ECO-03-8

Institute of Distance and Open Learning

Gauhati University

M.A./M.Sc. in Economics

Semester II

Paper VIII

Elements of Econometrics

Contents:

Course Advisors :

Prof. Kandarpa Das	:	HOD. Dept. of Foreign Language GU
Dr. Ratul Mahanta	:	Assistant Professor Department of Economics, GU

Contributors :

Jayashree Choudhury (Unit : 1 & 5)	:	Assistant Professor, Department of Economics Handique Girls' College, Guwahati
Shrutidhara Kashyap (Unit : 2)	:	Research Seholar Department of Economics, GU
Arati Bharali (Unit : 3 & 4)	:	Assistant Professor, Department of Economics ADP College, Nagaon

Editorial Team :

Prof. Pranab Jyoti Das	:	Director, i/c, IDOL, GU
Dr. Ratul Mahanta	:	Assistant Professor Department of Economics, GU
Dipankar Saikia	:	Editor, SLM, IDOL, GU

Cover Page Designing:

Bhaskarjyoti Goswami	:	IDOL, Gauhati University

MA/M.Sc. Economics
Institute of Distance and Open Learning
GAUHATI UNIVERSITY

COURSE STRUCTURE

A student shall do a total number of sixteen papers in the four Semesters. Each paper will carry 100 marks - 20 marks for internal evaluation during the semester and 80 marks for external evaluation through end semester examination. All the papers in the First, Second and Third Semesters will be compulsory. The paper XIII and XIV of the Fourth Semester will also be compulsory. The remaining two papers for the Fourth Semesters will be chosen by a student from the optional papers. The names and numbers assigned to the papers are as follows.

First Semester

I	Microeconomics Theory
II	Macroeconomics Theory - I
III	Mathematical Methods for Economic Analysis-I
IV	Statistical Methods for Economic Analysis

Second Semester

V	Advanced Microeconomics
VI	Macroeconomic Theory -II
VII	Mathematical Methods for Economic Analysis-II
VIII	Elementary Econometrics

Third Semester

IX	Development Economics-I
X	International Economics
XI	Issues in Indian Economy
XII	Public Finance-I

Fourth Semester

XIII	Development Economics-II
XIV	Public Finance-II

Papers XV and XVI are optional

A student has to choose any two of the following courses.

(a) Population and Human Resource Development

(b) Econometric Methods

(c) Environmental Economics

(d) Financial System

Detailed Contents of this Paper

Paper - VIII
ELEMENTS OF ECONOMETRICS

Unit – 1: Sampling and Estimation

Concept of Sampling Distribution and Standard Error of a Statistic – Methods of Estimation – Principles of Moments, Least Square and Maximum Likelihood (Concept only) – Characteristics of a Good Estimator.

Unit –2: Statistical Inference

Testing of Hypothesis: Type I and Type II Errors, One-tailed and Two-tailed Tests – Test based on Standard Normal, t and Chi-Square Distributions.

Unit – 3: Linear Regression Model and Its Estimation

The Two Variable Model and its OLS Estimation – The General Linear Regression Model – Standard Assumptions – OLS Estimators and their Properties – The Coefficient of Determination – Maximum Likelihood Methods, Estimation and Properties.

Unit – 4: Inference from Linear Regression Estimation

Test of Hypothesis about Regression Coefficients and their Confidence Interval – Prediction with the Linear Regression Model.

Unit – 5: Further Topics in Linear Regression Model

Multicollinearity: Effects, Detection and Remedies – Specification Errors and their Consequences – Qualitative Factors and Dummy Variables – Introductions to Heteroscedasticity and Autocorrelation of Disturbances (Ideas only).

* * *

UNIT-1

SAMPLING AND ESTIMATION

STRUCTURE :

1.0 Introduction :

Econometrics means "economic measurement". 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 phenomenon. In this unit, we shall understand the concept of Sampling Distribution and standard error of a statistic. Again, a brief introduction of the Methods of Estimation i.e. principle of moments, Least Square and Maximum Likelihood will be studied Lastly the characteristics of a good estimator shall be analysed.

1.1 Objectives :

After reading this unit, you will be able to—

- Understand the concept of Sampling Distribution and Standard Error of a Statistic.
- Learn the different methods of estimation such as principle of

moments, Least Square and maximum likelihood.

- Analyse the various characteristics of a good estimator.

1.2 Concept of Sampling Distribution and Standard Error of a Statistic

Before understanding sampling distribution, we should understand the meaning of the following terms—

(a) Population or Universe : In any statistical investigation, the group of items or individuals under study is known as population or universe.

(b) Sample : A finite subset of the population, selected from it with the objective of investigating its properties is called a sample.

(c) Sample size : The number of units in the sample is known as sample size.

(d) Sampling : Sampling is a tool which enables us to draw conclusions about the characteristics of the population after studying only those objects or items that are included in the sample.

(e) Parameter : The statistical consists of the population like mean (μ), variance $\left(\sigma^2\right)$, Skewness (β_1), Kurtosis (β_2) etc are known as parameters. Parameters are functions of population values.

(f) Statistic : The statistical consants of the sample like mean $(\overline{x})$ variance $\left(S^2\right)$, skewness (b_1), kurtosis (b_2) etc are known as statistic. They are functions of the sample observations.

Sampling Distribution :

If we draw a sample of size 'n' from a given finite population of size 'N', then the total number of possible samples is:

$$N_{C_n} = \frac{N!}{n!(N-n)!} = k \text{ (say)}$$

We can compute some statistic, say, 't' for each of these **k** samples like mean $(\overline{x})$, variance $\left(S^2\right)$ etc.

The set of values of the statistic so obtained, one for each sample, is called the sampling distribution of the statistic. For example, statistic 't' may be regarded as a random variable which can take the values $t_1, t_2, \ldots, t_k$ and we can compute various statistical constants like mean, variance etc for its distribution.

For eg, Mean $= \bar{t} = \frac{1}{k}(t_1 + t_2 + ... + t_n) = \frac{1}{k}\sum_{i=1}^{n} t_i$

Variance $=$ Var (t) $= \frac{1}{k}\sum[t - E(t)]^2 = \frac{1}{k}\sum_{i=1}^{n}(t_i - \bar{t})^2$

Standard Error :

The standard deviation of the sampling distribution of a statistic is known as its standard error (S.E.).

Thus, the standard error of the statistic t is given by

$$S.E\,(t) = \sqrt{Var\,(t)} = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(t_i - \bar{t})^2}$$

Again, the reliability or efficiency of a sampling plan is determined by the reciprocal of the standard error of the estimate and is called the precision of the estimate. Thus, if t is a statistic, then

$$\text{Precission of } t = \frac{1}{S.E.(t)}$$

1.3. Methods of Estimation :

Estimation of population parameters iike mean, variance etc. is one of the important problems of statistical inference.

Some of the commonly used methods of estimation are enumerated bellow—

1.3.1 Principle of Moments :

This consists in equating the moments of the population to the sample moments and then

solving the equations so obtained to get the required estimates of the population parameters. For example, if we want to estimate the parameter 'p' of the binomial distribution, when 'n' is known, then we equate the mean of binomial distribution which is np, to the sample mean $(\bar{x})$, which gives

$$np = \bar{x} \Rightarrow p = \frac{\bar{x}}{n}$$

If both n and p are unknown, then we take the first two moments. Thus solving

$$\text{Mean} = np = \bar{x} \text{ and}$$
$$\text{Var} = npq = np(1-p) = s^2$$

for p and n, we get the corresponding estimates of the population parameters.

This technique is used if we have to estimate the theoretical frequencies of given distribution by fitting an appropriate probability distribution to it.

1.3.2 Principle of least squares :

Principle of least squares is the most extensively used methods of estimation.

Let us consider a two-variable function as

$$Y_i = \beta_1 + \beta_2 X_i + \mu_i \quad \text{..... (i)}$$

But (i) is not directly observable. So we estimate it from the sample regression function:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \mu_i$$
$$= \hat{Y}_i + \mu_i \quad \text{....... (ii)}$$

where $\hat{Y}_i$ is the estimated value of Y_i

Again (ii) $\Rightarrow \hat{\mu}_i = Y_i - \hat{Y}_i$

$$= Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$$

which shows that $\hat{\mu}_i$ (the residuals) are simply the differences between the actual and estimated Y values. Now principle of least squares consists in minimizing the sum of squares of the residuals ie deviations between the given observed values of the variable and their corresponding estimated values.

Thus we get

$$\frac{\partial \sum \mu_i{}^2}{\partial \beta_1} = 0 \text{ and } \frac{\partial \sum \mu_i{}^2}{\partial \beta_2} = 0$$

Solving the above, we get the following equations

$$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i$$

$$\sum Y_i X_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2 .$$

These are known as the normal equations. Solving them we get the estimates of $\hat{\beta}_1$ and $\hat{\beta}_2$.

1.3.3 Method of Maximum Likelihood (concept only) :

This is the most commonly used method for estimating the population parameters. It consists in maximizing the likelihood of probability of randomly obtaining a set of sample values.

Mathematically, let $x_1, x_2, ..., x_n$ be a random sample of size n from a population with probability function or p.d f. $p(x, \theta)$, where θ is the unknown parameter. Then, an estimate of θ is obtained on maximizing the likelihood function.

$$L = p(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} p(x_i, \theta)$$

Using the principle of maxima and minima, maxima likelihood estimator is found out by solving the following equations:

$$\frac{\partial L}{\partial \theta} = 0 \quad \text{and} \quad \frac{\partial^2 L}{\partial \theta^2} < 0$$

Since log L is a non-decreasing function of L, L and log L attain their extreme values at the same values of θ. The equation can be written as

$$\frac{1}{L} \cdot \frac{\partial L}{\partial \theta} = 0 \quad \Rightarrow \frac{\partial}{\partial \theta}(\log L) = 0$$

which is known as the likelihood equation for estimating θ.

1.4 Characteristics of a good estimator :

A particular value of a statistic which is used to estimate a given parameter is known as a point estimate or estimator of the parameter. A good estimate is one whose value is close to the true value of the parameter. Following are the characteristics of a good estimator:

(1) Unbiasedness (2) Consistency

(3) Efficiency (4) Sufficiency

(1) Unbiasedness :

A statistic $t = t(x_1, x_2, \ldots, x_n)$, a function of sample observations $x_1, x_2, \ldots, x_n$ is said to be an unbiased estimate of the corresponding population parameter θ, if $E(t) = \theta$.

i.e. if the mean value of the sampling distribution of the statistic is equal to the parameter. For example, the sample mean $(\bar{x})$ is an unbiased estimate of the population mean μ ; the sample proportion P is an unbiased estimate of the population proportion P, ie

$$E(\bar{x}) = \mu, \qquad E(P) = P.$$

If $E(t) \neq \theta$, then the statistic t is said to be a biased estimate of θ. Let $E(t) = b + \theta$ then 'b' is called the 'amount of bias' in the estimate. If $b>0$, i.e $E(t) > 0$, then t is said to be positively biased and if $b<0$ i e $E(t)<0$, it is said to be negatively biased.

(2) Consistency :

A statistic $t = t_n = (x_1, x_2, \ldots, x_n)$ based on a sample of size n is said to be a consistent estimator of the parameter θ if it converges in probability to θ, i.e, if $t_n \to \theta$ as $n \to \infty$.

Symbolically, $\lim\limits_{n\to\infty} P(t_n \to \theta) = 1$

For example, sample mean $\bar{x}$ is a consistent estimator of the population mean, sample variance s^2 is a consistent estimator of the population variance σ^2

NOTE : A statistic $t = t_n = t(x_1, x_2, \ldots, x_n)$ is a consistent estimator of the parameter θ if

$$\left.\begin{array}{r} E(t_n) \to \theta \\ \text{and} \quad \mathrm{Var}(t_n) \to 0 \end{array}\right\} \text{as } n \to \infty.$$

(3) Efficiency :

For more than one consistent estimators of a parameter θ, the efficiency criterion helps us to choose between them by considering the variance of the sampling distributions of the estimators. If t_1 and t_2 are consistent estimators of a parameter θ such that

Var (t_1) < Var (t_2) for all n, then t_1 is said to be more efficiency than t_2. Hence, an estimator with lesser variability is said to be more efficient and consequently more reliable than the other.

If t is the most efficient estimator of a parameter θ with variance V and t_1 is any other estimator with variance V_1, then the efficiency E of t_1 is defined as :

$$E \doteq \frac{V}{V_1}.$$

Also efficiency of any estimator cannot exceed unity.

(4) Sufficiency :

A statistic $t = t(x_1, x_2, \ldots., x_n)$ is said to be a sufficient estimator of parameter θ if it contains all the information in the sample regarding the parameter. In other words, a sufficient statistic utilities all the information that a given sample can furnish about the parameter.

Thus, these are the characteristics of a good estimater.

1.5 Summary :

The set of values of a statistic, is called a sampling distribution of the statistic. Again, the standard error is the standard deviation of the sampling distribution of a statistic. The various methods of estimation consists of the principle of moments, least square and the maximum likelihood method. Also, there are four basic characteristics of a good estimator i.e. Unbiasedness, consistency, efficiency and sufficiency.

1.6. Additional Readings :

1. Johnston, J., "Econometric Methods", Mc Graw Hill.

2. Gujaethi, D., "Basic Econometrics", Mc Graw Hill.

3. Salvatore, Dominick and Reagle, Darvick, "Statistics and Econometrics". Tata McGraw Hill.

1.7. Self Assessment Test

1. What do you mean by population or universe, sample, sample size, sampling, parameter and statistic.

2. Explain the concept of sampling distribution and standard error of a statistic.

3. What are the different methods of estimation. Briefly explain the principle of moments.

4. Briefly describe the concepts of least square and the method of maximum likelihood.

5. State and explain the basic characteristics of a good estimator.

●●●

UNIT -2

STATISTICAL INFERENCE

Contents :

2.1 Introduction :

The inductive inference method is the logic of drawing statistically valid conclusions about the population characteristics on the basis of a sample drawn from it in a scientific manner. But the generalisations of results on the basis of samples involve an element of risk. The risk of taking wrong decisions. So, modern theory of probability plays an important role in decision making. The branch of statistics which helps in arriving at the criterion for such decisions in known as testing of hypothesis. The theory of testing of hypothesis was initiated by J. Neyman and P.S. pearson and employs statistical tecniques to arrive at decisions where there is an element of uncertainty on the basis of a sample whose size is fixed in advance. In this chapter, various fundamental concepts of testing of hypothesis and defferent tests are discussed.

2.2 Objectives :

This unit is designed to help you understand the concept of satistical inference and its related ideas. After reading this unit you will be able to.

- Construct and test hypothesis.
- Know whether a statistical test is significant or not.
- Distinguish the large sample tests from small sample tests.
- Know or choose a particular test for testing hypothesis.
- Know about critical region and confidence interval.

2.3 Testing of Hypothesis :

The inductive inference is based on decision about the characterstics of the population on the basis of sample study. Such decision involve an attempt of risk, the risk of taking wrong decisions. The modern theory of probability plays a vita' role in decision making and the branch of statistics which helps us in ariving at the criterion for decision is known as testing of hypothesis.

A proceedure to assess the significance of a statistic or difference between two independent statistics is known as the test of significance. To test the significance we will use two types of hypothesis.

Null Hypothesis :

For any test of significance first step is to set up a hypothesis— a definite statement about the population parameter. Such a statistical hypothesis which is under test is usually a hypothasis of no difference and hence is called null hypothesis. It is userally denoted by Ho. According to R.A. Fisher," Null hypothesis is that hypotheses which is tested for possible rejection under the assumption that it is true."

Usually, the null hypothesis is expressed as an equality eg. H_o: q ¹ q_o

Alternative Hypothesis :

Any hypothesis which is complementary to the null hypothesis is called an alternative hypothesis. It is usually denoted by H_1. It is very importent to explicitly state the alternative hypothesis in respect of any null hypothesis H_o, because the acceptance or rejection of H_o is meaninful only if it is being tested against a rival hypothesis.

The alternative hypothesis against the null hypothesis H_o : q = q_o will be, H_1 : q ¹ q_o.

2.4 Type I and Type II Errors :

It is mentioned that the inductive inference consists in ariving at a decision to accept or reject a null hypothesis (H_o) after inspecting only a sample from it. As such an element of risk— the risk of taking wrong decisions is involved. In any test procedure the four possible mutually disjoint and inclusive decisions are—

(i) Reject H_o when actually it is not true, ie when H_o is false.

(ii) Aecept H_o when it is true.

(iii) Reject H_o when is true.

(iv) Accept H_o when it is false.

The decisions in (i) and (ii) are correct decisions while the decisions in (iii) and (iv) are wrong decisions.

True Slate		Decision from Sample	
		Reject Ho	Accept H_o
	Ho True	Wrong (Type I Error)	Correct
	Ho False (H_1 True)	Correct	Wrong (Type II Error)

Type I Error : The error of rejecting H_o when H_o is true.

Type II Error : The error of accepting H_o when Ho is false (ie H_1 is true)

We make type I error by rejecting a true new hypothesis and type II error by accepting a wrong new hypothesis. Symbolically,

P[Reject H_o when it is true] = P [Rejecting H_o/H_o]

=P [Type I error] = α

and P[Accept Ho when it is wrong] = P [accepting H_o/H_1]

= P [Type II error] = β

α and β are sizes of I and Type II error respectively.

In the terminology of industrial quality control while inspecting the quality of a manufactured lot, the type I error amounts of rejecting a good lot and type II error amounts to accepting a bad lot.

Accordingly,

α = P [Rejecting a good lot]

β = P[Accepting a bad lot]

The size of type I and type II errors are known as producer's risk and consumer's risk respectively.

An ideal test procedure would be one which is to planned as to safeguard against both these errors. But practically, in any given problem, it is not possible to minimize both these errors simultaneously. An attempt to decrease α results in an increase in β and vice:– Versa. Consequences of type II error are likely to be more serious than the consequences of type I error. Since the errors cannot be reduced simultaneously, a compromise is made by minimizing more serious errors after fixing up the less serious error. Thus, we fix α, the size of type I error and then try to obtain a criterion which minimizes β, the size of type II error. We have,

β = P[Type II error]

= P [Accepting H_o when is false or is true]

Now,

P[Accept H_o when is wrong] + P[Accept H_O when it is true] =1

$\Rightarrow$ P[Accept H_o when H_o is true = 1-P[Accept H_O when H_O is wrong]

$=1-\beta$

Obviously, when H_0 is true it is ought to be accepted. Hence, minimizing β. Amounts to maximizing $(1-\beta)$, which is called the 'power of the test'. Hence the usual practice in testing of hypothesis is to fix α, the size of type I error and then try to obtain a criterion which minimizes β, the size of type II error or maximizes $(1-\beta)$, the power of the test.

Level of Significance :

The maximum size of type I error which we are prepared to risk is known as the level of significance. It is denoted by,

P[Rejecting H_o when H_o is true] = α

Commonly used levels of significance in practice are 5% (0.05) and 1% (0.01). If we adopt 5% level of significance it implies that in 5 samples out of 100, we are likely to reject a correct H_o. In other words, this implies that we are 95% confident that our decision to reject H_0 is correct. Level of significance is always fixed in advance before collecting the sample information.

Whan we reject a null hypothesis H_0 we have certain confidence in our decision which depends on the level of significance employed. Thus, at α Level of significance, the degree of confidence in our decision is $(1-\alpha)$, which is called the 'confidence coefficient'. However when we accept H_0 we do not have any confidence in our decision.

Critical Region :

The statistics which lead to the rejection of the Ho gives us a region called Critical Region (C) or Rejection Region (R) while those which lead to the acceptance of H_0 gives us a region called Acceptance Region (A).

2.5. One Tailed and Two Tailed Tests :

In any test, the critical region is represented by a portion of the area under the probability curve of the sampling distribution of the test statistic.

