:

Error sum of squares for the first 8 observations = 1.893

Error sum of squares for the second 8 observations = 20.3

$$S_1^{\,2} = \frac{\sum \hat{\varepsilon}_i^{\,2}}{n_1 - 2} = \frac{1.893}{8-2} = 0.3155 \qquad S_2^{\,2} = \frac{\sum \hat{\varepsilon}_i^{\,2}}{n_2 - 2} = \frac{20.3}{8-2} = 3.383$$

Calculate the Goldfeld-Quandt test statistic as:

$$\frac{S_2^{\,2}}{S_1^{\,2}} = \frac{3.383}{0.3155} = 10.72$$

We compare this value with F_α (n_1-2, n_2-2) for a given level of significance α.

For α = 0.01, $F_{0.01}$ (6, 6) = 8.47

For α = 0.05, $F_{0.05}$ (6, 6) = 4.28

Decision: Since F_{Cal} = 10.72 is greater than both tabulated values, we reject the null hypothesis of homoscedasticity at both 1 % and 5 % significance levels.

2. The White test

This involves applying OLS to: $\hat{\varepsilon}_i^{\,2} = \delta_0 + \delta_1 X_i + \delta_2 X_i^{\,2} + u_i$ and computing the coefficient of determination R^2_W. These yields R^2_W = 0.878. The White test statistic is:

$$\chi^2_{Cal} = nR^2_W = 20\ (0.878) = 17.56$$

We compare this value with $\chi^2_\alpha(p)$ for a given level of significance α.

• For α = 0.01, $\chi^2_{0.01}(2)=9.210$

For α = 0.05, $\chi^2_{0.05}(2)=5.991$

Decision: Since $\chi^2_{Cal}=17.56$ is greater than both tabulated values, we reject the null hypothesis of homoscedasticity at both 1% and 5% significance levels.

6.3.4. Solutions (corrections) for Heteroscedasticity

A. When the error variance is known

Consider the model:

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$$

$Where\ E(\varepsilon_i^2) = \sigma_i^2$ And σ_i^2 is known for i = 1, 2, …, n. We make the following transformation:

$$\frac{Y_i}{\sigma_i} = \beta_1\left(\frac{1}{\sigma_i}\right) + \beta_2\frac{X_i}{\sigma_i} + \frac{\varepsilon_i}{\sigma_i}$$

$$\Rightarrow Y_i^* = \beta_1^* + \beta_2 X_i^* + \varepsilon_i^* \ldots\ldots (*)$$

$Where\ \varepsilon_i^* = \frac{\varepsilon_i}{\sigma_i}$. The transformed error term ε_i^* is homoscedastic since:

$$E(\varepsilon_i^{*2}) = E\left(\frac{\varepsilon_i^2}{\sigma_i^2}\right) = \frac{E(\varepsilon_i^2)}{\sigma_i^2} = \frac{\sigma_i^2}{\sigma_i^2} = 1$$

Thus we can apply OLS to equation (*) to get regression coefficient estimates that are BLUE. This estimation method is known as **weighed least squares (WLS)** since each observation is weighted (multiplied) by $\frac{1}{\sigma_i}$. The major difficulty with WLS is that σ_i^2 are rarely known. Sometimes we can estimate σ_i^2 from the sample.

B. When error variances vary directly with an independent variable

Suppose the variance of the i^{th} observation is proportional to the square of the explanatory variable X_i, that is, $Var(\varepsilon_i) = \sigma^2 X_i^2$. We can transform the model as:

$$\frac{Y_i}{X_i}=\left(\frac{\beta_1}{X_i}\right)+\beta_2+\frac{\varepsilon_i}{X_i}=\beta_2+\left(\frac{\beta_1}{X_i}\right)+\frac{\varepsilon_i}{X_i}$$

$$\Rightarrow Var\left(\frac{\varepsilon_i}{X_i}\right)=E\left(\frac{\varepsilon_i}{X_i}\right)^2=\frac{E\left(\varepsilon_i^{\,2}\right)}{X_i^{\,2}}=\frac{\sigma^2 X_i^{\,2}}{X_i^{\,2}}=\sigma^2$$

Hence, the variance of the disturbance term is constant, and we can apply OLS by regressing $\frac{Y_i}{X_i} on \frac{1}{X_i}$. Note that the estimated constant term and slope in the transformed model will be the values of $\hat{\beta}_2$ *and* $\hat{\beta}_1, respectively.$

Weighted least squares (WLS)

All of the tests indicate that the disturbances are heteroscedastic. Thus, the regression coefficients obtained by OLS are not efficient. In such cases, we have to apply weighted least squares (WLS) estimation. The weights can be obtained from the sample at hand or from some prior knowledge. In our example we will estimate the weights $\left(\sigma_i\right)$ from the sample.

First order the data based on the absolute magnitude of the explanatory variable (income) and apply OLS estimation and obtain the residuals $\left(\hat{\varepsilon}_i\right)$. We then order the residuals based on the absolute magnitude of the explanatory variable (income). Next we divide the residuals into three parts: the first and second parts consisting of seven residuals and the third part consisting of six residuals. The variance of each part is computed as: $\hat{\sigma}_i^{\,2}=\frac{1}{n_i}\Sigma\hat{\varepsilon}_i^{\,2}$, where n_i is the number of residuals in the i^{th} part, i = 1, 2, 3. The results are:

$$\hat{\sigma}_1^{\,2}=0.225894\Rightarrow\hat{\sigma}_1=0.475283$$

$\hat{\sigma}_2^2 = 1.330623 \Rightarrow \hat{\sigma}_1 = 1.153526$

$\hat{\sigma}_3^2 = 3.363032 \Rightarrow \hat{\sigma}_1 = 1.833857$

The next step is to divide the values of the dependent variable, the independent variable and the constant term (a vector 1's) in the i[th] part by $\hat{\sigma}_i$:

$$\frac{Y_i}{\hat{\sigma}_i} = \alpha\left(\frac{1}{\hat{\sigma}_i}\right) + \beta\left(\frac{X_i}{\hat{\sigma}_i}\right) + U_i \quad \textit{Where} \quad U_i = \frac{\varepsilon_i}{\hat{\sigma}_i}.$$

$\frac{1}{\hat{\sigma}_i}$	$\frac{X_i}{\hat{\sigma}_i}$	$\frac{Y_i}{\hat{\sigma}_i}$
2.1040	13.0449	12.8345
2.1040	17.0425	16.8321
2.1040	21.6713	21.6713
2.1040	25.4585	25.4585
2.1040	29.6665	27.5625
2.1040	34.5058	31.1393
2.1040	38.2929	37.6618
0.8669	17.4248	17.1648
0.8669	19.3320	17.2515
0.8669	20.8925	18.7252
0.8669	22.6263	22.1061
0.8669	24.5335	21.6727
0.8669	26.0939	25.4004
0.8669	28.0011	27.0475
0.5453	18.8128	18.0494
0.5453	19.9579	17.3405
0.5453	20.7214	18.2675
0.5453	21.9210	21.1576
0.5453	23.0661	22.1937
0.5453	24.3749	21.0485

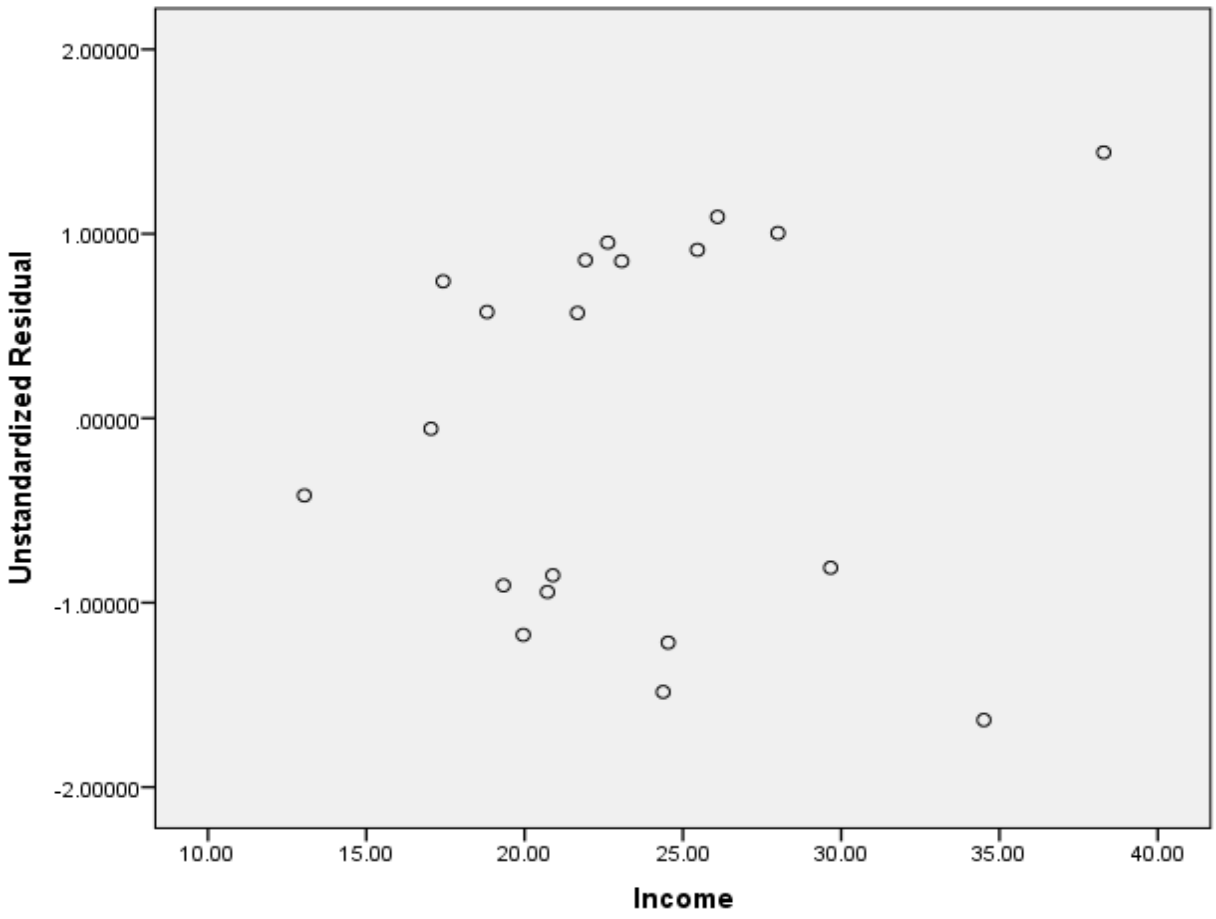
We then run an OLS regression of $\frac{Y_i}{\hat{\sigma}_i}$ on $\frac{1}{\hat{\sigma}_i}$ and $\frac{X_i}{\hat{\sigma}_i}$ without a constant term. The results are:

$$Y_i = 0.659 \; + \; 0.910 X_i$$

$$(1.826) \quad (44.048)$$

$$R^2 = 0.998 \qquad \hat{\sigma}^2 = 1.094$$

The plot of the residuals of the transformed model against the explanatory variable (income) is shown below. It can be seen that the spread of the residuals has no increasing or decreasing pattern, i.e., there is no heteroscedasticity.