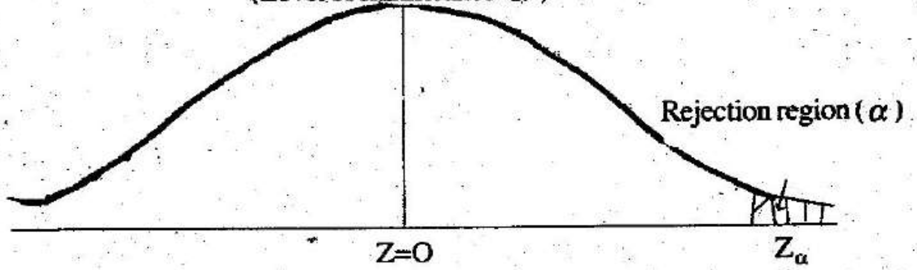
A test of any to statistical hypothesis where the alternative hypothesis is one tailed (right tailed or left tailed) is called a one tailed test. For example a test for testing the mean of a population

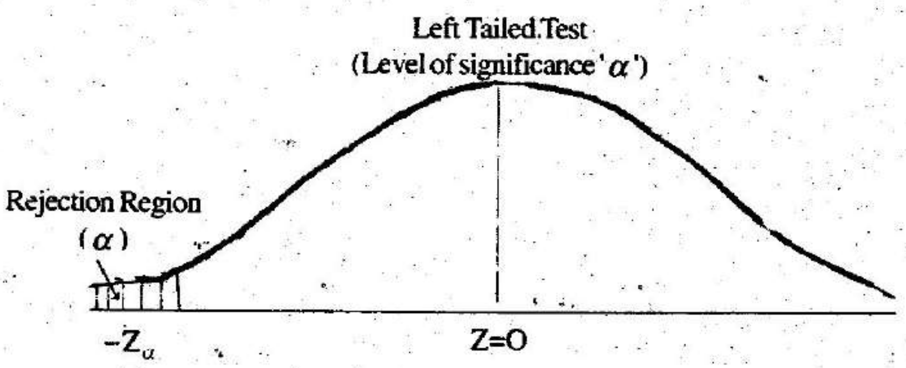
$$H_0 : \mu = \mu_0$$

Against the alternative hypothesis,

$H_1 : \mu > \mu_0$ (Right tailed) or $H_1 : \mu < \mu_0$ (left tailed) is a single tailed test. In the right tailed test ($H_1 : \mu > \mu_0$) the critical region lies entirely in the right tail of the sampling distribution of $\overline{x}$, , while for the left tailed test ($H_1 : \mu < \mu_0$) the critical region is entirely in the left tail of the distribution of $\overline{x}$.

Right tailed test
(Level of significance 'α')





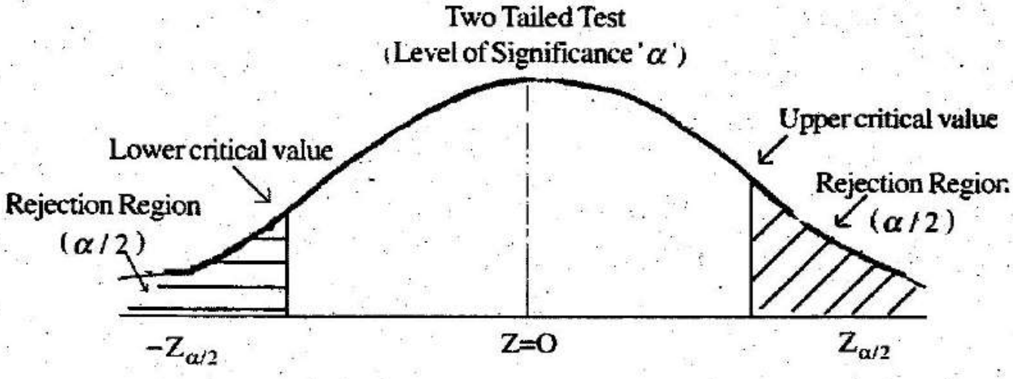
A test of statistical hypothesis where the alternative hypothesis is two tailed such as:

$$H_0: \mu = \mu_0$$

against the alternative hypothesis

$$H_1: \mu \neq \mu_0 \quad (\mu > \mu_0 \text{ and } \mu < \mu_0)$$

is known as two tailed test and in such a case the critical region is given by the portion of the area lying in both the tails of the probability curve of the test statistic.



In a particular problem, whether one tailed or two tailed test is to be applied depends entirely on the nature of the alternative hypothesis. If the alternative hypothesis is two tailed, we apply two tailed test and if the alternative hypothesis is one tailed, we apply one tailed test.

Let us understand,

In the language of significance tests, a statistic is said to be statistically significant if the value of the test statistic lies in the critical region. In this case the null hypothesis is rejected. A test is said to be statistically insignificant if the value of the test statistic lies in the acceptance region.

Critical Values of Z

Critical values ($(Z\alpha)$)	level of significance		
	1%	5%	10%
Two-tailed test	$\lvert Z\alpha\rvert = 2.58$	$\lvert Z\alpha\rvert = 1.96$	$\lvert Z\alpha\rvert = 1.645$
Right-tailed test	$Z\alpha = 2.33$	$Z\alpha = 1.645$	$Z\alpha = 1.28$
Left-failed test	$Z\alpha = -2.33$	$Z\alpha = -1.645$	$Z\alpha = -1.28$

2.6 Steps for testing hypothesis :

The steps which are followed for testing hypothesis are the following—

Step 1.

Set up null hypothesis H_o.

Step 2.

Set up alternative hypothesis H_1. It will determine whether the test is single tailed or two-tailed.

Step 3

Choose the applecapleciate level of significance.

Step 4

Compute the test statistic

$$Z = \frac{t - E(t)}{S.E.(t)}$$

Under the null hypothesis H_o,

The commonly used tests are based on standard normal, t, F and chi-Square distributions.

Step 5

Compare the computed value of Z in step 4 with the tabulated value of Z at the given level of significance α.

If the computed value of Z is less than Z_α, it is not significant and the null hypothesis is accepted. On the other hand, if the computed value of Z is greater than the critical value of Z, it is significant and the null hypothesis rejected at level of significance α.

2.7.Test based on Standard normal distribution :

The sampling distribution of the statistic $t = t(x_1, x_2, \ldots\ldots x_n)$, a function of the sample observations is asymptotically normal, i.e. the standard variate corresponding to the statistic t,

$$Z = \frac{t - E(t)}{S.E.(t)}$$

is asymptotically normally distributed N(0,1) as $n \to \alpha$.

In case of large samples that is when the sample size is greater than 30 (n>30), the normal test is applied which is based on the following fundamental property of the normal distribution.

If $X \sim N(\mu, \sigma^2)$, then $Z = \dfrac{X - E(X)}{\sigma_x} = \dfrac{X - \mu}{\sigma} \sim N(0,1)$

Sampling of variables :

In case of sampling of variables, the quantitative measure ments are taken on the sampling units like height, weight, age, diameter, income, expenditure etc. Each member of the population provides a value of the variable and the aggregate of these values constitutes the frequency distribution of the population with, say mean = μ, standard deviation = σ and so on.

Test of significance for a single mean :

If $x_1, x_2, \ldots\ldots x_n$ are the observations on the n sample units drawn at random from a normal population with mean μ and variance σ^2 ,, then the sample mean $X \sim N(\mu, \sigma^2/n)$.

$X \sim N(\mu, \sigma^2/n)$, asymptotically as $n \to \alpha$

Thus, $E(\bar{x}) = \mu$ and Var $(\bar{x}) = \sigma^2/n$

$$\therefore S.E(\bar{x}) = \sigma/\sqrt{n}$$

The standard normal variate corresponding to $(\bar{x})$ becomes

$$Z = \frac{\bar{x} - E(\bar{x})}{S.E.(\bar{x})} = \frac{\bar{x} - \mu}{\sigma|\sqrt{n}} \sim N(0,1)$$ for iarge samples.

In developing the test of significance for a single mean, we are interested—

(i) To test if the mean of the population has a specified value μ_0 (say) i.e. $\mu = \mu_0$

(ii) To test if the sample mean differs significantly from the hypothetical value of population mean i.e. to test the significance of the difference between $\bar{x}$ and μ.

(iii) To test if the given random sample has been drawn from a population with specified mean μ_0 and variance σ^2.

It is important to note here that if the population standard deviation σ is unknown, then we use its estimate provided by the sample variance given by

$\hat{\sigma}^2 = s^2 \Rightarrow \hat{\sigma} = s$ (for large samples)

confidence limits for μ :

95% and 99% confidence limits for the population mean μ are given by :

95% confidence limits:

$$\bar{x} \pm 1.96 S.E.$$

$$= \bar{x} \pm 1.96\ \sigma|\sqrt{n} = \bar{x} \pm 1.96\ s|\sqrt{n}$$

99% confidence limits :

$$\bar{x} \pm 2.58\ \sigma|\sqrt{n} = \bar{x} \pm 2.58\ s|\sqrt{n}$$

It is in case of random sampling drawn from a large (infinite) population.

In sampling from a finite population with size N, the corresponding limits are the following—

$$\bar{x} \pm 1.96\ \sigma|\sqrt{n}\ \sqrt{\frac{N-n}{N-1}} \text{ and } \bar{x} \pm 2.58\ \sigma|\sqrt{n}\ \sqrt{\frac{N-n}{N-1}}$$

Example :

A random sample of 100 students gave a mean weight of 58 Kilograms with standard deviation of 4 kg. Test the hypothesis that the mean weight in the population is 60 kg.

Solution :

Given. $n = 100$, $\bar{x} = 58$kgs. $s = 4$kgs

Null hypothesis:

$H_0 : \mu = 60$kgs i.e., the mean weight in the population is 60kgs.

Alternative hypothesis:

$H_1: \mu \neq 60\text{kgs}$ (two-tailed)

Under H_o,

the test statistic. $Z = \dfrac{\bar{x}-\mu}{\sigma|\sqrt{n}}$

$$= \frac{\bar{x}-\mu}{\sigma|\sqrt{n}} \quad \left[\because \hat{\sigma}^2 = s^2 = \hat{\sigma} = s \text{ for large samples}\right]$$

$$= \frac{58-60}{s|\sqrt{100}} = \frac{-2}{4/10} = \frac{-20}{4} = -5$$

Since the calculated value of $|Z| = |-5| = 5$ is greater than the critical value of Z i.e. 2.58 1% level of significance, 80, null hypothesis H_o is rejected at 1% level of significance.

Again, since $|Z| = 5$ is queater than the critical value of Z at 5% level of significance i.e 1.96, so it is significant and null hypothesis is rejected.

Hence, we conclude that the mean weight of the population is not 60kgs.

Example :

An educator claims that the average I.Q. of college students is at most 110, and that in a study made to test this claim 150 college students, selected at random, had an average I.Q. of 112.2 with a standard deviation of 7.2. Use the level of significance 0.01 to test the claim of the educator.

Solution :

Given $n = 150$, $\bar{x} = 111.2$, $s = 7.2$

Null hypothesis:

$H_o: \mu = 110$ i.e. the average I.Q. of college students is 110.

Alternative hypothesis:

$H_1: \mu > 110$

Under H_o,

test statistic, $Z = \dfrac{\bar{x}-\mu}{\sigma|\sqrt{n}}$

$$= \frac{\bar{x}-\mu}{s|\sqrt{n}} = \frac{111.2-110}{7.2|\sqrt{150}} = \frac{1.2}{\frac{7.2}{12.25}} = \frac{1.2 \times 12.25}{7.2} = 2.04$$

Since the calculated value of $Z = 2.04$ is less then the critical value of

Z at 1% level of significance for right tailed test i.e. 2.33, it is not significant and null hypothesis H_o is accepted. So, the educator's claim is valid.

Test of significance for difference of means :

Let us suppose that there are two independent random samples of sizes n_1 and n_2 from the two populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively. Let $\bar{x}_1$ and $\bar{x}_2$ be the corresponding sample means.

Then $\bar{x}_1 \sim N\left(\mu_1, \sigma_1^2 / n_1\right)$ and $\bar{x}_2 \sim N\left(\mu_2, \sigma_2^2 / n_2\right)$ asymptotically,

i.e. $n_1 \to \alpha$ and $n_2 \to \alpha$ for large samples.

So, $(\bar{x}_1 - \bar{x}_2)$ being the difference of two independent normal variables is also a normal variables with mean $(\mu_1 - \mu_2)$ and variance $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

So, the standardised variable Z corresponding to the statistic $\bar{x}_1 - \bar{x}_2$ is given by

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1 - \bar{x}_2)}{S.E.(\bar{x}_1 - \bar{x}_2)} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

under the null hypothesis, $H_0 : \mu_1 = \mu_2$ i.e. the population means are equal, the test statistic becomes

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

If σ_1^2 and σ_2^2 are unknown, then their estimates provided by the corresponding sample variances s_1^2 and s_2^2 respectively are used, i.e..

$\hat{\sigma}_1^2 = s_1^2$ and $\hat{\sigma}_2^2 = s_2^2$ (Since samples are large)

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1)$$

If the two independent samples are drawn from the same population, then the test statistic will be

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

If the common variance σ^2 is not known, then we use its estimate (for large samples) based on both the samples and given by:

$$\sigma^2 = \frac{n_1 s_1^{\,2} + n_2 s_2^{\,2}}{n_1 + n_2}$$

Example:

Given the following information relating to two places, A and B, test whether thrue is any significant difference between their mean wages:

	A	B
Mean Wages (Rs.)	47	49
Standard deviations (Rs.)	28	40
No. of workers	1000	1500

Solution :

Let X and Y be the wages (in Rs.) in two places A and B respectively. Given,

$n_1 = 1{,}000$	$\bar{x} = 47$	$s_x = 28$
$n_2 = 1500$	$\bar{y} = 49$	$s_y = 40$

Null hypothesis: $H_o: \mu_x = \mu_y$ i.e. there is no significant difference between the mean wages in places A and B.

Alternative hypothesis : $H_1: \mu_x \neq \mu_y$

(two-tailed)

Under H_o, the test statistic is

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^{\,2}}{n_1} + \frac{s_y^{\,2}}{n_2}}} = \frac{47 - 49}{\sqrt{\frac{(28)^2}{1.000} + \frac{(40)^2}{1500}}} = \frac{-2}{\sqrt{\frac{784}{1000} + \frac{1600}{1500}}}$$

$$= \frac{-2}{\sqrt{1.784 + 1.067}} = \frac{-2}{\sqrt{1.851}} = \frac{-2}{1.361} = -1.469 = -1.47$$

Since $|Z| < 2.58$, it is not significant at 1% level of significance. Hence we may accept H_0 and conclude that there is no significant difference in the mean wages at places A and B.

Again, $|Z| < 1.96$, it is not significant at 5% level of significance. So, the null hypothesis H_0 is accepted and there is no significant difference in the mean wages at places A and B.

Example :

If 60 M.A. Economics students are found to have a mean height of 63.60 inches and 50 M.Com. students a mean height of 69.51 inches would you conclude that the commerce students are taller than Economics students? Assume the S.D. of height of post-graduate students to be 2.40 inches.

Solution :

Let X denote the height (in inches) of M.A. Economics students and Y the height (in inches) of M.Com. students.

Given,

$$n_1 = 60 \qquad \bar{x} = 63.60$$
$$n_2 = 50 \qquad \bar{y} = 69.51$$

It is also given that the standard deviation of height of post-graduate students is 2.48 inches, i.e.

$$\sigma_x = \sigma_y = \sigma = 2.48$$

Null hypothesis $H_0 : \mu_x = \mu_y$ i.e. the mean heights of M.A. Economics and M.Com. students are equal.

Alternative hypothesis $H_1 : \mu_x < \mu_y$ (left-tailed)

Under H_0, the test statistic is

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}}} \sim N(0.1)$$

$$= \frac{\bar{x} - \bar{y}}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{63.60 - 69.51}{\sqrt{(2.48)^2\left(\frac{1}{60} + \frac{1}{50}\right)}} = \frac{-5.91}{2.48\sqrt{\left(\frac{50+60}{60 \times 50}\right)}}$$

$$= \frac{-5.91}{2.48 \times \sqrt{\left(\frac{110}{3000}\right)}} = \frac{-5.91}{2.48 \times 0.191} = \frac{-5.91}{0.474} = -\ 12.47$$

Since Z = – 12.47<–2.33 i.e.12.47>2.33, it is highly significant at 1% level of significance. So, null hypothesis is rejected and we can conclude that the comerce students are taller than economics students.

Sampling of attributes :

In this case the given population is divided into two mutually disjoint and exhaustive classes, one possessing a particular attribute under study and the other not possessing the attribute.

Test for single proportion :

Consider a population consisting of N units and let, the number of exists possessing the attribute under study be α. Hence, the number of units which do not possess the given attribute is $(N-\alpha)$.

P = Proportion of units in the population possessing the given attribute $= \frac{\alpha}{N}$

Q = Proportion of units in the population which do not possess the given attribute $= \frac{N-\alpha}{N} = 1 - \frac{\alpha}{N} \Rightarrow Q = 1 - P$

In sampling theory, the possession of an attribute by a sampling unit is termed as a success and P represents the probability of success in the population. Again, when a sampling unit does not possess the attribute, it is called failure and represented by Q.

If X is the number of units possessing the given attribute in a sample of size n drown from an infinite (large) population, then

p = Proportion of sampled units possessing the given attribute.

$$\Rightarrow p = \frac{X}{n}$$

q = Proportion of sampled units which do not possess the given attribute $\Rightarrow q = 1 - p$

It is important to note here that E(p) = P, i.e., sample proportion p is an unbiased estimate of the population proportion P.

and $S.E.(p)=\sqrt{\frac{PQ}{N}}$

Therefore, for large samples, the standard normal variate corresponding to statistic 'p' is

$$Z=\frac{p-E(p)}{S.E.(p)}=\frac{p-P}{\sqrt{\frac{PQ}{n}}}\sim N(0,1)$$

Note : If the sample is drawn from a finite population of size N, then

$$S.E.(p)=\sqrt{\left(\frac{N-n}{N-1}\right)\frac{PQ}{n}}$$

Example :

In a big city 325 men out of 600 men were found to be smokers. Does this information support the conclusion that the majority of men in the city are smokers?

Solution :

Given, n = 600

No. of smokers = 325

P = sample proportion of smokers $=\frac{325}{600}=0.5417$

Null hypothesis H_0 : The number of smokers and non-smokers are equal so that

P = population proportion of smokers in the city.

$$=\frac{1}{2}=0.5$$

$\therefore Q=1-P=1-0.5=0.5$

Alternative hypothesis $H_1:P>0.5$ (Right-tailed)

Under H_0, the test statistic is

$$Z=\frac{p-E(p)}{S.E.(p)}=\frac{p-P}{\sqrt{\frac{PQ}{n}}}\sim N(0,1)$$ (Since the sample is large)

$$=\frac{0.5417-0.5}{\sqrt{\frac{0.5\times0.5}{600}}}=\frac{0.0417}{0.0204}=2.04$$

Since the calculated value of $Z = 2.04$ is greater than the critical value of Z for right-tailed test at 5% level of significance i.e. 1.645, it is significant and the null hypothesis is rejected. So, we conclude that majority of men in the city are smokers.

Again, the calculated value of $Z = 2.04$ is less than the critical value of Z i.e. 2.33 for right tailed test at 1% level of significance, it is not significant and the null hypothesis is accepted at 1% level of significance.

Test of significance for difference of proportions :

Suppose we want to compare two large populations, A and B with respect to the prevalence of a certain attribute among their numbers. Let's take two independent large samples of sizes n_1 and n_2 from the populations A and B respectively. Let X_1 and X_2 be the observed number of successes i.e. the number of units possessing the given attribute in these samples respectively. Then.

p_1 = Observed proportion of successes in the sample from population $A = \frac{X_1}{n_1}$

p_2 = Observed proportion of successes in the sample from population $B = \frac{X_2}{n_2}$

Here, $E(p_1) = P_1$, $\qquad E(p_2) = P_2$

$$\text{Var. } (p_1) = \frac{P_1Q_1}{n_1}, \qquad \text{Var. } (p_2) = \frac{P_2Q_2}{n_2}$$

$$\text{Also, S.E. } (p_1 - P_2) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$$

$$Z = \frac{(p_1 - p_2) - E.(p_1 - p_2)}{S.E.(p_1 - p_2)}$$

$$= \frac{(p_1 - p_2) - (p_1 - p_2)}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}}$$

Under the null hypothesis $H_o : P_1 = P_2$ i.e. the population are the same, the test statistic for the difference of proportions becomes :

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

Since $P_1 = P_2 = P$ and $Q_1 = Q_2 = Q$

If P, the common population proportion (under Ho) is not known, we use its unbiased estimate provided by both the samples taken together, given by

$$\hat{P} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

Again, under the null hypothesis $H_o : p_1 = p_2$ i.e. the sample proportions are equal which implies that the difference in population proportions will not be revealed by the samples from these populations, the test statistic is

$$|Z| = \frac{|p_1 - p_2|}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \sim N(0,1)$$

Example :

In a certain district A, 450 persons were considered regular consumers of tea out of a sample of 1000 persons. In another district B, 400 were regular consumers of tea out of a sample of 800 persons. Do these facts reveal a significant difference between the two districts as for tea drinking habit is concerned? Use 5% level.