6.4. Autocorrelation

One of the assumptions of ordinary least squares is that the successive values of the random variable U are independent, i.e. the value which U assumes in any one period is independent from the value which it assumed in any other period. This assumption implies that the covariance of u_i and u_j is equal to zero

$$\text{Cov}(U_i U_j) = E[(U_i - E(U_i))(U_j - E(U_j)] = 0$$

This implies that,

$$E(U_i U_j) = 0, \qquad \text{since } E(U_i) = E(U_j) = 0$$

If this assumption is satisfied, then the U's does not exhibit serial correlation or no autocorrelation between elements of U. This means that when observations are made over time, the effect of the disturbance occurring at one period does not carry-over into another period. However, if this assumption is not satisfied, then we say that there is autocorrelation or serial correlation of the random variable U. In this case, the value of U in any particular period is correlated with its own preceding or succeeding element value.

Autocorrelation is a special case of correlation where the association is not between elements of two or more variables but between successive values of one variable, while correlation refers to the relationship between values of two or more different variables. Autocorrelation is a common problem in econometrics but the problem is serious when time series data is considered.

In addition to autocorrelation in random term, there may be also autocorrelation in most economic variables. However, in this section we will deal with the autocorrelation problem of the random variable U.

For instance, in a study of the relationship between output and inputs of a firm or industry from monthly observations, non-AC of the disturbance implies that the effect of machine breakdown is strictly temporary in the sense that only the current month's output is affected.

In case of cross-sectional data such as those on income and expenditure of different families, the assumption of non-AC means that if the expenditure behavior of one family is "disturbed" (for example by the visit of a relative or due to a wedding party), then this does not affect the expenditure behavior of any other family. Thus, it seems more plausible that the assumption is violated in case of time series data than cross-sectional data.

Another case where AC arises is when a lagged value of a variable is used as a regressor. For instance, if past consumption is used as a regressor in a consumption model, then this always

results in error AC. Intuitive reasoning: this month's spending is correlated with last month's spending due to habit formation over time.

There are a number of time series patterns or process that can be used to model correlated errors. The most common is what is known as "the first order autoregressive process" or AR (1).

Let U_t denotes the value that U assumes in period t and U_{t-1} denotes the value that U assumes in period t-1.

Assume there is a linear relationship between any two successive values of the random variable U and the model is given as follow:

$$U_t = \rho U_{t-1} + V_t \quad \text{.....} \quad (1)$$

Where: V_t is stochastic such that it satisfies the standard Ordinary Least Squares (OLS) assumptions, namely:

$$E(V_t) = 0$$

$$Var(V_t) = \sigma^2$$

$$Cov(V_t, V_{t+s}) = 0$$

Where the subscript "S" represent the exact period of lag.

Equation (1) represents a **first order autoregressive relationship**. A measure of the first order linear autocorrelation is given by the following autocorrelation coefficient:

$$r_{Ut\,Ut-1} = \frac{\Sigma \hat{U}_i \hat{U}_{t-1}}{\sqrt{\Sigma \hat{U}_t^{\,2} \; \Sigma \hat{U}_{t-1}^{\,2}}} \quad \text{.....} \quad (2)$$

Where $r_{Ut\,Ut-1}$ is an estimate of the population autocorrelation coefficient $\rho_{Ut\,Ut-1}$ which measures the correlation of the true population of U's.

If the value of U in any period depends on its own value in the preceding period alone, then we say that the U's follow a first order autoregressive scheme and the relation is in the form of:

$$U_t = f(U_{t-1})$$

If, on the other hand, the value of U depends on the values of the two previous periods, then we say that the U's follow a second order autoregressive scheme and so on and the relation is in the form of:

$$U_t = f(U_{t-1}, U_{t-2})$$

In most cases, we have a first order autoregressive autocorrelation with a linear relationship between successive values of U's and are given as:

$$U_t = \rho_1 U_{t-1} + V_t$$

Where: ρ_1 is the coefficient of autocorrelation relationship.

V_t is a random variable satisfying all assumptions of classical regression model, namely

$$E(V) = 0 \qquad E(V^2) = \sigma_V{}^2 \qquad E(V_i\, V_j) = 0$$

If we apply the method of ordinary least squares, we obtain

$$\hat{\rho}_1 = \frac{\Sigma \hat{U}_i\, \hat{U}_{t-1}}{\Sigma \hat{U}_{t-1}{}^2} \quad \text{.....} \quad (3)$$

And autocorrelation coefficient $\rho_{Ut\,Ut-1}$ is given by

$$\rho_{U_t U_{t-1}} = \frac{\Sigma \hat{U}_i \hat{U}_{t-1}}{\sqrt{\Sigma \hat{U}_t^2 \; \Sigma \hat{U}_{t-1}^2}}$$.

For large sample size, $\Sigma \hat{U}_t^2 \cong \Sigma \hat{U}_{t-1}^2$ hence $\rho_{U_t U_{t-1}}$ will be given by

$$\rho_{U_t U_{t-1}} = \frac{\Sigma \hat{U}_i \hat{U}_{t-1}}{\Sigma \hat{U}_{t-1}^2} \qquad \text{...} \quad (4)$$

This implies that $\rho_{U_t U_{t-1}} = \hat{\rho}_1$

Because of this, the first order autoregressive model is given as:

$$U_t = \rho U_{t-1} + V_t$$

Where: ρ is the first order autocorrelation coefficient. If $\rho = 0$, then U_t is equal to V_t, i.e. U_t is not autocorrelated because Vt is not autocorrelated by the assumption of random term. If $\rho > 0$ successive errors are positively correlated and when $\rho < 0$ successive errors are negatively correlated.

6.4.1. Sources of Autocorrelation

Autocorrelated values of the disturbance term may be observed for many reasons. These are:

- **Omitted explanatory variables**

Most economic variables tend to be autocorrealted. If an autocorrelated variable has been excluded from the set of explanatory variables, then its influence will be reflected in the random variable U. This is called “quasi-autocorrelation”, since it is due to the autocorrelated pattern of the omitted explanatory variables and not because of the pattern of the values of the random variable U. If several

autocorrelated explanatory variables are omitted, then the random variable, U, may not be autocorrelated. This is because the autocorrelation patterns of the omitted variables may offset each other.

- **Mis – specification of the mathematical form of the model**

If we use a mathematical form which differs from the correct form of the relationship, then, the random variable may show serial correlation, if we chose a linear function while the correct form is non-linear, then the values of U will be correlated.

- **Mis – specification of the true random term U**

Many random factors like war, drought, weather conditions, strikes etc exert influence that are spread over more than one period of time. For example, the effect of weather conditions in agricultural sector will influence the performance of all other economic variables in several times in the future. A strike in an organization affects the production process which will persist for several future periods. In such cases, the values of U's become serially dependent, so that if we assume E (U_iU_j) = 0, then we mis-specify the true pattern of values of U. This type of autocorrelation is called "*true autocorrelation*".

- **Interpolation in the statistical observation**

Most time series data involve some interpolation and "smoothing process" to remove seasonal effect which does average the true disturbances over successive time periods. As a result, the successive values of U's are interrelated and show autocorrelation patterns. For instance, AC arises when a lagged value of a variable is used as a regressor.

The source of autocorrelation has a strong influence on selecting solution for the correction of autocorrelation. This means, the type of corrective action depends on the cause or source of autocorrelation. This means, the type of corrective action depends on the cause or source of autocorrelation.

6.4.2. Consequences of autocorrelation

When the disturbance term exhibits autocorrelation, the numerical value as well as standard errors of the estimates is affected. However, the estimates of the parameters do not have statistical bias, i.e. even when the residuals are serially correlated, the estimates of ordinary least squares are unbiased. When the disturbance terms are autocorrelated, the ordinary least squares variances of the estimates are likely to be larger than those of other methods and the variance of the random term may be underestimated, i.e. Minimum variance property of estimate is violated. This inflated variance will lead to accept insignificant estimate.

With autocorrelated values of the disturbance term, the prediction based on ordinary least squares estimates will be inefficient, i.e. they will have a large variance as compared with prediction based on estimated obtained from other methods. The variance of a prediction /forecast/ will depend on

the variance of the coefficient estimates and the variance of the random term, U. Since these variances are not minimum as compared with other methods, the standard error of the forecast from ordinary least squares will not have least value due to autocorrelated U's.

With autocorrelatedvalues of the disturbance term, the tests, confidence intervals using estimates and their variances is not efficient.

6.4.3. Tests for autocorrelation

There are two alternative methods to detect for autocorrelation. These are: visual observation of residuals and formula tests. Let's discuss each of them as follows:

A. Visual observation

In this method, we can identify the existence of autocorrelation by plotting the residuals either against their own lagged values or against time. This test will be described as follows:

Suppose $Y_t = f(X_t)$

Step 1: regress Y_t on X_t's using the method of ordinary least squares and obtain the residual estimates $\hat{U}_t = Y_t - \hat{Y}_t$

Step 2: To identify whether there is autocorrelation or not, either plot the residuals against time or their own lagged values and observe for systematic pattern.

One method that can be used for the detection of autocorrelation is to plot the regression residuals $\hat{U}_t$'s against time. If the $\hat{U}_t$'s in successive periods show a regular time pattern, then we conclude that there is autocorrelation. That is, if there is systematic pattern, then there is autocorrelation otherwise there is no autocorrelation.

B. Formal tests

Plotting the residuals either against their owned lagged values or against time is one of the methods to identify the existence of autocorrelation. It provides some rough idea about the existence of autocorrelation. However, there are other more accurate tests for autocorrelation.

These tests are described as follows:

- **Von Neumann ratio and**
- **Durbin-Watson**

In this course, we will discuss the second method of tests for autocorrelation, i.e. Durbin-Watson. This test is the most celebrated and widely used test for autocorrelation problem in a given model. Durbin and Watson have suggested a test which is applicable to small samples. This test is appropriate only for the first order autoregressive schemes. The test is described as follows:

Assume: $Y_t = \alpha + \beta X + U_t$ and the first order auto regressive between U_t is given by:

$$U_t = \rho U_{t-1} + V_t$$

Where; $E(V_t) = 0 \quad and \quad Var(V_t) = \sigma_v^2 \quad Cov(V_t, V_{t-s}) = 0$

Test the following hypothesis:

$H_0: \rho = 0$ (The U's are not autocorrelated with a first order autoregressive scheme) against the alternative hypothesis

$H_0: \rho \neq 0$ (The U's are autocorrelated with a first order autoregressive scheme)

To test the null hypothesis, we use the Durbin-Watson statistic given as below

$$d = \frac{\sum_{t=2}^{n} \left(\hat{U}_t - \hat{U}_{t-1} \right)^2}{\sum_{t=1}^{n} \hat{U}_t^{\,2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (5)$$