Solution :

Given, $n_1 = 1000$, $n_2 = 800$

p_1 = Sample proportion of tea drinkers in district A $= \frac{450}{1000} = 0.45$

p_2 = Sample proportion of tea drinkers in district B $= \frac{400}{800} = 0.5$

Null hypothesis $H_o : P_1 = P_2$ i.e. there is no significant difference between the two districts as far as tea drinking habit is concerned.

Alternative hypothesis $H_1 : P_1 \neq P_2$ (two-tailed)

Under H_o, the test statistic is

$$Z = \frac{p_1 - p_2}{S.E.(p_1 - p_2)}$$

$$= \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

$$\text{Now, } \hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{450 + 400}{1000 + 800} = \frac{850}{1800} = 0.472$$

$$\therefore \hat{Q} = 1 - \hat{P} = 1 - 0.472 = 0.528$$

$$\therefore Z = \frac{0.45 + 0.5}{\sqrt{0.472 \times 0.528\left(\frac{1}{1000} + \frac{1}{800}\right)}} = \frac{-0.05}{\sqrt{0.25 \times 2.25 \times 10^{-03}}}$$

$$= \frac{-0.05}{0.0237} = 2.11$$

Since the calculated value of Z = 2.11 is greater than the critical value of Z at 5% level of significance for two-tailed test i.e. 1.96, it is significant at 5% level of significance. Hence, the null hypothesis is rejected and we conclude that there is significant difference between the two districts as far as tea drinking habit is concerned.

Example :

In two large populations there are 30% and 25% respectively of fair haired people. Is this difference likely to be hidden in samples of 1,200 and 900 respectively from the two populations?

Solution :

Given, $n_1 = 1200$, $n_2 = 900$

$p_1 = 30\% = 0.30 \Rightarrow Q_1 = 0.70$

$p_2 = 25\% = 0.25 \Rightarrow Q_2 = 0.75$

Under H_o; that the difference in population proportions is likely to be hidden in samples, the test statistic is

$$|Z| = \frac{P_1 - P_2}{\frac{P_1 - Q_1}{n_1} + \frac{P_2 - Q_2}{n_2}} \sim N(0,1)$$

$$= \frac{0.30-0.25}{\sqrt{\frac{0.30\times0.70}{1200}+\frac{0.25\times0.75}{900}}} = \frac{0.05}{\sqrt{0.000175+0.000208}}$$

$$= \frac{0.05}{\sqrt{0.000383}} = \frac{0.05}{0.0196} = 2.55$$

Since $|Z| > 1.96$, it significant at 5% level of significance. Hence null hypothesis is rejected and we conclude that the difference in population proportions is not likely to be hidden in these samples, i.e., these samples would reveal the difference in population proportions.

Check your progress :

1. What are the assumptions of large sample tests?
2. What do you mean by sampling for attributes? Develop the large sample test for testing the significance of an observed sample proportion.
3. Explain the large sample test of significance for mean.
4. Distinguish between large sample and small sample tests of sig nificance.
5. Are small sample tests valid for large samples?

2.8 Tests of significance based on 't' distribution :

The sampling distribution of any statistic in the standardized form is asymptotically normally distributed. For example,

$$Z=\frac{\bar{x}-E(\bar{x})}{S.E.(\bar{x})}=\frac{\bar{x}-\mu}{\sigma|\sqrt{n}}\sim N(0,1) \text{ as } n\rightarrow\alpha$$

But if sample size n is small, then the distributions of standardized result, we can not apply normal test. So, to deal with small samples, now techniques and tests of significance known as it is very difficult to distinguish between small samples and large samples. Generally, a sample is termed as small if $n \le \alpha$ It is important to note here that 'Exact Sample Test' can be used even for large samples but large sample tests can not be used for small samples.

The basic fundamental assumptions of all the exact sample tests are—

(1) The parent population from which the sample is drawn is normally distributed.

(2) The sample is random and independent of each other.

Student's 't' distribution :

If $x_1, x_2, \ldots\ldots x_n$ is a random sample of size n from a normal population with mean μ, and variance σ^2, then student's 't' statistic is defined as——

$$t = \frac{\overline{x} - \mu}{S|\sqrt{n}}$$

$\overline{x} = \frac{\Sigma x}{n}$, is the sample mean.

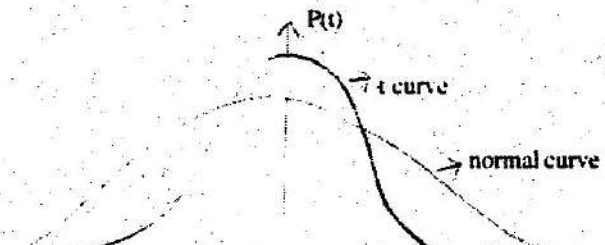
$S^2 = \frac{1}{n-1}\Sigma(x - \overline{x})^2$ is an unbiased estimate of the population variance σ^2 't' statistic defined above follows student's t-distribution with V=(n-1) degrees of freedom and with probability density function (p.d.f.)

$$P(t) = \text{Const.}\frac{1}{\left(1+\frac{t^2}{V}\right)\frac{(V+1)}{2}} - \alpha < t < \alpha$$

We can write.

$$t = \frac{\overline{x} - \mu}{S|\sqrt{n}} \sim t_{n-1}$$

comparison between normal curve and corresponding 't' curve.

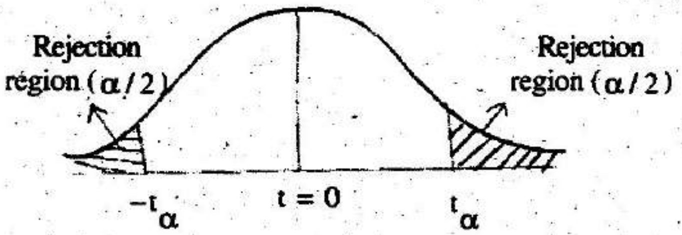


Critical values of 't'

The Critical values of t at the level of significance α and required degrees of freedom V for two-tailed test are given by

$$P[|t| > t_v(\alpha)] = \alpha$$

$$\Rightarrow P[|t| \leq t_v(\alpha)] = 1-\alpha$$



Since t-distribution is symmetric about t = 0, the significant values at the level of significance α for a single tailed (right or left) test can be obtained from the table of two-tailed test by looking the value at level of significance 2α, i.e. $t_v(0.05)$ for single-tailed test = $t_v(0.10)$ for two-tailed test and = $t_v(0.01)$ for single-tailed test = $t_v(0.02)$ for two-tailed test.

Application of t-distribution :

(1) t-test for significance of single mean, population variance being unknown.

(2) t-test for the significance of the difference between two sample means, the population variances being equal but unknown.

(3) t-test for significance of an observed sample correlation coefficient.

Test for single mean :

In this case, we are interested to test :

(1) If the given normal population has a specified value of the population mean, say μ_o.

(2) If the sample mean $\bar{x}$ differs significantly from specified value of population mean.

(3) If a given random sample $x_1, x_2, \ldots\ldots x_n$ of size n has drawn from a normal population with specified mean, μ_o.

The test statistic is

$$t = \frac{\bar{x} - \mu_o}{S/\sqrt{n}} \sim t_{n-1}$$

Computing the test statistic, we compare it with the tabulated value of t for (n-1) d.f. at certain level of significance. If calculated $|t|$ is greater than tabulated 't', we say that it is significant and H_0 is rejected. Again, if calculated $|t|$ is less than tabulated t, H_0 is accepted at the given level of significance.

Assumptions for student's t-test :

Student's t-test is based on the following assumptions—

(1) The parent population from which the sample is drawn is normal.

(2) The sample observations are independent, i.e. the given sample is random.

(3) The population standard deviation σ is unknown.

Example :

A random sample of size 20 from a normal population gives a sample mean of 42 and sample standard deviation of 6. Test the hypothesis that the population mean is 44.

Solution :

Given, $n = 20, \ \overline{x} = 42, \ s = 6$

Null hypothesis $H_0 : \mu = 44$ i.e. the sample mean $\overline{x} = 42$ does not differ significantly from the population mean $\mu = 44$.

Alternative hypothesis $H_1 : \mu \neq 44$ (two-tailed)

Under H_o, the test statistic is

$$t = \frac{\overline{x} - \mu}{\sqrt{\frac{S^2}{n}}} = \frac{\overline{x} - \mu}{\sqrt{\frac{S^2}{n-1}}} \sim t_{n-1} = t_{19}$$

$$\therefore t = \frac{42 - 44}{\sqrt{\frac{62}{20-1}}} = \frac{-2}{\sqrt{\frac{36}{19}}} = \frac{-2}{\sqrt{1.89}} = \frac{-2}{1.37} = -\ 1.46$$

Since the calculated value of $|t| <$ the critical value of $t_{0.05}$ for 19 degrees of freedom = 2.09, it is not significant and H_o is accepted. We can conclude that the sample mean does not differ significantly from the population mean.

Again, $|t| <$ the tabulated $t_{0.01}$ = 2.86 for 19 d.f., it is also not significant at 1% level of significance. So, H_o is accepted and we may conclude that sample mean does not differ significantly from the population mean.

Example :

Prices of shares of a company on the different days in a month were found to be

66, 65, 69, 70, 69, 71, 70. 63, 64 and 68.

Discuss whether the mean price of the shows in the month is 65.

Solution :

Null hypothesis, $H_o : \mu = 65$ i.e., the mean price of the shows in the month is 65.

Alternative hypothesis $H_o : \mu \neq 65$ (two-tailed)

x	66	65	69	70	69	71	70	63	64	68	Total
d=x–69	–3	–4	0	1	0	2	1	–6	–5	–1	–15
d^2	9	16	0	1	0	4	1	36	25	1	93

$$\bar{x} = A + \frac{\Sigma d}{n} = 69 + \frac{(-15)}{10} = 69 - \frac{-15}{10} = 69–1.5 = 67.5$$

$$S^2 = \frac{1}{n-1}\left[\Sigma d^2 - \frac{(\Sigma d)^2}{n}\right] = \frac{1}{10-1}\left[93 - \frac{(-15)^2}{10}\right] = \frac{1}{9}\left[93 - \frac{225}{10}\right]$$

$$= \frac{1}{9}[93 - 22.5] = \frac{1}{9}[70.5] = 7.83$$

Under H_o, the test statistic is

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{S^2}{n}}} = \frac{67.5 - 65}{\sqrt{\frac{(7.83)^2}{10}}} = \frac{2.5}{\sqrt{\frac{61.31}{10}}} = \frac{2.5}{\sqrt{6.131}} = \frac{2.5}{2.48} = 1.008$$

Since t =1.008< tabulated $t_{0.05} = 2.26$ for 9 d.f., it is not significant at 5% level of significance and the null hypothesis is accepted. We can conclude that the mean price of the shows in the month is 65.

Example :

A sample of size 9 from a normal population gives $\bar{x} = 15.8$ and $s_x^{\ 2} = 10.3$. Find 95% and 99% interval for population mean.

Solution :

Given, $\bar{x} = 15.8,\ s_x^{\ 2} = 10.3,\ n = 9$

95% confidence limits for population mean μ are given by

$$\bar{x} \pm t_{0.05} \times \frac{S}{\sqrt{n}}$$

Now,

$$\frac{S}{\sqrt{n}} = \frac{s}{\sqrt{n-1}} = \sqrt{\frac{10.3}{9-1}} = \sqrt{\frac{10.3}{8}} = 1.135$$

$$\begin{aligned} \bar{x} \pm t_{0.05} \times \frac{S}{\sqrt{n}} \\ &= 15.8 \pm 2.31 \times 1.135 \quad [\therefore t_{0.05} \text{ for 8 d.f. } = 2.3.1] \\ &= 15.8 \pm 2.622 = (13.178,\ 18.422) \\ &\text{i.e. } 13.178 < \mu < 18.422 \end{aligned}$$

Again.

99% confidence limits are given by

$$\begin{aligned} \bar{x} \pm t_{0.01} \times \frac{S}{\sqrt{n}} \\ &= 15.8 \pm 3.36 \times 1.135 \quad [\therefore t_{0.01} \text{ for 8 d.f. } = 3.36] \\ &= 15.8 \pm 3.8136 = (11.9864,\ 19.6136) \\ &\text{i.e. } 11.9864 < \mu < 19.6136 \end{aligned}$$

t-tst for difference of means :

Suppose $x_1, x_2, \ldots\ldots x_{n1}$ and $y_1, y_2, \ldots\ldots y_{n2}$ are two independent random samples drawn from two normal populations having the same means, the population variances being equal. We set up the null hypothesis $H_o : \mu_x = \mu_y$ i.e. the samples have been drawn from the normal population with the same means. In other words, the sample means $\bar{x}$ and $\bar{y}$ do not differ significantly. Under the assumption that $\sigma_1^{\ 2} = \sigma_2^{\ 2} \Rightarrow \sigma^2$ Population variance are equal but unknown, the test statistic under H_o is

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n1} + n_2 - 2$$

$$\bar{x} = \frac{1}{n_1}\Sigma x, \qquad \bar{y} = \frac{1}{n_2}\Sigma y$$

$$S^2=\frac{1}{n_1+n_2-2}\left[\Sigma(x-\bar{x})^2+\Sigma(y-\bar{y})^2\right]$$ is an unbiased estimate of the common population variance σ^2 based on both the samples.

But comparing the calculated value of t with the critical value of t for n_1+n_2-2 d.f. at the given level of significance, the null hypothesis is either rejected or accepted.

Assumptions—

(1) Parent populations from which samples have been drawn are normally distributed.

(2) The two samples are random and independent of each other.

(3) $\sigma_1^2=\sigma_2^2=\sigma^2$ i.e., population variances are equal and unknown.

Example :

Two salesmen A and B are working on a certain district. From a sample survey conducted by the head office, the following results are obtained.

	A	B
No. of sales	20	18
Average sales (in Rs.)	170	205
Standard deviation (in Rs.)	20	25

Test whether there is any significant difference in the average sales between the two salesmen.

Solution : Given, $n_1=20$ $\qquad n_2=18$

$\bar{x}=170$ $\qquad \bar{y}=205$

$s_x=20$ $\qquad s_y=25$

Null hypothesis $H_o:\mu_x=\mu_y$ i.e. there is not any significance difference in the average sales between the two districts A and B.

Alternative hypothesis $H_1:\mu_x\neq\mu_y$ under H_o, the test statistic is

$$t=\frac{\bar{x}-\bar{y}}{\sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\sim t_{n_1+n_2-2}=t_{36}$$

Again,

$$S^2=\frac{n_1s_x^2+n_2s_y^2}{n_1+n_2-2}=\frac{20\times400+18\times625}{36}=\frac{8000+11{,}250}{36}$$

$$= \frac{19,250}{36} = \frac{19,250}{36} = 534.72$$

$$\therefore t = \frac{170-205}{\sqrt{534.72\left(\frac{1}{20}+\frac{1}{18}\right)}} = \frac{-35}{\sqrt{534.72 \times 0.106}} = \frac{-35}{\sqrt{56.68}}$$

$$= \frac{-35}{7.53} = -4.65$$

Since $|t| = 4.65$ is greater than $t_{0.01}$ and $t_{0.05}$ for 36 d.f., it is significant at 1% and 5% level of significance. So, H_0 is rejected and we can conclude that there is significant difference in average sales between two districts.

Example :

Samples of two types of electric light bulbs were tested for length of life and following data were obtained.

	Type I	**Type II**
Sample No.	$n_1 = 8$	$n_2 = 7$
Sample Means	$\bar{x}_1 = 1234$ hrs.	$\bar{x}_2 = 1.0364$ hrs.
Sample S.D.	$s_1 = 36$ hrs.	$s_2 = 40$ hrs.

Is the difference in the means sufficient to warrant that type I is superior to type II regarding length of life?

Solution :

Given that there are two types of bulbs.

	Type I	**Type II**
Sample No.	$n_1 = 8$	$n_2 = 7$
Sample Means	$\bar{x}_1 = 1,234$ hrs.	$\bar{x}_2 = 1,036$ hrs.
Sample S.D.	$s_1 = 36$ hrs.	$s_2 = 40$ hrs.

Null hypothesis $H_0 : \mu_1 = \mu_2$ i.e. there is not any difference in the two types of bulbs regarding the length of life or, the avarage mean life of the two types of bulbs, are equal.

Alternative hypothesis : $H_1 : \mu_1 > \mu_2$ (Right-tailed)

Under H_0, the test statistic is

$$t = \frac{\overline{x}_1 - \overline{y}_2}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2} = t$$

$$S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = \frac{8\times(36)^2 + 7\times(40)^2}{8+7-2} = \frac{10.368 + 11.200}{13}$$

$$= \frac{21,568}{13} = 1659.08$$

$$t = \frac{1234-1036}{\sqrt{1659.08\left(\frac{1}{8}+\frac{1}{7}\right)}} = \frac{198}{\sqrt{444.396}} = \frac{198}{21.08} = 9.39$$

Since t = 9.39> the tabulated value $t_{0.05} = 2.16$ and $t_{0.01} = 3.01$ at 13 d.f., so it is significant at both 5% and 1% level of significance. Null hypothesis is rejected and we may conclude that type I bulbs are superior to type II bulbs.

Paired t-test :

In the t-test for difference of means, the two samples were independent of each other. In a particular case where—

(i) The sample sizes are equal i.e. $n_1 = n_2 = n$ (say) and

(ii) observations $(x_1, x_2, \ldots\ldots x_n)$ and $(y_1, y_2, \ldots\ldots y_n)$ are not completely independent but they are dependent in pairs i.e. the pairs of observations $(x_1, y_1), (x_2, y_2), \ldots\ldots (x_n, y_n)$ correspond to the 1st, 2nd,.......n-th unit respectively.

Let $x_1, x_2, \ldots\ldots x_n$ be the sales of the product in 'n' departmental stores for a certain period before advertisement campaign and $y_1, y_2, \ldots\ldots y_n$ be the corresponding sales of the same product for same period in the same departmental stores respectively. Suppose (x_i, y_i), i=1,2,....n is the pair of sales in the ith departmental store before and after advertisement.

Let $d_i = x_i = y_i$, $i = 1,2,\ldots n$ denote the difference in the observations for the ith unit. Under the null hypothesis that the increments are just by chance and not due to advertisement campaign, $H_0: \mu_x = \mu_y$, the test statistic is $t = \frac{\bar{d}}{S/\sqrt{n}} \sim t_{n-1}$. $d = x - y$. $\bar{d} = \frac{1}{n}\Sigma d$

$$S^2 = \frac{1}{n-1}\Sigma\left(d-\bar{d}\right)^2 = \frac{1}{n-1}\left[\Sigma d^2 - \frac{(\Sigma d)^2}{n}\right]$$

Example :

Memory capacity of 10 students was tested before and after training. State whether the training was effective or not from the following scores :

Roll No.	:	1	2	3	4	5	6	7	8	9	10
Before training	:	12	14	11	8	7	10	3	0	5	6
After training	:	15	16	10	7	5	12	10	2	3	8

Solution : Memory capacity of students before training (x) and after training (y) are paired together.

Null hypothesis $H_0 : \mu_1 = \mu_2$ i.e. mean scores before training and after training are same or there is no significant change in memory capacity after the training programme.

Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$

Under H_o, the test statistic is $t = \frac{\bar{d}}{S/\sqrt{n}} \sim t_{n-1} = t_9$

Roll No.	x	y	d=(x-y)	d^2
1	12	15	-3	9
2	14	16	-2	4
3	11	10	1	1
4	8	7	1	1
5	7	5	2	4
6	10	12	-2	4
7	3	10	-7	4
8	0	2	-2	4
9	5	3	2	4
10	6	8	-2	4
			$\Sigma d = -12$	$\Sigma d^2 = 84$

$$\therefore \bar{d} = \frac{\Sigma d}{n} = \frac{-12}{10} = -1.2$$

$$S^2 = \frac{1}{n-1}\left[\Sigma d^2 - \frac{(\Sigma d)^2}{n}\right] = \frac{1}{10-1}\left[84 - \frac{(-12)^2}{10}\right]$$

$$= \frac{1}{9}\left[84 - \frac{144}{10}\right] = \frac{1}{9}[84 - 14.4] = \frac{1}{9}\times 69.6 = 7.73$$

$$|t| = \frac{|\bar{d}|}{\sqrt{S^2/n}} = \frac{1.2}{\sqrt{\frac{7.73}{10}}} = \frac{.1.2}{\sqrt{0.773}} = \frac{1.2}{0.879} = 1.365$$

Since $|t| = 1.365 <$ the tabulated $t_{0.05} = 2.26$ for 9 d.f., it is not significant at 5% level of significance. So, null hypothesis is accepted and we conclude that there is no significant change in memory capacity after the training programme or, training programme is not effective in enhancing memory capacity of the students.

t-test for significance of an observed sample correlation coefficient :

Suppose that a random sample (x_i, y_i), $i = 1,2,....,n$ of size n has been drawn from a bevariate normal distribution. Let, r be the observed sample correlation coefficient. Prof. R.A. Fisher proved that under the null hypothesis $H_o : P = O$, i.e., the variables are uncorrelated in the population, the statistic

$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2} \sim t_{n-2}$$

95% confidence limits for t :

$$r \pm t_{0.05}(n-2) \times S.E.(r) = r \pm t_{0.05}(n-2) \times \frac{(1-r^2)}{\sqrt{n}}$$

99% confidence limits for t :

$$r \pm t_{0.01}(n-2) \times S.E.(r) = r \pm t_{0.01}(n-2) \times \frac{(1-r^2)}{\sqrt{n}}$$

Example :

A random samples of 27 pairs of observations from a normal population gives a correlation coefficient of 0.6. Is it likely that the variables in the population are uncorrelated?

Solution :

Given $n = 27$ and $r = 0.6$

Null hypothesis $H_o : P = 0$ i.e. the variables are uncorrelated in the population. Alternative hypothesis $H_1 : P \neq 0$, i.e. (two-tailed) under H_o, the test statistic is

$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2} \sim t_{n2} = t_{25}$$

$$\therefore t = \frac{0.6}{\sqrt{1-(0.6)^2}} \times \sqrt{27-2} = \frac{0.6}{\sqrt{1-0.36}} \times \sqrt{25} = \frac{0.6}{\sqrt{0.64}} \times 5$$

$$= \frac{0.6}{0.8} \times 5 = 0.75 \times 5 = 3.75$$

Since $t = 3.75$ is greater than the tabulated $t_{0.05}$ for 25 d.f = 2.06, it is significant at 5% level of significance we reject the null hypothesis and conclude that the variables are not uncorrelated in the population.

Again $t = 3.75 > t_{0.01} = 2.79$ for 25 d.f, it is significant at 1% level of significance. So, null hypothesis is rejected and we conclude that the variables are correlated in the population.

Note :

95% confidence limits for t :

$$0.6 \pm 2.06 \times \frac{1-(0.6)^2}{\sqrt{25}} = 0.6 \pm 2.06 \times \frac{0.64}{5} = 0.6 \pm 2.06 \times 0.128$$

$$= 0.6 \pm 0.264 = (0.336, 0.864)$$

99% confidence limits for t :

$$0.6 \pm 2.79 \times \frac{1-(0.6)^2}{\sqrt{25}} = 0.6 \pm 2.79 \times 0.128 = 0.6 \pm 0.357$$

$$= (0.243, 0.957)$$

Cheek your progress

1. What do you mean by Exact Sample Tests?
2. How the Exact Sample Tests different from harge Sample Tests? Can these tests be used for large samples also?
3. Define student's t-statistic.
4. State the properties of the t-distribution.
5. Explain the t-test for testing the significance of the difference between two sample means.
6. What do you understand by the paired t-test? Under what conditions this test be applied?
7. Explain the t-test for testing the significance of an observed sample correlation.

2.9 Tests based on Chi-square distribution :

If the sample size n is small, the sampling distribution of the statistic in its standardised form is not normal and so the normal test can not be applied. Hence, for small samples, we apply the exact sample tests like t, z, F and Chi-square tests.

Chi-square distribution :

The square of a standard normal variables is called a Chi-square variate with 1 degree of freedom. If X is a random variable following normal distribution with mean μ and standard deviation σ then $\frac{X-\mu}{\sigma}$ is a standard normal variate. Therefore, $\left(\frac{X-\mu}{\sigma}\right)^2$ is a Chi-square (denoted by the Greek Letter χ^2) variate with 1 d.f.

If $x_1, x_2, \ldots.. x_v$ are V independent random variables following normal distribution with means $\mu_1, \mu_2, \mu_3, \ldots.. \mu_v$ and standard deviations $\sigma_1, \sigma_2, \sigma_3, \ldots.. \sigma_v$ respectively, then the variate

$$\chi^2 = \left(\frac{X_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{X_2-\mu_2}{\sigma_2}\right)^2 + \ldots\ldots \left(\frac{X_v-\mu_v}{\sigma_v}\right)^2$$

$$= \sum_{i=1} \left(\frac{\lambda_i - \mu_i}{\sigma_i} \right)$$

follows Chi-square distribution with V degrees of freedom.

Application of the χ^2 distribution :

Chi-square distribution has a number of applications. some of them are as follows :

(i) Chi-square test of goodness of fit.

(ii) χ^2 test for independence of attributes.

(iii) To test if the population has a specified value of the variance σ^2.

Conditions for the validity of Chi-square test :

(1) N. the total frequency. should be reasonably large. say greater than 50.

(2) The sample observations should be independent. That is. no individual item should be included twice or more in the sample.

(3) The constraints on the cell frequencies. if any, should be linear.

(4) No theoretical frequency should be small. Preferably each theoretical frequency should be larger than 10 but in any case not less than 5. In case if it is less than 5, we use the technique of 'pooling' which consists in adding the frequencies which are less than 5 with the preceding or succeeding frequencies so that their sum is greater than 5 and adjust the degrees of freedom accordingly.

Degrees of freedom :

The degree of freedom denotes the extent of independence enjoyed by a given set of observed frequencies. Suppose we are given a set of n observed frequencies which are subjected to k independent constraints, then

d.f. = (Number of frequencies) —

(Number of independent constraints on them)

$\Rightarrow V = n - K$

Hence if we are given n frequencies $(O_1, O_2, ...O_n)$ subject to the linear constraint $\Sigma O = \Sigma E = N$. then for the application of χ^2 test, $V = n-1$.

If the given frequency distribution is used to compute the parameters of a theoretical distribution and if these parameters are used to obtain the theoretical (expected) frequencies of the distribution, then we subtract 1 degrees of freedom for each parameter estimated and poisson distribution to the given set of data, we lose 1 d.f. for applying χ^2 test because for binomial distribution we have to estimate only one parameter P, n being given in the data. For poisson distribution, we have to estimate only one parameter m. But for normal distribution, we have to estimate two parameters μ and σ^2 and consequently we lose two degrees of freedom for applying χ^2 test. Moreover, if in case binomial distribution, the hypothetical value of P is given, we do not lose any degrees of freedom.

Again, if any of the theoretical cell-frequencies is less than 5, we pool it with proceeding or succeeding cell frequency is greater than 5. In general, the degrees of freedom for the χ^2 test of goodness of fit are given by

$V = n - 1 - K_1 - K_2$ where

(1) 1 d.f. is lost due to the linear constraint $\Sigma O = \Sigma E = N$.

(2) K_1 is the number of parameter computed and used in estimating the theoretical frequencies of the distribution.

(3) K_2 is the number of d.f. lost in pooling of cell frequencies which are less than 5.

Chi-square test of goodness of fit :

Suppose we have to test whether the results obtained from some experiment Support a particular hypothesis or theory. This is called χ^2 test of goodness of fit.

Under the null hypothesis that there is no significant difference between the observed (experimental) and the theoretical or hypothetical values, the test statistic as proved by Karl Pearson.

$$\chi^2 = \sum_{i=1}^{n}\left(\frac{O_i - E_i}{E_i}\right)^2 = \left(\frac{O_1 - E_1}{E_1}\right)^2 + \left(\frac{O_2 - E_2}{E_2}\right)^2 + \ldots\ldots \left(\frac{O_n - E_n}{E_n}\right)^2$$

Follows χ^2 distribution with $v = n - 1$ degrees of freedom where

$O_1, O_2, \ldots\ldots O_n$ are the observed frequencies and $E_1, E_2, \ldots\ldots E_n$ are the corresponding expected or theoretical frequencies.

If the computed value of χ^2 is less than the corresponding tabulated value of χ^2 for (n-1) d.f. at the given level of significance then it is not significant and null hypothesis will be accepted. This means that the discrepancy between observed values and the expected values may be attributed to chance i.e. fluctuations of sampling and there is good correspondence between theory and experiment. If the calculated value of χ^2 is greater than the tabulated value, it is said to be significant. Then we can conclude that the experiment does not support the theory.

The observed and expected frequencies have a very important relation.

$$\sum_{i=1}^{n} O_i = \sum_{i=1}^{n} E_i = N \quad \text{or,} \quad \Sigma O = \Sigma E = N$$

For numerical problems, it is useful to use $\chi^2 = \Sigma\left(\frac{O^2}{E}\right) - N$

Proof : $\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$

$$= \Sigma\left[\frac{O^2 + E^2 - 20E}{E}\right] = \Sigma\left[\frac{O^2}{E} + E - 20\right] = \Sigma\frac{O^2}{E} + \Sigma E - 2\Sigma 0$$

$$= \Sigma\frac{O^2}{E} + N - 2N = \Sigma\frac{O^2}{E} - N$$

It should be noted that the χ^2 test depends only on the set of observed and expected frequencies and on degrees of freedom V. It does not make any assumption regarding the parent population from which the observations are taken. Since χ^2 does not involve any population parameters it is termed as a statistic and the test is known as Non-Parametric test.

Example :

In a set random numbers, the digits 0,1,.....9 were found to have the following frequencies :

Digit	:	0	1	2	3	4	5	6	7	8	9
F	:	43	32	38	27	38	52	36	31	39	24

Test whether they are significantly different those expected on the hypothesis of uniform distribution.

Solution :

Null hypothesis set up the null hypothesis that the digits 0,1,2,3,....9 are uniformly distributed, i.e. all the digits occur equally frequently.

Then under the null hypothesis, the expected frequency for each of the digits 0,1,2,3,...,9 is

$$\frac{43+32+38+27+38+52+36+31+39+24}{10}=\frac{360}{10}=36$$

Digits	Oservied frequency (O)	Expected frequency (O)	$(O-E)$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
0	43	36	7	49	1.36
1	32	36	-4	16	0.44
2	38	36	+2	4	0.11
3	27	36	-9	81	2.25
4	38	36	2	4	0.11
5	52	36	16	256	7.11
6	36	36	0	0	0
7	31	36	-5	25	0.694
8	39	36	3	9	0.25
9	24	36	-12	144	4
	360	360	0		16.325

$$\chi^2=\Sigma\left[\frac{(O-E)^2}{E}\right]=16.325$$

Since we are given 10 frequencies subjected to only one linear constraint $\Sigma O=\Sigma E=360$, Degrees of freedom $10-1=9$

Tabulated value of χ^2 for 9 d.f. at 5% level of significance is 16.92. Since calculated value of χ^2 = 16.325 is less than the tabulated value 16.92, it is not significance and null hypothesis is accepted at 5% level of significance. So, we conclude that the digits 0,1,2,...,9 are uniformly distributed i.e. all the digits occur equally frequently.

Example :

The figures below are (a) the frequencies of a distribution and (b) the frequencies of the normal distribution having the same mean, standard deviation and total frequency as in (a)

(a)	1	12	66	220	495	792	924	792	495	220	66	12	1
(b)	2	15	66	210	484	799	944	799	484	210	66	15	2

Apply χ^2 test of goodness of fit.

Solution : We set up the null hypothesis that normal distribution is a good fit to the given frequency distribution, i.e. the observed frequencies (O) and the theoritical normal frequencies (E) do not differ significantly.

O	E	(O - E)	$(O - E)^2$	$\frac{(O-E)^2}{E}$
$\left.\begin{matrix}1\\12\end{matrix}\right\}13$	$\left.\begin{matrix}2\\15\end{matrix}\right\}17$	–4	16	0.94
66	66	0	0	0.00
220	210	10	100	0.48
495	484	11	121	0.25
792	799	–7	49	0.06
924	944	–20	400	0.42
792	799	–7	49	0.06
495	484	11	121	0.25
220	210	10	100	0.48
66	66	0	0	0.00
$\left.\begin{matrix}12\\1\end{matrix}\right\}13$	$\left.\begin{matrix}15\\2\end{matrix}\right\}17$	–4	16	0.94
4096	4096	0		3.88

$$\therefore \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 3.88$$

Here we are given 13 frequencies, d.f. $= 13 - 1 - 2 - 2 = 0$

1d.f. being lost because of the linear constraint $\Sigma O = \Sigma E = 4096$, 2d.f.

are lost because the parameters m and σ^2 are estimated from the given distribution and used in computing the expected (theoritical) frequencies of the normal distribution, 2d.f. are lost in pooling because the first and the last frequencies are less than 5 and they are pooled with succeeding and prociding frequency respectively so that the resulting frequencies are greater than 5.

Tabulated value of χ^2 for 8d.f. at 5% level of significance is 15.507. Since calculated value of $\chi^2 = 3.88$ is much less than the tabulated value i.e. 15.507, it is highly non-significant. Hence we may acceot the null hypothesis at 5% level of significance and conclude that the normal distribution is a good fit to the given distribution.

Chi-square test for independence of attributes :

Lets suppose that the given population consisting of N items is divided in π mutually exclusive and exhaustive classes $A_1, A_2, \ldots\ldots A_r$ with respect to the attribute A so that random selected item belongs to one and only one of the attributes $A_1, A_2, \ldots\ldots A_r$. Again, suppose that the same population is devided into S mutually disjoint and exhaustive classes $B_1, B_2, \ldots\ldots B_s$ with respect to another attribute B so that an item selected at random possesses one and only one of the attributes $B_1, B_2, \ldots\ldots B_s$. The frequency distribution of the items belonging to the classes $A_1, A_2, \ldots\ldots A_r$ and $B_1, B_2, \ldots\ldots B_s$ can be represented in the following $r \times s$ manifold contingency table.

$r \times s$ **manifold contingency table**

A/B	B_1	B_2	....	B_j	....	B_s	Total
A_1	(A_1B_1)	(A_1B_2)	....	(A_1B_j)	....	(A_1B_s)	(A_1)
A_2	(A_2B_1)	(A_2B_2)	....	(A_2B_j)	....	(A_2B_s)	(A_2)
⋮	⋮	⋮		⋮		⋮	⋮
A_i	(A_iB_1)	(A_iB_2)	....	(A_iB_j)	....	(A_iB_s)	(A_i)
⋮	⋮	⋮		⋮		⋮	⋮
A_r	(A_rB_1)	(A_rB_2)	....	(A_rB_j)	....	(A_rB_s)	(A_r)
Total	(B_1)	(B_2)	....	(B_j)	....	(B_s)	

(A_i) is the frequency of the ith attribute A_i i.e: it is the number of persons possessing the attribute (A_i), $i = 1,2,....,r(B_j)$ is the number of persons possessing the attribute B_j, $j = 1,2,.....s$ and (A_iB_j) is the number of persons possessing both the attributes A_i and B_j.

Under the null hypothesis that the two attributes A and B are independent, the expected frequency for (A_iB_j) is given by

$$E[A_iB_j] = N.P[A_iB_j]$$

$$= N.P[A_i \cap B_j]$$

$$= NP[A_i].P[B_j]$$

$$= N \times \frac{(A_i)}{N} \times \frac{(B_j)}{N}$$

$$= \frac{(A_i)(B_j)}{N}$$

It implies that the expected frequency for any of the cell frequencies can be obtained on multiplying the row totals and column totals in which the frequency occurs and dividing the product by the total frequency N.

Here, the test statistic χ^2 follows χ^2 distribution with $(r-1)\times(s-1)$ degrees of freedom. Comparing the calculated value of χ^2 with the tabulated value for $(r-1)\times(s-1)$ d.f. at certain given level of significance we reject or retain the null hypothesis of independence attributes.

Degrees of freedom for $r \times s$ contingency table :

For $r \times s$ contingency table, the table number of frequencies is $n = r \times s$ = rs. These n frequencies are subjected to the following linear constraint :.

(1) r row totals $(A_1),(A_2)$,.....(A_r) are fixed.

(1) s column totals $(B_1),(B_2)$,.....(B_r) are fixed.

Again $\sum_{i=1}^{r} (A_i) = \sum_{j=1}^{s} (B_j) = N$

So, the number of independent constraints is $K = r + s - 1$.

For $r \times s$ contingency table, d.f. are

$$V = n - K = n - (r + s - 1) = rs - (r + s - 1) = rs - r - s + 1$$
$$= (r - 1)(s - 1)$$

For 2×2 contingency table.

degrees of freedom $=(2-1)\times(2-1)=1$

For 4×5 contingency table,

degrees of freedom $=(4-1)\times(5-1)=3\times4=12$ and so on.........

Example :

1000 students at college level were graded according to their I.Q. and the economic conditions of their homes. Use χ^2 test to find out whether ther is any association between economic conditions at home and I.Q.

	I.Q.		
Economic Conditions	**High**	**Low**	**Total**
Rich	460	140	600
Poor	240	160	400
Total	700	300	1000

Solution :

We set up the null hypothesis that the two attributes economic conditious and I.Q. of the students are independent. The expected frequencies are—

$$E(460)=\frac{700\times600}{1000}=420, \qquad E(140)=\frac{300\times600}{1000}=180$$

$$E(240)=\frac{700\times400}{1000}=120, \qquad E(160)=\frac{300\times400}{1000}=120$$

The value of Chi-square is obtained below :

O	E	$(O-E)$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
460	420	40	1600	3.81
140	180	–40	1600	8.80
240	280	–40	1600	5.71
160	120	40	1600	13.33
1000	1000	0		31.74

$$\therefore \chi^2=\Sigma\left[\frac{(O-E)^2}{E}\right]=31.74$$

Since the calculated value of $\chi^2 = 31.74$ is much greater than the tabulated value of $Y^2 = 3.84$ for V=1, it is highly significant. So, the null hypothesis that the two attribute are independent is rejected and we can conclude that they are highly associated.

Example :

Out of a sample of 120 persons in a village. 76 persons were administered a new drug for preventing influenza, and out of them, 24 pensons were attacked by influenza. Out of those who were not administered the new drug, 12 persons were not affected by influenza. Use Chi-square test for finding out whether the new drug is effective or not.

Solution : The above data can be arranged in the following 2×2 contingency table.

New during	Effect of influenza		
	Attacked	Not-attacked	Total
Administered	24	76–24=52	76
Not administered	44–12=32	12	120–76 = 44
Total	24+32=56	52+12=64	120

We set up the null hypothesis that the attributes 'attack by influenza' and the 'administration of the new drug' are independent, i.e. the new durg is not effictive in controlling influenza.

Under the null hypothesis of independence, the expected frequencies are—

$$E(24)=\frac{56\times76}{120}=35.47,\qquad E(32)=\frac{56\times44}{120}=20.53$$

$$E(52)=6\frac{44\times76}{120}=40.53,\qquad E(12)=\frac{44\times64}{120}=23.47$$

O	E	$(O-E)$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
24	35.47	–11.47	131.5609	3.709
32	20.53	11.47	131.5609	6.408
52	40.53	11.47	131.5609	3.246
12	23.47	–11.47	131.5609	5.605
120	120	0		18.968

$$\therefore \chi^2=\Sigma\left[\frac{(O-E)^2}{E}\right]=18.968$$

d.f = (2−1) (2−1) = 1

Since the calculated value of χ^2 i.e. 18.968 is much greater than the tabulated value $\chi^2_{00.5}$ for 1 d.f. = 3.84, it is highly significant and the null hypothesis is rejected at 5% level of significance. So we conclude that the new durg is difinitely effective in controlling influenza.

χ^2 test for the population variance :

Suppose we are interested to test if the given normal population has a specified variance $\sigma^2 = \sigma_0^{\ 2}$ (say).

We set up the null hypothesis $H_0 : \sigma^2 = \sigma_0^{\ 2}$. If $x_1, x_2, x_3, \ldots, x_n$ are the observations from a random sample of size n from the given population, then under the null hypothesis H_0, the statistic

$$\chi^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma_0^{\ 2}} = \frac{ns^2}{\sigma_0^{\ 2}}$$

follows χ^2 distribution with (n−1) d.f. where

$$S^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2$$

Now, comparing the calculated value of χ^2 with the tabulated value for (n−1) d.f. at certain level of significance, the null hypothesis is rejected or retained.