$\hat{U}_t$ is the estimate of U_t

The value of **d** lies between **0** and **4**. When d=2, P= 0. Thus, testing H_0: P = 0 is equivalent to testing H_0: d = 2. This can be shown as follows:
Let us expand the above expression to prove the above statement

$$d = \frac{\sum_{t=2}^{n}\left(\hat{U}_t - \hat{U}_{t-1}\right)^2}{\sum_{t=1}^{n}\hat{U}_t^{\,2}} = \frac{\sum_{t=2}^{n}\left(\hat{U}_t^{\,2} + \hat{U}_{t-1}^{\,2} + -2\hat{U}_t\hat{U}_{t-1}\right)}{\sum_{t=1}^{n}\hat{U}_{t-1}^{\,2}}$$

But for large sample, $\sum_{t=1}^{n}\hat{U}_t^{\,2} \cong \sum_{t=2}^{n}\hat{U}_{t-1}^{\,2} \cong \sum_{t=1}^{n}\hat{U}_{t-1}^{\,2}$

And substituting in the above formula, we obtain

$$d = \frac{\sum_{t=1}^{n}\hat{U}_{t-1}^{\,2} + \sum_{t=1}^{n}\hat{U}_{t-1} - 2\sum_{t=1}^{n}\hat{U}_t\hat{U}_{t-1}}{\sum_{t=1}^{n}\hat{U}_{t-1}^{\,2}} = 2\left(\frac{\sum_{t=1}^{n}\hat{U}_{t-1}^{\,2} - \sum_{t=1}^{n}\hat{U}_t\hat{U}_{t-1}}{\sum_{t=1}^{n}\hat{U}_{t-1}^{\,2}}\right)$$

$$d \equiv 2\left(1 - \frac{\Sigma\hat{U}_t\hat{U}_{t-1}}{\Sigma\hat{U}_{t-1}^{\,2}}\right), \text{ but } \hat{\rho} = \frac{\Sigma\hat{U}_t\hat{U}_{t-1}}{\Sigma\hat{U}_{t-1}^{\,2}}$$

$$d \cong 2\left(1 - \hat{\rho}\right)$$

Where $\hat{\rho}$ is the estimate of the population coefficient ρ and it is defined in the range

$$-1 \le \hat{\rho} \le 1$$

From this expression, we observe the following points:

- *When* $\hat{\rho} = -1,\ d = 4$

- *When* $\hat{\rho}=1,\ d=0$
- *When* $\hat{\rho}=0,\ d=2$
- *When* $-1<\hat{\rho}<1,\ 2<d<4$

Therefore, observe that the value of **d** lies between **0** and **4**. If there is no autocorrelation (i.e. $\hat{\rho}$ = 0 or d = 2), then we accept the null hypothesis that there is no autocorrelation in the function.

- *If* $\hat{\rho}=1$ *or* $d=0$, *then we have perfect positive autocorrelation.*
- *If* $0<d<2$ *or* $0<\hat{\rho}<1$, *then there is some* degree *of positive autocorrelation.*

If $\hat{\rho}=-1$ *or* $d=4$, *then we have perfect negative autocorrelation.*

If $2<d<4$ *or* $-1<\hat{\rho}<0$, *then there is some* degree *of negative autocorrelation.*

Therefore, in the Durbin-Watson test of the null hypothesis of zero autocorrelation ($\hat{\rho}=0$ is the same as testing the hypothesis d = 2. To test the hypothesis we follow this procedure:

Step 1: Fit using the method of ordinary least squares Y_t on the explanatory variables to get e_t and e_{t-1}.
Step 2: Compute the empirical value of the Durbin-Watson statistic using sample residuals e_t's.
Step 3: For a given value of significance level, α find the expected value of d or d-tabulated. The d-tabulated depends on the significance level, sample size and the number of explanatory variables excluding the constant term.
Step 4: Compare the computed value with the critical values of d. The problem of this test is that the exact distribution of d is not known. However, Durbin and Watson have established upper (du) and lower limits (dL) for the significance levels of d which are appropriate to test the

hypothesis of zero first order autocorrelation against the alternative hypothesis of positive first order autocorrelation.

Durbin and Watson have formulated these upper and lower values at 5% and 1% level of significance. The table assumes that the U's are normal, homoscedastic and not autocorrelated. In the table n is the sample size and k represents the number of exogenous explanatory variables being estimated. Compare the computed d value from the regression residual with du and dL in the Durbin-Watson table and with their transformations (4-du) and (4-dL).

The comparison using dL and du investigates the possibility of positive autocorrelation and the comparison with (4-dL) and (4-du) investigate the possibility of negative autocorrelation.

The decision rules based on the values are given as follows:

- If the computed d-value is less than dL, we reject the null hypothesis of no autocorrelation and accept there is positive autocorrelation of first order ($\rho > 0$)
- If the computed d-value is greater than (4-dL), we reject the null hypothesis of no autocorrelation and accept there is negative autocorrelation of first order ($\rho < 0$)
- If the computed d-value lies between du and (4-du), we accept the null hypothesis of no autocorrelation ($\rho = 0$)
- If the computed d-value lies either between dL and du or between (4-du) and (4-dL), then the test is inconclusive.