Example :

A sample of 15 values shows the standard deviation to be 6.4. Does this agree with hypothesis that the population standard deviation is 5, the population being normal.

Solution :

We set up the null hypothesis that population standard deviation is $\sigma = 5$. Given, n = 15, s = 6.4

$$\therefore \chi^2 = \frac{ns^2}{\sigma^2} = \frac{15 \times (6.4)^2}{25} = \frac{15 \times 40.96}{25} = \frac{614.4}{25} = 24.576$$

Since the calcuolated value of χ^2 =24.576 is greater than the tabulated value of $\chi^2_{0.05}$ for 14 d.f. = 23.635, it is significant. So, the null hypothesis is rejected and we conclude that population standard deviation is not 5.

Check your progress :

1. What is χ^2 test of goodness of fit?

2. State the conditions for validity of χ^2 test.

3. What do you mean by degrees of freedom?

4. What is pooling?

5. Describe the χ^2 test for independence of attributes.

6. What is a contingency table?

7. Explain how χ^2 test can be used for testing a hypothesis that population has a specified variance $\sigma_0{}^2$.

2.10 Additional Readings :

1. S.C. Gupta, Fundamentals of Statistics, Himalaya Publishing House.

2. S.P. Gupta, Statistical Methods, Sultan Chand & Sons.

3. A.L. Nagar and R.K. Das, Basic statistic, OUP.

4. Padmalochan Hazarika, Essential Statistics for Economics and Commerce, Akansha Publishing House

5. D. Salvatore, Mathematics and Statistics, Schaum's Series, Tata Mc Graw Hill.

UNIT-3

Linear Regression Model and its Estimation

Contents :

3.1 Introduction :

The term regression was introduced by Francis Galton. The modern interpretation of regression is Regression analysis is concerned with the study of the dependence of one variable, the dependent variable, on one or more other variables, the explanatory variables, with a view to estimating and /or predicting the (Population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the letter.

In a regression model, to examine the dependence of the dependent variable on the explanatory variables, first the coefficients (parameters of the regression model) of the explanatory variables should be estimated. There is a possibility of getting numerous estimates of each coefficient. But out of all these only that estimate will be selected which is best as well as significant. The significant coefficients of the explanatory variable has some effect on the dependent variable and otherwise not.

However, the validity of the estimates of the parameters depends upon some standard assumptions. Violation of these classical Linear Regression Model (CLRM) assumptions will result in some problems like auto-correlation, hetero scadasticity and multi collinearity etc.

In the literature the dependent and independent variable has different names. Such as,

Dependent Variable :

Explained variable, predictand. Regressand, Response, Endogenous. Outcome, controlled variable

Independent variable.

Explanatory variable. predictor. Regressor. Stimulus. Exogenous. covariate. control variable.

3.2 Objectives :

This unit is designed to help you understand the concept of linear regression model and its estimation and its related ideas. After reading this unit you will be able to,

- Construct linear regression model.
- Understand different assumptions and properties of OLS and GLS.
- Describe how far one independent variable is able to explain the dependent variable.
- Estimate maximum likelihood methods

3.3 The Two Variable Regression Model :

Typically a detailed econometric model contains a number of equations and each equation contains a number of variables. Given this complexity of the model. some complicated problems arise when we estimate the parameters of the model. Hence, for simplicity the simplest possible case of econometric model which consists of just one equation in two variable will be discussed first.

Denoting the variable by Y and X the two variable linear regression equation will be,

$$Y = f(x). \quad .... (3.1.1)$$

Where Y is the dependent variable and X is the explanatory or independent variable. Here, Variations in Y are explained in terms of the variations in X. Here. Y is said to be regressed to X.

3.4 The Linear Specification :

Equation (3.1.1) merely indentifies the variable X, which is thought to influence the other variable Y.

The simplest relationship between two variables is a linear one, namely.

$$Y = \alpha + \beta x \quad (3.2.1)$$

where α and β are unknown parameters indicating the intercept and slope of the function. The value of α and β will specify the relationship between X and Y.

The linear specification is useful not only because it is simple but also for the fact that many non linear functional forms can be linearised by mathematical manipulations. This means that linear specification is not necessarily a severe restriction. For instance, let a non linear relation—

$$Y = \alpha x^{\beta}$$

Taking logarithms on both the side of the relation we will get,

$$\log Y = \log \alpha + \beta \log X$$

Which is a linear function in both log Y and log X. Denoting in T and in X by and V respectively, we have,

$$W = \alpha + \beta v. \text{ which is linear.}$$

Even when such linearisation is not possible, the linear specification of a non linear function over a small range. Thus the linear specification has a much wider area of applicability than what appears at the first sight.

3.5 Introduction of the Ramdom Disturbance :

Conventional economic theory usually postulets exact functional relationship between variables like the equation (3.2.1). Such an exact relation will give a specific value of Y for any given value of X. But emperical experience are in general not quite exact. To allow for the inexact relationship between economic variables, the econometiecian modifies the deterministic relationship as,

$$Y = \alpha + \beta x + \mu \quad (3.3.1)$$

Here, with the deterministic part of the model a random part or stochastic random disturbance term usually denoted by U. The deterministic part of the model can be represented by a straight line like RL in figure 1.1.

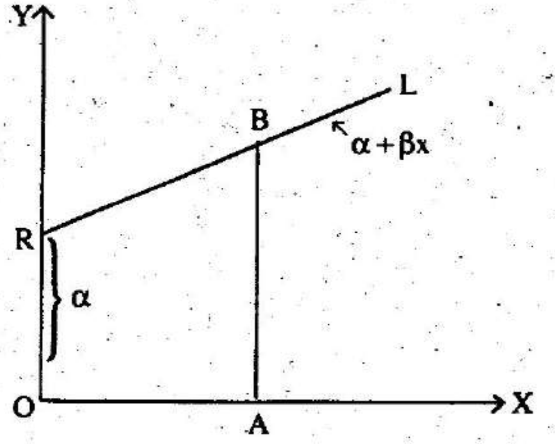


Figure-1.1

Such a line gives the expected value of Y for any given value of X. Here, X takes the values OA, Y is expected to take the value AB $(= \alpha + \beta X)$. But in reality, Y may take a value different from AB because of the influence of the disturbance term U. Smaller the values of U, higher is the possibility of Y taking values closer to $\alpha + \beta X$ and vice-versa. Generally U is assumed to take smaller values with higher probability.

3.6 Interpretation of the Parameters :

Let, the two variable linear regression model is,

$$Y = \alpha + \beta x + \mu.$$

The parameter β is the slope of the regression line is $\beta = \frac{\partial Y}{\partial X}$ or the rate of change in Y to per unit change in X. Thus, β capture the effect of the explanatory variable X on the dependent variable. Keeping other factors constant.

The parameter α is the intercept of the regression line. Intercept is the value of the dependent variable when the value of the independent variable is zero. In econometrics the meaning of α is quite different. Here α Implies the mean effect of those factors which are not explicitly involved in the model. That is in econometrics the parameter α Is to be interpreted as the mean effect of the factors other than which constitute the value of Y.

3.7 Rationalisation of the Disturbance Term :

The stochastic nature of the regression model implies that for every value of X there is a whole probability distribution of Y. In other words the value of Y can never be predicted exactly. This uncertainty regarding the value of Y arises because of the presence of the stochastic term U in the model.

The insertion of the random disturbance term U in the regression model can be justified on the basis of the following arguments.

1. There may be several factors, which influence the dependent variable besides explanatory variables. It is practically impossible to introduce explicitly all the factors influencing the dependent variable. It may be difficult to quantify some of the factors. Even when factors are quantifiable, obtaining statistical data on them may be so small that one would gain very little by introducing them explicitly in the model. Generally, the mean effect of the factors not explicitly included in the model can be captured through the disturbance term U.

2. Though economic theory assumes human behavior to be rational, in practice the same may not be perfectly rational. This randomness of the human behavior can be captured through the disturbance term.

3. A part of the disturbance U may be the errors of observation or measurement present in the observation of the dependent variable Y.

4. The disturbance term may represent such an error which may be due to the imperfect specification of the form of the model.

5. U may represent the error which may be due to the aggregation of data.

6. The disturbance term may represent wrong functional form.

3.8 Exercise :

1. In a linear regression model $Y_t = \alpha + \beta x_t + \mu_t$. explain the nature of the random variable μ_t And give justification for its presence in the model.

[M.A. Previous, 2003 GU]

2. It is hypothesized that household expenditure on energy depends not only on income but also on whether the household is located in a rural or urban area. Formulate a suitable regression model to list the hypothesis of the parameters in the regression equation.

[M.A Previous, 2006 G.U]

3.9 The Least Square Principle of Estimating Regression Parameters :

In order to estimate the parameters α and β of the regression model, one requires observations on the dependent as well as esplanatory variable. Given the observations, the estimation can be done by a method of Ordinary Least Square (OLS).

Let $(X_1, Y_1),(X_2, Y_2)$, ..., (X_n, Y_n) be n pairs of observations on X and Y. (For example for a regression on income, these could be income and consumption expenditures data collected from N sample household. Now using the regression model the t-th observation can be represented as,

$$Y_t = \alpha + \beta x_t + \mu_t \quad \text{............} \; t = 1,2,......,n$$

The value of parameters α and β are not given, so we are to estimate the values of α and β. The estimated value of Y denoted as $\hat{Y}$ is given by,

$$\hat{Y}_t = \hat{\alpha} + \hat{\beta} x_t$$

Where $\hat{\alpha}$ and $\hat{\beta}$ are the estimated values of parameters. The values of $\hat{\alpha}$ and $\hat{\beta}$ should be selected in such a way so that it gives the best possible fit.

The estimated value in general will not be indentical with the observed value of Yt due to the random disturbaance term. Denoting the residue in Y_t (in excess of Y_t over $\hat{Y}_t$) but $\hat{U}_t$. we have

$$\hat{U}_t = Y_t - \hat{Y}_t = Y_t - \hat{\alpha} - \hat{\beta} x_t$$

For n observations we get n residuals $\hat{u}_1, \hat{u}_2, ..., \hat{u}_n$.

That is residual is the difference between the estimated value and observed value, Residual can be positive or negative.

Let us understand

- If U_t is positive, then $Y_t > \hat{Y}_t$
- If U_t is negative, then $Y_t < \hat{Y}_t$

If the residuals are large it means that the estimate is not good, but if residuals are small then it means that the estimate is good. Better the estimation of α and β, nearer will be the estimated value to the observed value and hence smaller would be the magnitude of the residuals than together. Thus choosing the estimators $\hat{\alpha}$ and $\hat{\beta}$, one may look to minimise the overall magnitude of the residuals. But in doing so the first problem is to get a measure of overall magnitude of the rasiduels. A simple sum of the residuals will not serve the purpose as negative residuals will cancel positive ones and we can get a very small sum even when actual residuals are quite large. A sum of their absolute values is sum of the residuals after ignoring sings, will not have this problem. But ignoring of signs makes this sum unsuitable for further algebric manipulation. Moreover, in case of minimisation of absolute deviation, we can't distinguish the deviations within deviations. Therefore the reduction of one unit of deviation from large or small deviation are squared out, the greater deviations get further magnified. Hence while minimising the sum of squares of deviation, more importance will be given to reduce deviation which are already further off. Therefore the sum of squares of residuals is taken as the measure of overall magnitude of the residuals.

Thus in OLS method the regression parameters are so estimated that the sum of squares of residuals becomes minimum. In the present case the sum of squares of residuals (S) becomes,

$$S=\sum_{t=1}^{n}\hat{U}_t^2=\sum_{t=1}^{n}\left(Y_t-\hat{\alpha}-\hat{\beta}x_t\right)^2$$

Given the observations $(X_1, Y_1),(X_2, Y_2), \ldots\ldots\ldots, (X_n, Y_n)$ the sum of squares of residuals is a function of $\hat{\alpha}$ and $\hat{\beta}$ ie for different pairs of $\hat{\alpha}$ and $\hat{\beta}$ the value of S will be different. The OLS estimatores of α and β are those values of $\hat{\alpha}$ and $\hat{\beta}$ for which S takes the smallest possible value.

Mathematically, the first order conditions for minimisation of S with respect to $\hat{\alpha}$ and $\hat{\beta}$ are

$$\frac{\partial S}{\partial\hat{\alpha}}=0 \text{ and } \frac{\partial S}{\partial\hat{\beta}}=0$$

$$\text{Now } \frac{\partial S}{\partial\hat{\alpha}}=0 \Rightarrow \frac{\partial S}{\partial\hat{\alpha}}\sum\left(Y_t-\hat{\alpha}-\hat{\beta}x_t\right)^2=0$$

$$\Rightarrow 2\sum\left(Y_t-\hat{\alpha}-\hat{\beta}x_t\right)(-1)=0$$

$$\left.\begin{array}{r}\Rightarrow\sum\left(Y_t-\hat{\alpha}-\hat{\beta}x_t\right)=0\\ \Rightarrow\sum\hat{U}_t=0\end{array}\right\} \quad \ldots\ldots\ldots\ (3.4.1)$$

(Deviding both sides by-2)

Similarly,

$$\frac{\partial S}{\partial \hat{\alpha}} = 0 \text{ and } \frac{\partial S}{\partial \hat{\beta}} = 0$$

Now $\frac{\partial S}{\partial \hat{\beta}} = 0$

$$\left.\begin{aligned} &\Rightarrow \sum \left(Y_t - \hat{\alpha} - \hat{\beta}x_t\right) x_t = 0 \\ &\Rightarrow \sum \hat{U}_t x_t = 0 \end{aligned}\right\} \quad \text{..........} \; (3.4.2)$$

The equations (3.1) and (3.2) and called OLS normal equations. From (3.1) we have,

$$\sum Y_t - n\hat{\alpha} - \hat{\beta}\sum x_t = 0 \Rightarrow n\hat{\alpha} = \sum Y_t - \hat{\beta}\sum x_i$$

$$\Rightarrow \hat{\alpha} = \overline{Y} - \hat{\beta}\hat{x} \quad \text{............}(3.4.3)$$

From (3.2) we have,

$$\sum Y_t X_t - \hat{\alpha}\sum X_t - \hat{\beta}\sum X_t^2 = 0$$

Next substituting for $\hat{\alpha}$ from (3.3), we have

$$\Sigma Y_t X_t - \hat{Y}\Sigma X_t + \hat{\beta}\overline{X}\Sigma X_t - \hat{\beta}\Sigma X_t^2 = 0$$

$$\Rightarrow \hat{\beta}\left(\Sigma X_t^2 - \overline{X}\Sigma X_t\right) = \Sigma X_t X_t - \hat{Y}\Sigma X_t \quad \Rightarrow \hat{\beta} = \frac{\Sigma Y_t X_t - \overline{Y}\Sigma X_t}{\Sigma X_t^2 - \overline{X}\Sigma X_t}$$

This can be varified as,

$$\Sigma\left(Y_t - \overline{Y}\right)\left(X_t - \overline{X}\right) = \Sigma Y_t X_t - \overline{Y}\Sigma X_t$$

$$\text{and } \Sigma\left(X_t - \overline{X}\right)^2 = \Sigma X_t^2 - \overline{X}\Sigma X_t$$

Thus, $\hat{\beta}$ can be written as,

$$\hat{\beta} = \frac{\Sigma\left(Y_t - \overline{X}\right)\left(Y_t - \overline{Y}\right)}{\Sigma\left(X_t - \overline{X}\right)^2} = \frac{\Sigma x_t y_t}{\Sigma x_t^2} \quad \text{...........} \; (3.4.4)$$

where $x_t = X_t - \overline{X}$ and $y_t = Y_t - \overline{Y}$

Equation (3.4.3; and (3.4.4) represent OLS estimators of α and β. Here conditions for minimisation of S are satisfied automatically.

3.10 The Standard OLS Assumptions :

The OLS estimators possess cerntain desirable properties provided some standard assumptions are stisfied. These assumptions are briefly discussed below.

(i) The random disturbance term has a zero mean for observation in the model,

$$Y_t = \alpha + \beta X_t + u_t, \ \ \Sigma U_t = 0 \ \forall \ t$$

For any given value of the explanatory variable say X_{t1} the disturbance term u_t can take many alternative values, u_t being a random variable, there are specific probabilities of it taking various different possible values of u_t are so distributed that the expected value become equal to zero. For example, suppose we are able to take many repeated observation in which the X value remain specific at X_t. Though X value remains unchanged in these observation the random disturbance can vary across these repeated observations. In some cases the disturbance will be positive and in some other cases it will take negative values. But the frequency of the different values will be such that the positive values tend to be balanced by the negative values ie for each positive value of U there will be a coresponding negative value of some magnitude so that average value tends to be zero.

(ii) The disturbance term has a content variance for all observations. As explained in the above assumption for any specific value of X, say X_t, the disturbance term has a distribution, ie a whole range of possible values with associated probabilities. It is assumed that varience of the different distributions of the disturbances corresponding to the different X values are same. Thus,

$$VU_t = E(U_t - EU_t)^2 = \sigma^2 \forall t$$

$$\Rightarrow EU_t^2 = \sigma^2 \forall t \qquad [EU_t = 0 \text{ by assumption } (1)]$$

(iii) The disturbance in different observations are independently distributed. This means that the value taken by U_t in the observation t is not related to value taken by U_t in another observation S ie,

$$\text{Cov } U_t U_s = E[U_t - EU_t][U_s - EU_s] = 0 \ \forall \ t \neq s$$

$$\text{or } EU_t U_s = 0 \ [EU_t = 0 \ = \ EU_s \text{ by assumption } (1)]$$

(iv) Each disturbance term is normally distributed

$$\text{ie } U_t = N(0, \sigma^2) \ \forall \ t.$$

(v) The explanatory variable (S) is (are) nonstochastic. This means that each X_t value is fixed and there is no question of X_t having a distribution of different possible values with assigned probabilities (as we have in case of U_t).

The assumptions (i), (ii) and (iii) are sometimes jointly stated as U_ts. are independent and identicaly distributed with zero mean and variance σ^2 or in short "U_ts are iid $(0, \sigma^2)$"

(vi) The regression model is linear in the parameters

(vii) The number of observations n must be greater than the number of explanatory variables.

(viii) The regression model should be correctly specified. i.e. there should not be specificatin bias or error in the model used in emperical analysis.

(ix) There must not be perfect linear relationship among the explanatory variables. ie there should be no perfect multicollenearity.

(x) The X values in a given sample must not all be the same. Technically var (X) must be a finite positive number.

3.11 Properties of the OLS Estimators :

From the two variable linear regression model.

$Y_t = \alpha + \beta X_t + u_t$

The estimator of β is,

$$\hat{\beta} = \frac{\Sigma x_t y_t}{\Sigma x_t^2}$$

where, $x_t = X_t - \overline{X}$ and $y_t = Y_t - \overline{Y}$

Now, let us examine the properties of $\hat{\beta}$ under the standard OLS assumptions.

Putting $y_t = Y_t - \overline{Y}$ in $\hat{\beta}$, we have

$$\hat{\beta} = \frac{\Sigma x_t \left(Y_t - \overline{Y}\right)}{\Sigma x_t^2} = \frac{\Sigma x_t Y_t}{\Sigma x_t^2} - \frac{\overline{Y} \Sigma x_t}{\Sigma x_t^2} = \frac{\Sigma x_t y_t}{\Sigma x_t^2}$$

$$\left[\Sigma x_t = \Sigma\left(X_t = \overline{X}\right) = 0, \quad \text{being the sum of deviations from mean value}\right]$$

Thus

$$\hat{\beta} = \Sigma \left[\frac{x_t}{\Sigma x_t^2} \right] Y_t = \Sigma W_t Y_t \qquad \text{where } w_t = \frac{x_t}{\Sigma x_t^2}$$

This shows $\hat{\beta}$ as a linear combination of sample Y values. Hence $\hat{\beta}$ is said to be a linear estimator. For deriving further properties of $\hat{\beta}$, let us first note a few results about W_t.

$$W_t = 0 \text{ .... } (3.5.1)$$

$$\Sigma W_t X_t = \Sigma W_t x_t = 1 \text{ .... } (3.5.2)$$

$$\Sigma W_t^2 = \frac{1}{\Sigma x^2} \text{ .... } (3.5.3)$$

$$\left[W_t = \frac{x_t}{\Sigma x_t^2} \text{ can be varified} \right].$$

Main of $\hat{\beta}$:

$\hat{\beta} = \Sigma W_t Y_t$ and $Y_t = \alpha + \beta x_t + u_t$

we have,

$$\hat{\beta} = \Sigma W_t(\alpha + \beta x_t + u_t) = \alpha \Sigma W_t + \beta \Sigma W_t X_t + \Sigma w_t u_t$$

$$\text{or } \hat{\beta} = \beta + \Sigma W_t X_t \text{ .... } (3.5.4)$$

$$(\because \Sigma W_t = 0 \text{ and } \Sigma W_t X_t = 1)$$

$$E\hat{\beta} = E\beta + E\Sigma W_t U_t = \beta + \Sigma W_t EU_t$$

$$(E\beta = \beta, \text{ since } \beta \text{ is a constant})$$

[W_t is a function of X_ts. X_t being non-stochastic, W_t can be treated as a fixed constant for the purpose of taking the expectation of $W_t u_t$ ie $EW_t u_t = W_t Eu_t$]

Since $EU_t = 0$ by assumption 1

$$E\hat{\beta} = \beta \text{ .... } (3.5.5)$$

In other words, $\hat{\beta}$ is an unbiased estimator of β.

Variance of $\hat{\beta}$:

By difinition,

Variance of $\hat{\beta}$ is $\sigma^2\hat{\beta} = E\left(\hat{\beta} - E\hat{\beta}\right)^2$ Substituting (3.5.4) and (3.5.5) we get.

$$\sigma^2\hat{\beta} = E(\Sigma W_t - U_t)^2$$

$$= E\left[(W_1U_1 + \ldots\ldots + W_nU_n)(W_1U_1 + \ldots\ldots + W_nU_n)\right]$$

∵ Expectations of sum is equal to the sum of expectations.

$$= E\left[\sum_t W_t^2 U_t^2 + \sum_{t\neq s} W_t W_s U_t U_s\right]$$

$$= \sum_t W_t^2 U_t^2 + \sum_{t\neq s} E W_t W_s U_t U_s = \sum_t W_t^2 \sigma_t^2$$

∵ W_ts are non stochastic.

$$\left[\begin{array}{l} \because EU_t^2 = \sigma^2 \text{ by assumption (ii) and} \\ \quad EU_tU_s = 0 \ \forall\, t \neq s \text{ by assumption (iii)} \end{array}\right]$$

$$\text{or } \sigma^2\hat{\beta} = \sigma^2 \Sigma W_t^2$$

$$\text{or } \sigma^2\hat{\beta} = \frac{\sigma^2}{\Sigma W_t^2} \quad \ldots\ldots\ldots\ldots (3.5.6)$$

$$\left[\because \Sigma W_t^2 = \frac{1}{\Sigma x_t^2} \text{ by assumption (3.5.3)}\right]$$

3.12 BLUE Property :

Under the classical assumptions of the classical linear regression model, the least square estimates prossess some ideal or optimum properties. The best linear unbiased properties. The best linear unbiased properties of an estimator (BLUE) are.

(i) It is linear ie a linear function of a random variable. Such as the dependent variable in the regression model.

(ii) It is unbiased ie the average or expected value of the estimation is equal to the true value $\left(E\hat{\beta} = \beta\right)$

(iii) It has minimum variance in the class of all such linear unbiasd estimators. An unbiased estimater with the least varience is known as effecint estimaton.

In regression context it can be proved that the OLS estimators are best linear unbiased estimator (BLUE).

BLUE Property of $\hat{\beta}$:

Under the standard OLS assumption, $\hat{\beta}$, the OLS estimation of β is best linear unbiased estimator. It is best in the sense that among all linear unbiased estimator of $\beta, \hat{\beta}$ has the smallest variance and hence $\hat{\beta}$ is the most effecient.

To prove this let us begin with an arbitary linear estimator b of β such that

$$b = \Sigma C_t Y_t \quad \ldots\ldots\ldots\ldots\ldots (3.5.7)$$

We know that $\hat{\beta} = \Sigma W_t Y_t$

Denoting the difference between C_t and W_t by d_t, we have.

$$c_t = w_t + d_t \quad \ldots\ldots\ldots\ldots\ldots (3.5.8)$$

Now expanding b we have

$$b = \Sigma C_t(\alpha + \beta X_t + U_t)$$

$$\Rightarrow b = \alpha\Sigma C_t + \beta\Sigma C_t X_t + \Sigma C_t U_t \quad \ldots\ldots\ldots (3.5.9)$$

$$Eb = \alpha\Sigma C_t + \beta\Sigma C_t X_t + \Sigma C_t EU_t$$

$$\Rightarrow Eb = \alpha\Sigma C_t + \beta\Sigma C_t X_t \quad \ldots\ldots\ldots (3.5.10)$$

$$\because E(U_t) = 0 \text{ by assumption 1.}$$

For b to be unbiased ie for Eb to be β, we require

$\Sigma C_t = 0$ and $\Sigma C_t X_t = 1$

ie $\Sigma w_t + \Sigma d_t = 0$ and $\Sigma d_t X_t = 0$

$[\because \Sigma W_t = 0$ and $\Sigma W_t X_t = 1$ by relations (3.5.1) and (3.5.2) $]$

Now variance of b is given by

$$\sigma^2 b = E(b - Eb)^2$$

Substituting equations (3.5.9) and (3.5.10) we have,

$$\sigma^2 b = E(\Sigma C_t - U_t)^2$$

$$= E\left[(c_1 u_1 + \ldots\ldots\ldots + C_n U_n)(c_1 u_1 + \ldots\ldots\ldots + C_n U_n)\right]$$

$$= E\left(\sum_t C_t^2 U_t^2 + \sum_{t \neq s} C_t C_s U_t U_s\right) = \sigma^2 \Sigma C_t^2 \text{ as in case of } \sigma^2\hat{\beta}.$$

or $\sigma^2\hat{\beta} = \sigma^2\Sigma(w_t + d_t)^2 = \sigma^2\Sigma w_t^{\,2} + \sigma^2\Sigma d_t^{\,2} + 2\sigma^2\Sigma w_t d_t$

Now $\Sigma W_t d_t = \frac{1}{\Sigma x_t^{\,2}}\Sigma x_t d_t \qquad \therefore W_t = \frac{x_t}{\Sigma x_t^{\,2}}$

$$= \frac{1}{\Sigma x_t^{\,2}}(X_t - \overline{X})d_t = \frac{1}{\Sigma x_t^{\,2}}(\Sigma X_t d_t - \overline{X}\Sigma d_t) = 0$$

$\because$ (Unbiasedness of b implies that$\Sigma X_t d_t = 0 = \Sigma d_t$)

Thus,

$\sigma^2 b = \sigma^2\Sigma W_t^{\,2} + \sigma^2\Sigma d_t^{\,2} \Rightarrow \sigma^2\hat{b} = \sigma^2\hat{\beta} + \sigma^2\Sigma d_t^{\,2}$

σ^2 is always positive and $\Sigma d_t^{\,2}$ being sum of squared terms is always non negative ie

$\sigma^2\Sigma d_t^{\,2} \geq 0 \qquad \therefore \sigma^2 b \geq \sigma^2\hat{\beta}$

Equality will hold when $\sigma^2\Sigma d_t^{\,2} = 0$, ie when $\Sigma d_t^{\,2} = 0$. This will be the case when d_t are all individually equal to zero or when $C_t = W_t$ or $\hat{b} = \hat{\beta}$

Thus, it is proved that $\hat{\beta}$ is the linear unbiased estimator of β.

Properties of $\hat{\alpha}$:

From the two variable linear regression model $Y_t = \alpha + \beta x_t + u_t$. we have the estimator of α as

$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$

Substituting $\Sigma W_t Y_t$ for $\hat{\beta}$ we have

$$\hat{\alpha} = \frac{1}{n}\Sigma Y_t - \overline{X}\Sigma W_t Y_t \;= \Sigma\left(\frac{1}{n} - \overline{X}W_t\right)Y_t$$

Denoting $\left(\frac{1}{n} - \overline{X}W_t\right)$ by Z_t we have,

$$\hat{\alpha} = Z_t Y_t$$

This shows that like $\hat{\beta}, \hat{\alpha}$ is also linear in Y_t

Here $\Sigma Z_t = 1$ and $\Sigma Z_t X_t = 0$

Now substituting $Y_t = \alpha + \beta x_t + u_t$ in $\hat{\alpha}$ we have

$\hat{\alpha} = \Sigma Z_t(\alpha + \beta x_t + u_t) = \alpha\Sigma Z_t + \beta\Sigma Z_t X_t + \Sigma Z_t X_t = \alpha + \Sigma Z_t U_t$

Using the OLS standard assumptions and above results, it can be proved that

$$E\hat{\alpha} = \alpha \quad \text{and} \quad \sigma^2\hat{\alpha} = E(\Sigma Z_t U_t)^2 = \sigma^2 \Sigma Z_t{}^2 = \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\Sigma x_t{}^2}\right]$$

Denoting $\alpha = \Sigma Y_t Y_t$ as an arbitrary estimator of α, a proof can be developed in similar line as in the case of $\hat{\beta}$ to show that $\hat{\alpha}$ also happens to be the BLUE of α.

The fact that under the standard assumption OLS estimators are the best linear unbiased estimators is known in econometric literative as the Gauss Markov Theorem.

Finally, another result which may be of use in the covariance between $\hat{\alpha}$ and $\hat{\beta}$. By definition covariance

$$\left(\hat{\alpha},\hat{\beta}\right) = \sigma\hat{\alpha}\hat{\beta} = E\left\{(\hat{\alpha} - E\hat{\alpha})\left(\hat{\beta} - E\hat{\beta}\right)\right\}$$

$$= E\left\{(\hat{\alpha} - E\hat{\alpha})\left(\hat{\beta} - E\hat{\beta}\right)\right\} \quad \because E\hat{\beta} = \beta \quad \ldots\ldots\ldots (3.5.11)$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} \Rightarrow E\hat{\alpha} = \overline{Y} - \beta\overline{X}$$

Now, $\hat{\alpha} - E\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} - \overline{Y} + \beta\overline{X} = -\overline{X}\left(\hat{\beta} - \beta\right)$

Substituting the above result in (3.5.11) we have

$$E\left\{-\overline{X}\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)\right\} = -\overline{X}E\left(\hat{\beta} - \beta\right)^2 = -\overline{X}\,\mathrm{Var}\,\hat{\beta} = -\overline{X}\frac{\sigma^2}{\Sigma x_t{}^2}$$

Estimation of σ^2 :

The expressions for variances of $\hat{\alpha}$ and $\hat{\beta}$ and the covariance involve the unknown prameter σ^2.

Therefore to estimate variance and covariance, we have to σ^2 before hand. σ^2 is the variance of the distribution of the random term u_t. As different values of U_ts are not observable we can not get a sample of values of u_t which could have been used to estimate the population variance of u_t. But we can consider the OLS residuals $\hat{u}_t$ as proxy observations on u_ts. Then we can attempt to obtain an estimator of σ^2 on the basis of these proxy observations $\hat{u}_t$s.

To start with, reall that

$$\hat{U}_t = Y_t - \overline{Y}_t \Rightarrow \hat{u}_t = \alpha + \beta x_t + u_t - \left(\hat{\alpha} + \hat{\beta} x_t\right)$$

$$\Rightarrow \hat{u}_t = u_t - (\hat{\alpha} - \alpha) - (\hat{\beta} - \beta) x_t$$

The actual derivation process is a little tedious; but starting from the above expression and using various result already derived it can be shown that

$$E\Sigma\hat{u}_t^2 = (n-2)\sigma^2$$

In other words, $E\dfrac{\Sigma\hat{u}_t^2}{n-2} = \sigma^2$

Thus, an unbiased estimation of σ^2 may be defined as

$$\hat{\sigma}^2 = \frac{\Sigma\hat{u}_t^2}{n-2} = \frac{\Sigma\left(Y_t - \hat{\alpha} - \hat{\beta}X_t\right)^2}{n-2}$$

3.13 General Linear Regression Model :

OLS assigns equal wights or importance to each observation. Hence OLS estimator may not be the best estimator though it is unbiased. GLS (Generalised Least Squares) takes such information into account explicitly and is therefore capable of producing estimators that are BLUE. In short, GSL is OLS on the transformed variables that satisfy the standard least squares assumptions.

Let, there are k explanatory variables in the model. Then the model will be—

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 + \beta_3 X_{3t} + \dots\dots\dots + \beta_k + \beta_k X_{kt} + u_t \quad \dots\dots\dots 1$$

where t = 1, 2, ... n and n>k.

There are K parameters to be estimated (K = k +1). System of normal equations will consists of K equations, in which the unknowns are the parameters $\beta_1, \beta_2, \beta_3 \dots\dots\dots \beta_k$ and the known terms will be the sums of squares and the sums of products of all the variables in the structural equation.

For t =1,

$$Y_1 = \beta_1 + \beta_2 X_{21} + \beta_3 + \beta_3 X_{31} + \dots\dots\dots + \beta_k + \beta_k X_{k1} + u_1$$

For t = 2,

$$Y_2 = \beta_1 + \beta_2 X_{22} + \beta_3 + \beta_3 X_{32} + \dots\dots\dots + \beta_k + \beta_k X_{k2} + u_2$$

For t = n.

$$Y_n = \beta_1 + \beta_2 X_{2n} + \beta_3 + \beta_3 X_{3n} + \dots\dots\dots + \beta_k + \beta_k X_{kn} + u_3$$

In matrix form,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_1 + \beta_2 X_{21} + \beta_3 + \beta_3 X_{31} + \dots\dots\dots + \beta_k + \beta_k X_{k1} + u_1 \\ \beta_1 + \beta_2 X_{22} + \beta_3 + \beta_3 X_{32} + \dots\dots\dots + \beta_k + \beta_k X_{k2} + u_2 \\ \vdots \quad \vdots \quad \vdots \\ \beta_1 + \beta_2 X_{2n} + \beta_3 + \beta_3 X_{3n} + \dots\dots\dots + \beta_k + \beta_k X_{kn} + u \end{bmatrix}$$

$$\text{or} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_1 + \beta_2 X_{21} + \beta_3 + \beta_3 X_{31} + \dots\dots\dots + \beta_k + \beta_k X_{k1} \\ \beta_1 + \beta_2 X_{22} + \beta_3 + \beta_3 X_{32} + \dots\dots\dots + \beta_k + \beta_k X_{k2} \\ \vdots \\ \beta_1 + \beta_2 X_{2n} + \beta_3 + \beta_3 X_{3n} + \dots\dots\dots + \beta_k + \beta_k X_{kn} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

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

Thus. in matrix form we can write equation 1

$$Y = X\beta + u \quad \dots\dots\dots\dots 2$$

3.14 Assumptions of Generalised Least Square :

(i) or $E(U) = 0;\quad E(U) = \begin{bmatrix} E(u_1) \\ E(u_2) \\ \vdots \\ E(u_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$

The expectation of a vector/matrix is the expection of each element of the vector/matrix.

(ii) $E(UU') = \sigma^2 I_n$

$$E(UU') = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \begin{bmatrix} u_1 u_2 \ldots\ldots u_n \end{bmatrix} = \begin{bmatrix} u_1^2 & u_1u_2 & \ldots\ldots & u_1u_n \\ u_1u_2 & u_2^2 & \ldots\ldots & u_2u_n \\ & & \vdots & \\ u_nu_1 & u_nu_1 & \ldots\ldots & u_n^2 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & 0 \ldots\ldots & 0 \\ 0 & \sigma^2 \ldots\ldots & 0 \\ & \vdots & \\ 0 & 0 \ldots\ldots & \sigma^2 \end{bmatrix} \left(\because E\left(u_t^2 = \sigma^2\right) \text{ and } E\left(u_t u_s\right)_{t \neq s} = 0\right)$$

This is a double assumption, namely—

- Each u disturbance has the same variance.
- All disturbances are pair wise uncorrelated.

$$E\left(UU'\right) = \begin{bmatrix} \sigma^2 & 0 \ldots\ldots & 0 \\ 0 & \sigma^2 \ldots\ldots & 0 \\ & \vdots & \\ 0 & 0 \ldots\ldots & \sigma^2 \end{bmatrix} \begin{bmatrix} 1 & 0 \ldots\ldots & 0 \\ 0 & 1 \ldots\ldots & 0 \\ & \vdots & \\ 0 & 0 \ldots\ldots & 1 \end{bmatrix} = \sigma^2 I_n.$$

(iii) $U \sim N\left(0, \sigma^2 I_n\right)$ — the disturbances (U's) are normally distributed with each and every mean is equal to zero and variance is $0, \sigma^2 I_n$.

(iv) X is a non stochastic matrix and has full column rank P (x) = k – – rank of matrix X must be equal to K. This assumption states that the explanatory variables donot form a linearly independent set.

· Let us Understand,

Rank of a matrix is the number of linearly independent element in rows or columns of the matrix.

3.15 The Coeffecient of Determenation :

The coeffecient of determination tells how well the sample regression line fits the data. The coefferiant of determination or goodness of fit is the square of the correlation of coeffecient or R². This shows the percentage of the total variation in the dependent variable which can be explained by the independent variable.

Let us consider the simple linear regression of Y on X.

$Y_t = \alpha + \beta x_t + U_t$ 3.6.1

Total variation in Y_t; $\Sigma y^2 = \Sigma(Y_t - \overline{Y})^2$

Explained variation $= \Sigma\hat{y}_t^{\,2} = \Sigma(Y_t - \overline{Y})^2$

Unexplained variation $= \Sigma\hat{y}_t^{\,2} = \Sigma(Y_t - \overline{Y})^2$

From the residual on the error term we have

$\hat{U}_t = Y_t - \hat{Y}_t \Rightarrow Y_t = \hat{Y}_t + \hat{u}_t$ 3.6.2

Summing over t we have

$\sum_t Y_t = \sum_t \hat{Y}_t + \sum_t \hat{u}_t$ or $\Sigma Y_t = \Sigma\hat{Y}_t$

$\because\ \Sigma\hat{u}_t = 0$ by the first normal equation

Deviding both sides by n, we have

$$\overline{Y} = \overline{\hat{Y}} \quad .... 3.6.3$$

Substracting (3.6.3) from (3.6.2) we have,

$\left(Y_t - \overline{Y}\right) = \left(\hat{Y}_t - \overline{\hat{Y}}\right) + \hat{U}_t$ or $Y_t = \hat{Y}_t + \hat{U}_t$ 3.6.4

Squaring (3.6.4) on both sides and summing over the sample we have,

$\Sigma y_t^{\,2} = \Sigma\hat{y}_t^{\,2} + \Sigma\hat{u}_t^{\,2} + 2\Sigma\hat{y}_t u_t$

Now $\Sigma\hat{y}_t\hat{U}_t = \Sigma\left(\hat{Y}_t + \overline{\hat{Y}}\right)\hat{u}_t = \Sigma\hat{Y}_t\hat{u}_t - \overline{\hat{Y}}_t\sum\hat{u}_t = \Sigma\hat{Y}_t\hat{u}_t$

$\left(\because\ \Sigma\hat{u}_t = 0 \text{ by the first normal equation}\right)$

Substituting $Y_t = \hat{\alpha} + \hat{\beta}X_t$

$= \Sigma\left(\hat{\alpha} + \hat{\beta}X_t\right)\hat{u}_t = \hat{\alpha}\Sigma\hat{u}_t + \hat{\beta}\Sigma X_t\hat{u}_t = 0$

$\because\ \Sigma\hat{u}_t = 0 = \Sigma\hat{y}_t^{\,2} + \Sigma\hat{u}_t^{\,2}$ by I and II normal equation.

$\Sigma y_t^{\,2}$ represent the total variation of the actual Y values about ther sample mean. The term is called the total sum of squares (TSS). $\Sigma\hat{y}_t^{\,2}$ which is called the explained sum of squares (ESS) is the variation of estimated Y values about their mean $\Sigma\hat{y}_t$ tells how much variation in Y has been explained by the filted regression line $\Sigma\hat{u}_t^{\,2}$ or residual sum of squares (RSS) is the variation in Y which remains unexplained.

Deviding both sides by the total sum of squares $\Sigma y_t^{\,2}$, we have

$$1 = \frac{\Sigma\hat{y}^2}{\Sigma y_t^2} + \frac{\Sigma\hat{u}_t^2}{\Sigma y_t^2} \quad \text{or} \quad \frac{\Sigma\hat{y}_t^2}{\Sigma y_t^2} = 1 - \frac{\Sigma u_t}{\Sigma y_t}$$

The above expression represents the proportion of total variation in Y which is explained by the model. The ratio is called the coeffecient determination and is denoted by R^2.

$$\text{Thus, } R^2 = \frac{\Sigma\hat{y}_t^2}{\Sigma y_t^2} = 1 - \frac{\Sigma u_t^2}{\Sigma y_t^2}$$

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

The coeffecient of determination usually lies in the range of zero to unity. Higher the value of R^2 in the range, greater is the extent of variation explained by the fitted regression model. On the other hand, smaller value of R^2 implies weak explanatory power of the model.