The critical region of the Durbin- Watson is shown below:

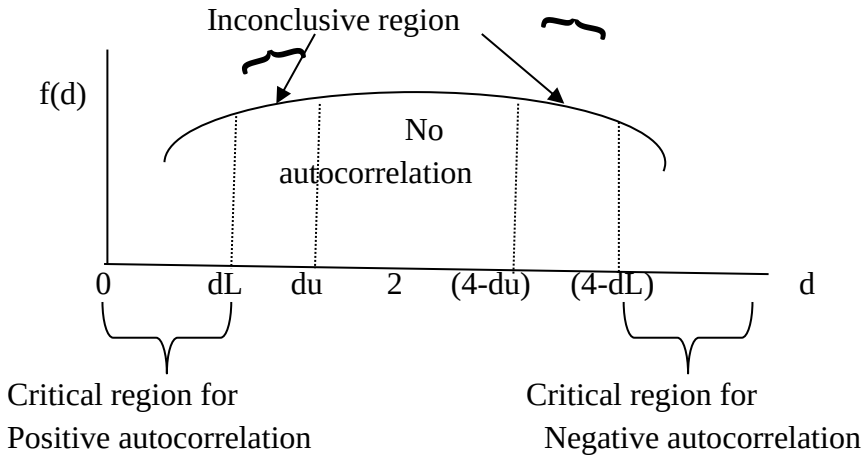


Fig. 6.1. The critical region of the Durbin-Watson

The Durbin-Watson statistic has the following shortcomings.

- d-statistic is not appropriate measure of autocorrelation if among explanatory variables there are lagged values of the endogenous variable. It should not be used if the model includes lagged values of dependent variables.

- The range of values of **d** over which the Durbin-Watson test is inconclusive is also a draw back to its application. If the computed d-value lies either between **dL** and **du** or between (4-du) and (4-dL), then the test is inconclusive. That means, it is impossible to conclude whether there is autocorrelation or not.
- It only detects first order autocorrelation, i.e. it is inappropriate for testing higher order serial correlation.
- The test requires that no heteroscedasticicty and no intercept term.

Example

Consider the data on expenditure, income and price

Expenditure (Y_i)	Income (X_1)	Price (X_2)
3.5	20	16
4.5	26	13
5	30	10
6	42	7
7	50	7
9	54	5
8	65	4
10	72	3
12	85	3.5
14	90	2

N= 10 and K=2 (number of explanatory variables excluding the constant term)

The model is fitted as: $\hat{Y} = -0.77 + 0.151X_1 + 0.089X_2$

To compute the value of ρ we need to first compute the values of U_t and U_{t-1} as follows:

Y	X_1	X_2	$\hat{Y}$	U_t	U_{t-1}	$U_t - U_{t-1}$	U_t^2	$(U_t - U_{t-1})^2$
3.5	20	16	3.66177	-.16177				
4.5	26	13	4.29966	.20034	-.16177	0.36211	0.04013612	0.131123652
5	30	10	4.63624	.36376	.20034	0.16342	0.13232134	0.026706096
6	42	7	6.17808	-.17808	.36376	-0.54184	0.03171249	0.293590586
7	50	7	7.38333	-.38333	-.17808	-0.20525	0.14694189	0.042127563
9	54	5	7.80860	1.19140	-.38333	1.57473	1.41943396	2.479774573
8	65	4	9.37714	-1.37714	1.19140	-2.56854	1.89651458	6.597397732
10	72	3	10.34305	-.34305	-1.37714	1.03409	0.1176833	1.069342128
12	85	3.5	12.34594	-.34594	-.34305	-0.00289	0.11967448	0.0000083521
14	90	2	12.96620	1.03380	-.34594	1.37974	1.06874244	1.903682468
					Sum		**4.9731606**	**12.54375315**

To test the following hypothesis we must compare the calculated value from the table value. That is,

$H_0: \rho = 0$ against the alternative hypothesis

$H_0: \rho \neq 0$ at $\alpha = 0.01$

The computed **d**-value is **2.52**

The tabulated d-value at $\alpha = 0.01$ are dL = 0.466, du = 1.333, 4-du = 4-1.333 = 2.667 and

4-dL= 4-0.466 = 3.534

Decision rule

Given the hypothesis for $\alpha = 0.01$, the decision rule will be given as:

dL	du	d	4-du	4-dL
0.466	1.333	2.52	2.667	3.534

Since du < d < 4-du, accept the null hypothesis, i.e. there is no evidence of autocorrelation. Therefore, $\rho = 0$.

Given the hypothesis for α = 0.05, the decision rule will be given as:

The tabulated d-value at α = 0.05 are dL = 0.697, du = 1.641, 4-du = 4-1.641 = 2.359 and

4-dL= 4-0.697 = 3.303

Decision rule

dL	du	4-du	d	4-dL
0.697	1.641	2.359	2.52	3.303

Since 4-du < d < 4-dL, there is inconclusive evidence

To overcome the shortcomings, Durbin-Watson test is amended in the following form:

- If the computed d-value is less than du or greater than (4-du), then we reject the null Hypothesis $(H_0: \rho = 0)$
- If the computed d-value lies between du and (4-du), then we accept the null hypothesis

In the amended test, the rejection (critical) region includes not only the values of d < dL and d> (4-dL) but also the inconclusive region in the original Durbin-Watson test.

This amendment is shown using graph as follows:

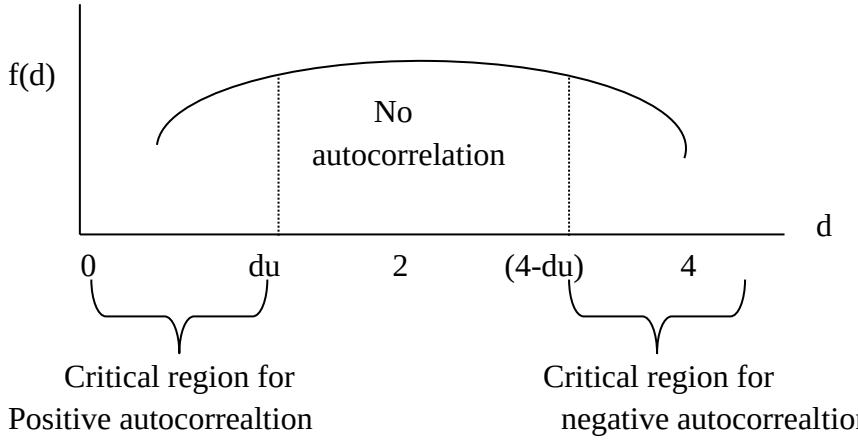


Fig. 6.2. The amended critical region of the Durbin-Watson

The Durbin-Watson test is inappropriate for testing higher order serial correlation or for other forms of autocorrelation like non-linear forms of serial autocorrealtion of the values of U_t.