It may be noted that R^2 happens to be equal to the square of simple correlation coefficient between observed and estimated Y values.

ie $R^2 = \sigma_{y\hat{y}}^2$

Again, for the two variable linear regression model R^2 is also equal to the square of the simple correlation co-effecient between the two variables.

ie $R^2 = \sigma^2_{xy}$.

3.16 Maximum Likelihood Methods :

This method is based on the idea that different population generate different samples and that any given sample is more likely to have come from same population than others.

Here, main objective to determine the proceedure by which we can compute the various unknown parameters of a given population on the basis of sample observations. These parameters can be mean, varience, slope and intercept. To estimate the parameters, likelihood function should be determined for the observation in sample and then maximises it with respect to the unknown parameters.

The sample in Maximum Likelihood Estimator (MLE) can be generated by various alternative population having different parameter values. Likelihood ie probability of different population are different. The parameter value of that population which has maximum likelihood of generating sample is known as likelihood estimates.

Let us consider the two variable linear regression model—

$$Y_t = \alpha + \beta X_t + u_t$$

We assume that the disturbance term or the random variable is distributed normally with zero mean and constant variance σ^2. ie $U \sim N(0, \sigma^2)$. Since u_t is assumed to be normally distributed.

Here, first we obtain the likelihood function. The joint probability density function (p.d.f) of Y, is called likelihood function (L).

$\therefore L = P(Y_1, Y_2 \ldots\ldots Y_n) = P(Y_1), P(Y_2), \ldots\ldots\ldots P(Y_n)$

Now

$Y_t = \alpha + \beta x_t + u_t$

$E(Y_t) = \alpha + \hat{\beta} X_t \qquad [\because E(u_t) = 0]$

$\therefore \text{Var}(Y_t) = E[Y_t - E(Y_t)]^2 = E[\alpha + \beta X_t + u_t - (\alpha + \beta X_t)]^2$

$$= E[\alpha + \beta X_t + u_t - \alpha - \beta X_t]^2 = E(u_t^2)$$

$\text{Var}(Y_t) = \sigma^2$

$\therefore Y_t \sim N[(\alpha + \beta X_t), \sigma^2]$

$$\therefore L = \prod_{t=1}^{n} P(Y_t) = \prod_{t=1}^{n} \frac{1}{\sqrt{\sigma^2 2\pi}} e^{-\frac{1}{2}\left(\frac{Y_t - \alpha - \beta X_t}{\sigma}\right)^2}$$

$$= \prod_{t=1}^{n} (\sigma^2)^{-n/2} (2\pi)^{-n/2} e^{-1/2\sigma^2 (Y_t - \alpha - \beta X_t)^2}$$

$$= (\sigma^2)^{-n/2} (2\pi)^{-n/2} e^{-1/2\sigma^2 \sum_{t=1}^{n} (Y_t - \alpha - \beta X_t)^2}$$

Taking logarithm

$$\ell = \ell nL = -n/2\, \ell n 2\pi - n/2\, \ell n\, \sigma^2 - 1/2\sigma^2 \sum_{t=1}^{n} (Y_t - \alpha - \beta X_t)^2$$

Which is the required likelihood function. We now maximise $\ell = \ell nL$ with respect to α and β and σ^2.

The justification of using logarithm is that since logarithm is a positive monotonic transformation, the parameter values that would maximise the likelihood function would also maximise the log of the likelihood function. Thus, the maximum likelihood estimator can be derived by maximising likelihood estimate $\ell = \ell nL$.

Let us understand,

$$\hat{\alpha}_{OLS} = \tilde{\alpha}_{MLE} \quad \text{and} \quad \hat{\beta}_{OLS} = \tilde{\beta}_{MLE}$$

$$\text{But, } \sigma^2{}_{OLS} \neq \sigma^2{}_{MLE}$$

3.17 Properties of Maximum Likelihood Estimators :

(i) Maximum likelihood estimators donot satisfy all the small sample properties of unbiased always.

(ii) The ML estimators are consistent.

(iii) The ML estimators are assymptotically normally distributed.

(iv) The ML estimators are asymptotically effecient among all the consistent estimators having smallest asymptotic variance.

(v) The ML estimate of a function $g(\theta)$ is $g(\hat{\theta})$ where $\hat{\theta}$ is the ML estimate of θ.

Let us understand,

Likelihood ratio is a test of maximum likelihood estimate. The likelihood ratio is the ratio of the restricted likelihood function to the unresticeted likelihood function.

3.18 Exercise :

1. State and prove the Gauss Markor theorem for OLS. Decompose the total variations in the dependent variable of a linear regression model and obtain the expression for the coeffecient of determination.

(G.U. Previous, '06)

2. What do you understand by BLUE property and regression parameter? Given the regression model $Y_t = \beta_t + \beta_2 x_t + U_t$ show that the least square parameters are best linear unbiased estimator under standard assumptions.

(G.U. MA Previous, '03, 04)

3. Derive OLS estimators of the parameters in the general linear regression equation. Prove that the estimators indeed minimise the sum of square of the residuals.

(G.U. MA Previous '05)

4. Why do you have a random variable in a regression model? What is the role in estimating regression parameters of a linear regression model?

(G.U. MA Privious '04)

5. When there is mis specification of explanatory variables of a linear regression model, show that the expected value of OLS estimator of the regression coeffecients is a linear combination of the true coeffecients.

(G.U. MA Previous '07)

6. Write down the general linear regression model and derive the OLS estimators of the vector of regression coeffecients. Prove that the estimators indeed maximise the sum of squares of residuals.

(G.U. MA Previous '08)

7. Outline the principle of maximum likelihood method of estimation. Why is it justified to obtain the estimators by maximising the logarithm of the likelihood function rather than the likelihood function itself?

(G.U. MA Previous '078)

8. Outline the principle of maximum likelihood estimation obtain the ML estimators of the parameters of the linear regression model. Explain the concept of likelihood ratio and give interpretation of its value.

(G.U. MA Previous '07)

9. How doyou measure the 'goodness' of fit of a regression model? Briefly explain.

(G.U. MA Previous '04)

10. Justify the presence of the random disturbance term in the regression model. Trace the role of standard assumptions about the disturbance term in characterisetion of estimation and setting up of the inference Proceedure.

(G.U. MA Previous '08)

11. What is meant when disturbance in a linear regression model are said to be identically and independently distributed. What further assumptions would you require to carry out inference from the OLS estimator.

(G.U. MA. Previous '05)

●●●

UNIT-4

INFERENCE FROM LINEAR REGRESSION ESTIMATION

Contents :

4.1 Introduction :

Simple Calculation of statistical data donot give a meaningful result. For this inference of calculated data should be made first. From inference of the data a layman is also able to understand the meaning and prediction of the statistical outcomes.

4.2 Objectives :

This unit is designed to help you understand the concept of inference from linear regression estimation and its related ideas. After reading this unit you will be able to,

- Test hypothesis about regression coefficients.
- Formulate confidence interval.
- Predict with linear regression model.
- Distinguish between point and interval production.

4.3 Test of Hypothesis about Regression Coefficients :

After obtaining the estimates of parameters, inference about the unknown parameter values should be made. To carry out the inference process first of all, the hypothesis about the parameter values should be formulated. The null hypothesis determines what type of inference should be made. After that, a particular test statistic should be find out. For setting up the test procedure it is necessary to ascertain the nature of the sampling distribution of the estimator.

Let us consider the sampling distribution of $\hat{\beta}$ Here,

$$\hat{\beta} = \sum W_t Y_t = \beta + \sum W_t u_t$$

This shows $\hat{\beta}$ as a linear combination of $U_1, U_2, \ldots\ldots U_n$. U_ts are normally distributed. $\hat{\beta}$ being a linear combination of U_t, $\hat{\beta}$ itself is also normally distributed. Since mean and variance of $\hat{\beta}$ are β and $\frac{\sigma^2}{\sum x_t^2}$ respectively. We have,

$$\hat{\beta} \approx N\left(\beta, \frac{\sigma^2}{\sum x_t^2}\right) \text{accordingly,} \quad \frac{\hat{\beta}-\beta}{\sigma\Big|\sqrt{\sum x_t^2}} \approx N(0, 1)$$

Thus,

$Z = \dfrac{\hat{\beta}-\beta}{\sigma\Big|\sqrt{\sum x_t^2}}$ could have been used as a test statistic for testing hypothesis about the value of the parameter β. But Z involves unknown parameter σ and hence cannot be calculated. Therefore, we have to substituted σ by its estimate $\hat{\sigma}$. Once we substitute σ by $\hat{\sigma}$, the above ratio will no longer remain a standard normal variate and conform to student t-distribution. Thus, $\dfrac{\hat{\beta}-\beta}{\sigma\Big|\sqrt{\sum x_t^2}} \sim t$ with (n – 2) degrees of freedom.

We know that, $\hat{\sigma}^2 = \dfrac{\sum \hat{U}_t^2}{n-2}$ i.e. $\hat{\sigma} = \sqrt{\dfrac{\sum\left(Y_t - \hat{\alpha} - \hat{\beta} x_t\right)^2}{n-2}}$

$\frac{\hat{\sigma}}{\sqrt{\Sigma x_t^2}}$ is the estimated standard deviation or standard error of $\hat{\beta}$.

Thus the above ratio can be re-written as,

$$t = \frac{\hat{\beta} - \beta}{SE(\hat{\beta})} = \frac{\text{estimator - parameter}}{\text{estimated standard error of estimator}}$$

Now for testing the null hypothesis H_0 ($\beta = b_0$) against the alternative hypothesis H_A $(\beta \neq b_0)$, the test statistic can be set up as $t = \frac{\hat{\beta} - b_0}{SE(\hat{\beta})}$

After calculating the test statistic, we shall compare the calculated value of the test statistic with its tabulated value. If at a given level of significance, the calculated value exceeds the tabulated value, we shall reject the null hypothesis. On the other hand, if the calculated value does not exceed the tabulated value, We shall accept the null hypothesis.

A hypothesis which is most commonly tested in econometrics is H_0 (β = b_0) against the alternative H_A $(\beta \neq b_0)$. The test statistic for the purpose is $t = \frac{\hat{\beta}}{SE\hat{\beta}}$

If H_0 ($\beta = 0$) is rejected the corresponding explanatory variable X is said to be significant. When X is significant, it implies that the variable X has some significant impact on the dependent variable Y. On the other hand, if we accept the null hypothesis, X will be significant. This means X has no significant effect on Y. When H_0 ($\beta = 0$) is rejected at 0.1 level of significicance, X is said to be highly significant.

A test procedure for hypothesis about the value of $\propto$ can be set up in a similar way. For instance, test the null hypothesis H_0 ($\propto = 0$) against the alternative H_A ($\propto \neq 0$), the test statistic would be $t = \frac{\hat{\propto}}{SE\,\hat{\propto}}$ with (n – 2) degrees of freedom.

4.4 Testing of overall significance of a multiple regression model :

Let us consider the following K variable linear regression model :

$$\gamma_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \ldots\ldots + \beta_k x_{kt} + U_k$$

Let H_0 be the null hypothesis that there is no significant difference between the slope coeffecientsand they are equal to zero.

Assuming H_0 to be true, the test statististic is given by

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

If the calculated value of F exceeds the tabulated value we reject the null hypothesis and accept the alternative hypothesis. This means that there is significant difference between the slope coefficients and they are not equal to zero.

Against, F test can be expressed in terms of R^2 as follows,

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

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

$$= \frac{(n-k)}{(n-1)} \times \frac{ESS/TSS}{1 - \frac{ESS}{TSS}}$$

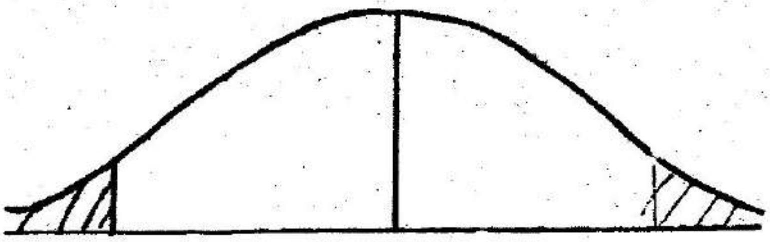
$$= \frac{(n-k)}{(n-1)} \times \frac{R^2}{(1 - R^2)} \quad \{\because R^2 = ESS/TSS\}$$

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

This means that F and R^2 are related to each other. When $R^2 = 0$ then F is also zero and larger the value of R^2, larger will be the value of F. When $R^2 = 1$, the value of F is infinity. Thus, the F test which is a measure of overall significance of the estimated regression is also a test of significance of R^2.

4.5 Confidence Interval for Parameter Values :

Due to sampling fluctuation, a single estimate of a parameter is likely to differ from its true value. Hence instead of relying on the single estimate we may construct an interval around that estimator. The confidence interval is



defined as a range of value which has a specific probability of including the true parameter value within its limit. If the sampling distribution of the estimator is known, one can easily fix the confidence interval for a parameter. Let us first fix a confidence interval for b.

In the previous section, we understand that when σ is substituted by $\hat{\sigma}$ in standard normal variate it conforms to the student t distribution. Accordingly we get,

$$\frac{\hat{\beta} - \beta}{SE\left(\hat{\beta}\right)} \sim t \text{ with (n-2) degrees of fredom.}$$

We know that the total area under the t-curve is one. Let us define $t_{.025}$ be the value of t to the right of which .025 of are under the t curve lies. Since 't' distribution is symmetrieal about zero, — $t_{.025}$ will be the value of t to left of which also .025 of area under 't' curve will lie.

Thus, the are under the curve between $-t_{.025}$ and $t_{.025}$ will be 1– .025 – .025 or .95. This implies,

$$\Pr\left\{-t_{.025} < t < t_{.025}\right\} = .95 \qquad \because \frac{\hat{\beta} - \beta}{SE\left(\hat{\beta}\right)} \sim t$$

We have. $\Pr\left\{-t_{.025} < \frac{\hat{\beta} - \beta}{SE\hat{\beta}} < t_{.025}\right\} = .95$

Now, $-t_{.025} < \frac{\hat{\beta} - \beta}{SE\,\hat{\beta}}$ implies $\hat{\beta} - \beta > -t_{.025}\,SE\hat{\beta}$

$$\Rightarrow \beta < \hat{\beta} + t_{.025}\,SE\hat{\beta}$$

and $\frac{\hat{\beta}-\beta}{SE\hat{\beta}} < t_{.025}$ implies $\hat{\beta}-\beta < t_{.025}\, SE\hat{\beta}$

$$\Rightarrow \hat{\beta} > \beta < t_{.025}\, SE\hat{\beta}$$

$\therefore$ we have $\Pr\left\{\hat{\beta} - t_{.025}\, SE\hat{\beta} < \beta < \hat{\beta} + t_{.025}\, SE\hat{\beta}\right\} = .95$

This is called the 95% confidence interval of β, because in 95 out of 100 causes the true parameter value of β will lie in interval so determined.

Similarly, intervals for alternative confidence levels can be set up. Obviously the interval will be wider for a higher confidence level (say 99%) but narrower for a lower confidence level (say 90%). Moreover, confidence intervals for other parameter $\propto$ can also be set up following a similar procedure.

Sometimes, one needs to construct a joint confidence interval for β_1 and β_2 such that with a confidence coefficient $(1-\alpha)$, say 95%, that interval includes β_1 and β_2 simultaneously.

Again, χ^2 (chi-square) distribution is used to establish confidence interval for σ^2 ie.

$$\Pr\left\{\chi^2_{1-\alpha/2} \le \chi^2 \le \chi^2_{\alpha/2}\right\} = 1-\alpha$$

4.6 Prediction with the Linear Regression Model :

Making prediction is one of the objectives of econometric research. Using the estimate of the dependent variable in an unobserved situation. When such prediction is made in a future date, it is called a forecast.

While making prediction, it is assumed that the prediction unit and the sample unit belong to the same population. In case of a forecast, based on model estimated from time series data, this assumption means that there is no structural change between the sample period and the forecast period. If this assumption is not satisfied we will get different parameter values (such as α, β in the two variable linear regression model) for the sample period as well as for the prediction period. In that case, prediction made on the basis of estimates of parameters from the sample will no longer be valid.

Let, the two variable linear regression model for the purpose of making prediction is

$$Y_t = \alpha+\beta X_t + U_t$$

This satisfies all the standard OLS assumptions. Let us assume that the estimates $\hat{\alpha}$ and $\hat{\beta}$ are based on n observations denoted by

$$(X_1, Y_1), (X_2, Y_2) .... (X_n, Y_n).$$

Now, we have to predict the value of Y for the unit P which lies outside the sample of n observations.

Suppose the value of X for the unit P, ie X_p can be either known or projected by following some well defined criterion. Given the values of X_p, $\hat{\alpha}$ and $\hat{\beta}$, the value of Y for the unit P, ie, the point prediction of Y_p is merely its estimated value based on these three values. Thus the point prediction is

$$\hat{Y}_p = \hat{\alpha}+\hat{\beta}X_p$$

This is called point prediction since $\hat{Y}_p$ is a specific value for the given values of $\hat{\alpha}$, $\hat{\beta}$ and X_p. Since $\hat{Y}_p$ is only a specific estimated value for some given values, it is likely to be different from its true value. The actual value of Y_p will be

$$Y_p = \alpha+\beta X_p + U_p$$

The difference between the actual and predicted value of Y_p is said to be the prediction error. Thus the prediction error (e) is

$$\ell = Y_p - \hat{Y}_p = \alpha+\beta X_p + U_p - \hat{\alpha} - \hat{\beta}X_p$$

$$= U_p - (\hat{\alpha} - \alpha) - X_p\left(\hat{\beta} - \beta\right)$$

Taking expectation of the prediction error, we have

$$E_e = EU_p - E(\hat{\alpha} - \alpha) - X_p E\left(\hat{\beta} - \beta\right)$$

$$= EU_p - (E\hat{\alpha} - \alpha) - X_p\left(E\hat{\beta} - \beta\right)$$

$$= 0 - (\alpha - \alpha) - X_p\left(\hat{\beta} - \beta\right)$$

$$= 0 \qquad \left[\because EU_p = 0 \text{ by assumption}\right]$$

Of the model and $\hat{\alpha}$ and $\hat{\beta}$ are unbiased estimation of α and β under the standard assumptions. Then, if the assumptions of the model and the assumption underlying the prediction exercise are satisfied, the expected value of the predicted error is zero and the prediction is unbiased.

Now variance of the prediction error is

$$\sigma^2_e = E(e - Ee)^2 = E(e - 0)^2 = E.e^2$$

$$= EU^2_p + E(\hat{\alpha} - \alpha)^2 + X^2_p E(\hat{\beta} - \beta)^2$$

$$+ 2X_p E(\hat{\alpha} - \alpha)(\hat{\beta} - \beta) - 2EU_p(\hat{\alpha} - \alpha) - 2X_p EU_p(\hat{\beta} - \beta)$$

$$= \sigma^2 + \sigma^2_{\hat{\alpha}} + X^2_p \sigma^2_{\hat{\beta}} + 2X_p \sigma_{\hat{\alpha}\hat{\beta}} + 0$$

$\therefore EU_p^{\,2} = \sigma^2$ and U_p will be independent of $\hat{\alpha}$ and $\hat{\beta}$.

Under the standard OLS assumption.

Now,

$$\sigma^2_e = \sigma^2 + \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum x^2_t}\right) + X^2_p \frac{\sigma^2}{\sum x^2_t} - 2X_p \frac{\overline{X}}{\sum x^2_t} \sigma^2$$

$$= \sigma^2 \left(1 + \frac{1}{n} + \frac{\overline{X}^2}{\sum x^2_t} + \frac{X^2_p}{\sum x^2_t} - 2X_p \frac{\overline{X}}{\sum x^2_t}\right)$$

$$= \sigma^2 \left[1 + \frac{1}{n} + \frac{(X^2_p + \overline{X}^2 - 2X_p\overline{X})}{\sum x^2_t}\right]$$

$$= \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_p + \overline{X})^2}{\sum x^2_t}\right]$$

From the above expression it is clear that the variance of prediction error (σ^2_e) is smallest when $X_p = \overline{X}$, ie the value of X for prediction unit is equal to the mean of sample. But as the gap between X_p and $\overline{X}$ widens, the variance starts increasing. Rising variance means that the point prediction becomes increasingly unreliable for units for which the difference between the value of independent variable for prediction unit and sample unit gets enlarged. Therefore it will be better to have an interval prediction.