6.4.4. Solutions for Autocorrelation

The solution to be adopted to remove the effects of autocorrelation depends on the sources of autocorrelation. For example, if the source of autocorrelation is omitted variables, then the appropriate solution is to include these omitted variables in the set of explanatory variables.

The simplest way to detect whether autocorrelation is due to omitted variables or not, is to regress the residuals, U_t's against variables which on a priori grounds might be relevant explanatory variables.

If the source of autocorrelation is mis-specification of the mathematical form, then the solution is to change the initial form. It can be investigated by regressing the residuals against higher powers of the explanatory variables and re-examining the resulting new residuals.

Once autocorreltation is detected by applying any test, the appropriate corrective method is to obtain an estimate of the ρ's and apply the method of ordinary least squares to a set of transformed data.

In order to estimate ρ's one can use different methods. Let's discuss these methods.

A. ρ can be estimated from Durbin-Watson d statistic

Recall that there is the following approximate relationship between the d statistic and ρ:

$$d \cong 2\,(1-\hat{\rho})$$

from which we can obtain

$$\hat{\rho} = 1 - \frac{d}{2}$$

Once the **d** statistic is computed, we can easily obtain an approximate estimate of ρ from the above given relationship.

B. ρ can be estimated from OLS residuals, U_t

Recall that the first-order autoregressive scheme is given as:

$$U_t = \rho U_{t-1} + V_t$$

Since the U's are not observable, we can use their sample counterparts (U's) and run the following regression and obtain the estimates of ρ as:

$$U_t = \hat{\rho} U_{t-1} + V_t \qquad Where, \hat{\rho}\ is\ an\ estimator\ of\ \rho$$

Correcting for error autocorrelation of AR (1) scheme

Consider the model:

$$Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t \qquad t = 1, 2, ...T ..... \quad (*)$$

Where the errors are generated according to the AR (1) scheme:

$\varepsilon_t = \rho \varepsilon_{t-1} + u_t$ Where u_t fulfils all assumptions of the CLRM. Suppose by applying any one of the above tests you come to the conclusion that the errors are autocorrelated. What to do next?

Lagging equation (*) by one period and multiplying throughout by ρ, we get:

$$\rho Y_{t-1} = \rho\beta_1 + \rho\beta_2 X_{t-1} + \rho\varepsilon_{t-1} ..... \quad (**)$$

Subtracting equation (**) from equation (*), we get:

$$Y_t - \rho Y_{t-1} = \beta_1(1-\rho) + \beta_2(X_t - \rho X_{t-1}) + (\varepsilon_t - \rho\varepsilon_{t-1})$$

$$Y^* = \beta_1^* + \beta_2 X_t^* + u_t \quad ..... \quad (***)$$

$$Where\ u_t = \varepsilon_t - \rho\varepsilon_{t-1}.$$

The above transformation is known as the Cochrane-Orcutt transformation. Since u_t fulfils all assumptions of the CLRM, we can apply OLS to equation (***) to get estimates which are BLUE.

Chapter 7: Non-linear Regression and Time Series Econometrics

7.1. Non-linear regression models: overview

Introduction

Nonlinearities can arise in two different ways. In a first case, the model is still linear in the parameters but nonlinear in its explanatory variables. This means that we include nonlinear functions of X_i as additional variables, for example age_i^2 and age could be included in an equation. The resulting model is still linear in the parameters and can still be estimated by OLS. In a second case, the model is nonlinear in its parameters and estimation is less easy.

Previously we have fitted, by OLS model which was of the type:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$$

This model is useful not only when the relationship between the dependent and explanatory variables is a linear one, but also in case where it can be transformed to linearity. For instance, the Cobb-Douglas production function relating an output, Y, to inputs $X_2 \ldots X_k$ is of the form

$$Y = \alpha_1 X_2^{\beta_2}, ... X_k^{\beta_k}$$

If we take logs of both sides of this equation and add an error term, we obtain a regression model:

$$\ln(Y_i) = \beta_1 + \beta_2(\ln X_{2i} + ... + \beta_k(\ln K_{ki}) + \varepsilon_i$$

Where $\beta_1 = \ln(\alpha_1)$. This specification is now linear in logs of the dependent and explanatory variables and, with this small difference, all the techniques of the previous chapters apply. There are, however, some functional forms which cannot be transformed to linearity. In other words, there are many situations in which a model of this form is not appropriate and too simple to represent the true relationship between the dependent (or response) variable Y and the independent (or predictor) variables X_1, X_2 ... and X_n.

Example: $E(Y_i / X_i) = \beta_1 + \beta_2 X_i^2$ is a linear (in the parameters) regression model. To see this, let us suppose X takes the value 3. Therefore, $E(Y / X = 3) = \beta_1 + 9\beta_2$ which is

obviously linear in $\beta_1 \; and \; \beta_2$. Now consider the model $E(Y_i / X_i)=\beta_1+\beta_2{}^2 X_i$. Now suppose X = 3; then we obtain $E(Y_i / X_i)=\beta_1+3\beta_2{}^2$, which is nonlinear in the parameter β_2. The preceding model is an example of nonlinear (in the parameter) regression model.