To derive interval prediction, let us first mention that the prediction, let us first mention that the prediction error $e = U_p - (\hat{\alpha} - \alpha) - X_p(\hat{\beta} - \beta)$ is normally distributed since it is a linear combination of U_p, $\hat{\alpha}$ and $\hat{\beta}$ of all which are also normally distributed.

$$\therefore \frac{e - Ee}{\sqrt{\sigma^2_e}} \sim N(0,1)$$

i.e. $$\frac{Y_p - \hat{Y}_p}{\sigma\sqrt{\left[1+\frac{1}{n}+\frac{(X_p-\overline{X})^2}{\sum x^2_t}\right]}} \sim N(0,1)$$

$$\because Ee = 0 \quad \text{and} \quad e = Y_p - \hat{Y}_p$$

In the above expression σ is an unknown parameter. If we substitute σ by its estimate $\hat{\sigma}$, we shall have,

$$\frac{Y_p - \hat{Y}_p}{\hat{\sigma}\sqrt{\left[1+\frac{1}{n}+\frac{(X_p-\overline{X})^2}{\sum x^2_t}\right]}} \sim t \quad \text{with } (n-2) \text{ degree of freedom}$$

where $$\hat{\sigma} = \frac{\sum_{t=1}^{n}\left(Y_t - \hat{\alpha} - \hat{\beta}X_t\right)^2}{n-2}$$

Finally, the 95% confidence interval for Y_p will be,

$$\hat{Y}_p \pm t_{.025}\ \hat{\sigma}\sqrt{\left[1+\frac{1}{n}+\frac{(X_p-\overline{X})^2}{\sum x^2_t}\right]}$$

Thus, specifying such an interval for the value of Y for the observed situation P we have an interval prediction.

4.7 Distinction between Point Prediction and Interval Prediction :

There are some difference between the point prediction and interval prediction.

When a particular value of independent variable is used to estimate a dependent variable, it is known as point prediction. Here, prediction is made about the value of dependent variable for a particular future value of independent variable.

On the other hand, when an attempt is made to make some reasonable probability statement, the unknown dependent variable at a certain level of significance, then it is called interval prediction. Here, there are two limits under which predictors value of dependent variable will lie.

Let us understand,

(i) While making prediction on the basis of an estimated econometric model, it is assumed that the unit for which prediction is made and the sample on which the model is estimated belong to the same population.

(ii) In case of forecast based on estimated model estimated from time series data, this mean that there is no change between the forecast period and the period for which the model has been estimated.

4 . 8 Exercise :

1. Distinguish between the point prediction and interval prediction. Mention the crucial assumption underlying any prediction based on linear regression estimation.

(G.U. MA Previous '06)

2. Show how the estimate of linear regression model can be used to generate a point prediction. Analyses the quality of the prediction under the standard assumption. Extend your exercise to derive an interval prediction with 95% confidence interval.

(G.U, MA Previous '07)

3. You are given a two variable linear regression model : $Y_t = \alpha + \beta X_t + U_t$

Where α and β are parameters and 'U_t' is the stochastic variable with usual assumptions. what statistical test would you use in testing H_o : β =0 against $H_o : \beta \neq 0$.

●●●

UNIT-5

FURTHER TOPICS IN LINEAR REGRESSION MODEL

STRUCTURE

5.0 Introduction :

Multicollinearity implies a correlation between the some or all explanatory variables of a regression model. In this unit we will we able to study the effects, detection and remedies of Multicollinearity. Again a study of types of specification errors and their consequences will be made. Also, in this unit a brief note on dummy variables will be provided. Lastly, an introduction on the concepts of hetroscedasiticity and autocorrelation of disturbances will be made.

5.1 Objectives :

After going through this unit, one will be able to—

- Understand the concept of Multicollinearity, its effects, detection and remedies.
- Examine the types of specification errors and their consequences.
- Analyse the uses of qualitative factors or dummy variables.
- Get an idea about Hetroscedasiticity and autocorrelation of disturbances.

5.2. Multicollinearity :

The term Multicollinearity was coined by Ragner Frisch. Originally, it means the existence of a perfect or exact, linear relationship among some or all explanatory variables of a regression model.

However, the term Multicollinearity is used in a broader sense to include the case of perfect Multicollinearity, as well as the case where the x variables are Intercorelated but not perfectly so.

For the k-variable regression involving explanatory variable $x_1, x_2, ..., x_k$ (where $x_1 = 1$ for all observations to allow for the intercept term), an exact relationship is said to exist if–

$$\lambda_1 x_1 + \lambda_2 x_2 + ... + \lambda_k x_k = 0$$

where $\lambda_1, \lambda_2, ..., \lambda_k$ are constants such that not all of them are zero.

In case of less than perfect Multicollinearity, we have,

$$\lambda_1 x_1 + \lambda_2 x_2 + ... + \lambda_k x_k + v_i = 0,$$

where v_i is a statistic error term.

5.2.1 Effects of Multicollinearity :

Theoretical consequences of Multicollinearity

1. In case of near Multicollinearity the ordinary least square (OLS) estimators are unbiased.

2. Collinearity does not destroy the property of minimum variance.

3. Multicolleairity is essentially a sample (regression) phenomenon in the sense that even if the X variables are not linearly related in the population, they may be so related in the particular sample at hand.

Practical consequences of Multicollinearity

1. Although BLUE (Best Linear Unbiased Estimators), the OLS estimators have large variances and covariance, making precise estimation difficult

2. Also, the confidence intervals tend to be much wider, leading to the acceptance of the "Zero null hypothesis" more readily.

3. The 't' ratio of one or more coefficients tends to be statistically significant

4. Although the t ratio of one or more coefficients is statistically insignificant, R^2, the overall measure of goodness of fit, can be very high.

5. The OLS estimators and their standard errors can be sensitive to small change in the data.

5.2.2 Detection of Multicollinearity :

Although there is no unique method of detecting Multicollinearity, we have some rules of thumb for its detection. Following are some of there rules:

1. High R_2 but few significant t ratios—

If R^2 is high, the F test in most cases will reject the hypothesis that the partial slope coefficients are simultaneously equal to zero, but the individual 't' tests will show that none or very few of the partial slope coefficients are statistically different from zero.

2. High pair-wise correlation among regressions—

If the pair wise or zero order correlation coefficient between two regressions is high, then Multicollinearity is a serious problem.

3. Examination of Partial correlation–

In the regression of Y on x_2, x_3, x_4, in the model

$$y_i = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \mu_1,$$

a finding that $R^2_{1.234}$ is very high but $r^2_{12.34}$, $r^2_{13.24}$ and $r^2_{14.23}$ are comparatively low may imply that the variables x_2, x_3, and x_4 are highly intercorrelated.

4. Auxiliary regressions—

One way of finding out which x variable is related to other x variables is to regress each x_i on the remaining x variables and compute the corresponding R^2, known as R_i^2. Each one of these regression is called an auxiliary regression, auxiliary to the main regression of Y on the x's.

We now compute the 'F' test. If computed F excludes the critical F_i at chosen level of significance, it means that the particular X_i is collinear with other X's, and if it does not exceed the critical F_i, it is not collinear with the other X's. Also instead of testing auxiliary R^2 values, we may use Klien's rule of thumb, which says that multicollinearity may be a problem only if the R^2 obtained from an auxiliary regression is greater than the overall R^2, i.e. that obtained from the regression of Y on all the regressions.

5. Eigenvalues and condition index—

Eigenvalues and the condition index can be used to deteet Multicollinearity. From Eigen values, we can derive the condition number k defined as

$$k = \frac{\text{Maximum eigen value}}{\text{Minimum eigen value}}$$

and the condition index (C.I) defined as

$$CI = \sqrt{\frac{\text{Maximum eigen value}}{\text{Minimum eigen value}}} = \sqrt{k}$$

Now, if 100<k<1000, there is moderate to strong Multicollinearity and if it exceeds 1000 there is severe Multicollinearity. Again, if the CI i.e $10 < \sqrt{k} < 30$, there is moderate to strong Multicollinearity and if it exceeds 30 there is severe Multicollearity.

Thus above mentioned are some of the methods of detecting Multicollinearity.

5.2.3 Remedial Measures :

The solutions which may be adopted if Multicollinearity exists in a function depends on the severity of the collinearity problem. Some rules of thumb may be adopted as follows—

1. Apriority Information :

Suppose, we have the model

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

where Y = consumption X_2 = income and X_3 = wealth. Here income and wealth are highly collinear and suppose $\beta_3 = 0.10\beta_2$ i.e rate of change in consumption w.r.t wealth is $\frac{1}{10}$th the corresponding rate w.r.t income. We can now run the following regression—

$$y_i = \beta_1 + \beta_2 x_{2i} + 0.10\beta_2 x_{3i} + u_i$$

$$= \beta_1 + \beta_2 + u_i \quad \text{where} \quad x_i = x_{2i} + 0.10x_{3i}$$

Once we obtain $\hat{\beta}_2$ we can estimate $\hat{\beta}_3$ from the postulated relationship between $\hat{\beta}_2$ and $\hat{\beta}_3$ Apriority information would come from previous empirical work in which the collinearity problem happens to be less serious.

2. Combining cross sectional and time series data :

Combining cross-sectional and time series data is also known as pooling the data. Suppose the demand for automobiles in given by,

$$\ell n\, Y_t = \beta_1 + \beta_2 \ell n\, P_t + \beta_3 \ell n\, I_t + u_t$$

where Y = no of cars sold, P = average price, I = income, t = time

If we use cross sectional data we can obtain a fairly reliable estimate of the income elasticity β_3 because in such data, which are at a point in time the prices do not vary much. Let the crosssectionally estimated income elasticity be $\hat{\beta}_3$, then the proceeding time series regression may be written as

$$Y_t^* = \beta_1 + \beta_2 \ell n\, P_t + u_t \quad \text{where} \quad Y^* = \ell n\ Y - \hat{\beta}_3 \ell n\ I.$$

and represents that value of Y after removing from it the effect of income. We can now obtain the estimate of the price elasticity β_2 but the pooling technique may create problems of interpretation.

3. Dropping a variable and specification bias :

When faced with Multicollinearity one of the simplest way is to drop one of the choice variables. But this may result in specification bias which arises from incorrect specification of the model used in the analysis. Hence, the remedy may be worse than the disease.

4. Additional or new data :

Since Multicollinearity is a sample feature it is possible that in another sample involving the same variables collinearity may not be as serious as in the first sample but sometimes simple increasing the size of the sample may intensify the collinearity problem.

5. Reducing collinearity in polynomial :

It practice, it has been found that if explanatory variables are expressed in deviation form substantially reduces Multicollinearty.

6. Other statistical techniques :

Statistical techniques such as factor analysis and principle components such as ridge regression are employed to solve the problem of collinearity.

5.3. Specification Errors and their consequences :

A model should be correctly specified otherwise specification bias may occur. Specification error arises due to the following reasons——

1. Omission of a relevant variable :

Let a model be,

$$Y_i = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t \quad \text{................} \text{ (A)}$$

But instead of this, suppose one uses the model,

$$Y_i = \alpha_1 + \alpha_2 x_{2t} + v_t \quad \text{where} \quad v_t = u_t + \beta_3 x_{3t} \quad \text{.........} \text{ (B)}$$

Here since the first model (A) is true, adopting the second model (B) will constitute a specification error of omitting a relevant variable (X_{3t}).

2. Inclusion of an unnecessary or irrelevant variable :

Consider a model,

$$\gamma_i = \lambda_1 + \lambda_2 x_{2i} + \lambda_3 x_{3i} + \lambda_4 x_{4i} + u_t \quad \text{................} \text{ (C)}$$

If (A) is true, then model (c) constitutes a specification error of including an unnecessary or irrelevant variable.

3. Adopting the wrong functional from :

Consider another model,

$$\ell_n Y_i = \alpha_1 + \alpha_2 x_{2i} + \alpha_3 x_{3i} + u_{5i} \quad \text{................ (D)}$$

In relation to (A, model (B) also constitutes a specification bias of usage of the wrong functional form.

4. Errors of Measurement :

Let a model be such that

$$Y_i^* = \beta_1^* + \beta_2^* x_{2i}^* + \beta_3^* x_{3i}^* + u_i^* \quad \text{................ (E)}$$

where $y_i^* = y_i + \varepsilon_i$ and $x_i^* = x_i + w_i$.

ε_i and w_i being the errors of measurement (E) implies that instead of using the true Y_i and X_i we use their proxies, Y_i^* which contain errors of measurement. Thus we commit the errors of measurement bias.

Thus, above mentioned were the different types of specification errors.

Consequences of specification errors :

1. Consequence due to omission of a relevant variable

Suppose that the true model is

$$Y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i \quad \text{.............. (i)}$$

But if fit the following model

$$Y_i = \alpha_1 + \alpha_2 x_{2i} + v_i \quad \text{......... (ii)}$$

Now the consequences of omitting X_3 are as follows:

i) If the left out variable X_3 is correlated with the included variable X_2, i.e, r_{23} is non-zero $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are biased as well as inconsistent i.e $E(\hat{\alpha}_1) \neq \beta_1$ and $E(\hat{\alpha}_2) \neq \beta_2$

ii). Even if X_2 and X_3 are uncorrelated $(r_{23} = 0)$, $\hat{\alpha}_1$ is still biased, although $\hat{\alpha}_2$ is now unbiased.

iii). The disturbance variance σ^2 incorrectly estimated.

iv). The variance of $\hat{\alpha}_2$ is a biased estimator of the variance of the true estimator $\hat{\beta}_2$.

v). The confidence interval and hypothesis testing procedures are likely to give misleading conclusions about the statistical significance of the estimated parameters.

2. Consequence of inclusion of an Irrelevant variable

Let us assume that a true model is

$$Y_i = \beta_1 + \beta_2 x_{2i} + u_i \quad \text{............. (i)}$$

But we fit the model

$$Y_i = \alpha_1 + \alpha_2 x_{2i} + \alpha_3 x_{3i} + v_i \quad \text{......... (ii)}$$

Here we commit the specification error of including an unnecessary variable in the model

The consequences of this specification error are—

1. The OLS estimators of the parameters of the incorrect model (ii) are all unbiased and consistent i.e

$$E(\alpha_1) = \beta_1,\ E(\hat{\alpha}_2) = \beta_2 \text{ and } E(\hat{\alpha}_3) = \beta_3 = 0$$

2. The error variance σ^2 is correctly estimated.

3. The usual confidence interval and hypothesis testing procedures remain valid.

4. The estimated α's will be generally inefficient, i.e, their variances will be generally larger than those of the $\hat{\beta}$'s of the true model.

Thus, these were the two consequences of omitting a relevant variable and inclusion of a relevant variable under specification error.

5.4 Qualitative Factors and Dummy variables :

A dummy variable is a variable which we construct to describe the development or variation of the variable under construction. They are used as proxies for other variables which cannot be measured in any particular case various reasons.

1. Dummy variables as proxies to qualitative (categorical) factors.

· Dummy variables are commonly used as proxies for qualitative factors such as profession religion, sex, region etc. For example, let us consider the demand function,

$$D_i = b_1 + b_1 X_{1i} + b_2 X_{2i} + \mu_i$$

where X_1 = income and X_2 = dummy variable for region ($b_2 > 0$)
Here $b_2 = 1$ for a person living in a town
and $b_2 = 0$ for a person living in a rural area.

2. Dumr·.y variables are proxies to numerical factors. Dummy variables may be used as proxies for quantitative factors, when no observations on these factors are available or when it is convenient to do so. For example, suppose we want to measure the savings function $S = f(Y)$. We assume that people become more thrifty as they grow old, and so the dummy variable for 'age' may be assigned the value of zero, if the person belongs to the first age group i.e between 20-35 years of age, and the value 1 if the person belongs to the second age group. The savings function assumes the form

$$S_i = b_0 + b_1 x_i + b_2 z_i + u_i$$

where, X_i = Income, and Z_i = dummy variable 'age' ($b_2 > 0$.)

3. Use of dummy variables for measuring the shift of a function over time.

A shift of a function that the constant intercept changes in different periods, while the other co-efficient remain constant. Such shifts may be taken into account by the introduction of a dummy variable in the function.

4. Use of dummy variables for measuring the change of parameters (slopes) over time

It is known that over long periods of time or in abnormal (war) years not only do the function shift (their constant intercept changes) but also their slopes may be expected to change Here elasticity's and propensities of a function may be captured by introducing appropriate dummy variables in the function.

5. Use dummy variables as proxies for the dependent variable.

The dependent variable of a function may be used as a dummy variable. For example, suppose we want to measure the determinants of car-ownership. Some people will have cars while others will not. Assuming that the determinates of ownership are income and profession, we have—

$$C = b_0 + b_1 Y + b_2 A + u$$

where C = crowners or non-owners, Y = income

A = dummy variable for profession.

Here, the dependent variable, C, will be a dummy variable which may be assigned the value 1 for a person who owns a car, and 0 for a person who does not. In this case, the dependent variable is dichotomous.

6. Use of dummy variables for seasonal adjustment of time series. One of the most common applications of dummy variables is in removing seasonal variations in time series.

5.5. INTRODUCTIONS TO HETROSCEDASITICITY AND AUTOCORRELATION OF DISTURBANCES:

HEREROSCEDASTICITY

One of the important assumptions of the classical linear regression model about the random variable is that its probability distribution remains the same over all the observations of X and in particular the variances of each u_i is the same for all the values of the explanatory variable. This is the assumption of homoscedosticity or equal (homo) spread (scadasticity) i.e. equal variance.

Symbolically,

$$\text{Var}(u_i) = E[u_i - E(u_i)]^2$$

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

If the assumption of homoscedosticity is not satisfied in any particular cause we say that the u's are heteroscedastic. Symbolically,

$$E(u_i) = \sigma^2_{u_i} \quad (\text{not constant})$$

where the subscript i signifies the fact that the individual variances may all be different.

AUTOCORRELATION

Another assumption of Ordinary Least Square is that the successive values of the random variable are temporally independent i.e that the values which u assumes in any one period is independent of the values which it assumed in any previous period. This assumption implies that,

$$\begin{aligned}\mathrm{Cov}(u_i u_j) &= E\left[\{u_i - E(u_i)\}\{u_j - E(u_j)\}\right] \\ &= E\left[u_i u_j\right] \qquad \left(\because E(u_i) = E(u_j) = 0 \text{ by assumption}\right) \\ &= 0\end{aligned}$$

If this assumption is not satisfied i.e if the value of u in any particular period is correlated with its own preceding value, we say that there is autocorrelation or serial correlation between the random variable. Autocorrelation refers to the relationship between successive residuals of the same variable.

5.6 Summary :

Multicollinearity implies a correlation between some or all explanatory variables of a regression model. Also there are various theoretical and practical consequences of Multicollinearity. There are also various rules for the detection of Multicollinearity. There are also certain rules of thumb to solve the problem of Multicollinearity to solve the problem of Multicollinearity. Again specification errors arise because models are not correctly specified. It occurs due to a number of causes. Also the consequence of omission of a relevant variable and inclusion of an irrelevant variable has been studied.

Dummy variables are variables which are used as proxies to qualitative factors, numerical factors, measuring shift of a function over time, measuring change in parameters, proxies for dependent variable and for seasonal adjustment of time series. Again homoscedosticity and no autocorrelation are two assumptions of classical linear regression model, and violation of which leads to hetroscedasiticity and autocorrelation.

5.7 Additional Readings

1. Johnston, J. "Econometric Methods", Mc Graw Hill

2. Gujarathi, D., "Basic Econometric", Mc Graw Hill.

3. Salvatore, D and Reagle, D, "Statistic and Econometrics", Tata Mc Graw Hill.

4. Gupta, S. C., "Fundament ants of Statistics".

5.8 Self Assessment Test

1. What do you understand by Multicollinearity. What are its effects?

2. Explain the various methods of detection of Multicollinearity. Also explain the remedies to solve Multicollinearity.

3. Brief explain the meaning of specification error, the types of specification errors, and their consequences.

4. Write a short note an qualitative factors and dummy variables.

5. Define Hetroscedasiticity and autocorrelation of disturbances.

●●●