When we are led to a model of nonlinear form, we would usually prefer to fit such a model whenever possible, rather than to fit an alternative, perhaps less realistic, linear model. Any model which is not of the form given above will be called a nonlinear model.

7.2. Time Series Analysis

I. Introduction

One objective of analyzing economic data is to predict or forecast the future values of economic variables. One approach to do this is to build a more or less structural economic model, describing the relationship between the variable of interest with other economic quantities, to estimate this model using a sample of data, and to use it as the basis for forecasting and inference.

A time series is a sequence of numerical data in which each item is associated with a particular instant in time. Time series data is, as its name suggests, ordered by time. One can quote numerous examples: monthly unemployment, weekly measures of money supply, daily closing prices of stock indices, and so on. In regression analysis involving time series data, if the regression model includes not only current but also the lagged (past) values of the explanatory variables, then it is called a **distributed lag-model**. While if the model includes one or more lagged values of the dependent variable among its explanatory variables, then it is called an **autoregressive model**. Thus,

$$Y_t=\beta_0+\beta_1 X_{t-1}+\beta_2 X_{t-2}+...+\beta_k X_{t-k}+u_t \; ..... \; (7.1)$$

represents the general form of a distributed lag model. Whereas the following is an example of autoregressive model.

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 Y_{t-1} + u_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (7.2)$$

Both distributed lag model and the autoregressive models are dynamic models because the influence of the explanatory variable on the dependent variable is distributed over a number of past values of the explanatory and dependent variables.

In this sub section, we will discuss the importance of lag in economics. Then we provide the techniques to estimate a model with lagged values of the explanatory variables. That means it provides techniques for estimation a distributive lagged model.

II. The role of "time" or "lag"

The number of lags, s, may be either finite or infinite. Assume that the β's have a finite sum, i.e.

$$\sum_{i=0}^{s} \beta_i = \beta$$

Lagged values of the variables are important explanatory variables, because most economic variables are influenced by past patterns of the variable. For example, take the consumption function. It postulates that the current level of consumption depends on past levels of consumption and current and past levels of income, i.e.

$$C_t = f\,(C_{t-1}, Y_t, Y_{t-1}, X_{1t}, X_{2t}\ldots)$$

The investment function postulates that it depends on past outputs, on expectation about future profits, on capital stock and other factors, i.e.

$$I_t = f\,(Q_t, Q_{t-1}, Q_{t-2}, \ldots, \pi_t, K_{t-1}, i_t, \ldots)$$

Where: Q is the level of output
π is profit
K is capital stock
i is interest rate

Very often, dependent variable responds to the explanatory variables with a lapse of time. Such a lapse of time is called **a lag**.

Lags are important for decision making especially, for government officials to know how fast, after how many time periods the economic units will react to changes of various policy variables

(instruments). For example, how fast will producers or consumers react to the imposition of sales tax and other incentives for investment? How fast will investors react to changes in the interest rate? Similarly, the lags involved in the demand function following a change in the policy instruments like price, quantity or advertising of a firm are important for managerial decisions.

Lagged variables are one-way for taking into account the length of time in the adjustment processes of economic behavior and for handling of expectations about future events. However, economic theory never suggests the precise number of lags that should be included in a function, even if it recognizes the importance of time lags. The researcher will choose among different lags patterns the one that gives the most satisfactory fit on the basis of statistical criteria.

III. Estimation of Distributed lag models

Assume that Y depends on the value of X over **s** periods. This is called a **finite (lag) distributed lag model** because the length of lag is specified.

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + ... + \beta_s X_{t-s} + u_t \quad \text{.....} \quad (7.3)$$

However, if the model is given as:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + ... + u_t \quad \text{.....} \quad (7.4)$$

Such a model is called an **infinite (lag) model**.

To estimate the β's in a infinite model, we may adopt two approaches: ad hoc estimation and a prior restrictions on the β's by assuming that β's follows some systematic pattern.

Suppose the model includes only lagged values of the exogenous variable (s) in the set of explanatory variables, make the usual assumptions about the error term U.

$$U \sim N(0, \sigma_U^2)$$

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

$$E(U_i X_j) = 0 \quad \text{for } j = 1, 2, 3 \ldots K$$

Since the explanatory variable X_t is non-stochastic, it is uncorrelated with the disturbance term U_t. Thus, X_t, X_{t-1} and so on are non-stochastic and the ordinary least squares can be applied.

This method suggests that to estimate such a model, one may proceed sequentially., i.e. first regress Y_t on X_t, then regress Y_t on X_t and X_{t-1} and so on. This sequential procedure stops when the

regression coefficients of the lagged variables start becoming statistically insignificant and/or the coefficient of at least one of the variables changes signs from positive to negative or vice versa.

However, the following problems will arise in attempting to apply this approach.

- If the number of lags is large and the sample is small (in the case of time series data), we may be unable to estimate the parameters because there will be no adequate degrees of freedom to carry out the statistical tests of significance.

- There will be a multicollinearity problem, since there is strong correlation between successive values of the same variable. With strong collinearity, the values of the estimates will be imprecise and their standard errors will be large so that we may be led to mis-specification of the model by dropping variables.

To avoid these problems, various methods have been suggested to reduce the number of lagged variables. This is achieved by imposing restrictions on the β's and constructing new variables from a linear combination of the lagged variables. The methods differ in the weights which are used in constructing these new variables.

One of the most popular distributed lag models with endogenous lagged variables is Koyck's Geometric lag scheme. This model assumes that the weights/lag coefficients are declining continuously following the pattern of geometric progression.