MODULE 1

1

THE CONCEPT OF A RANDOM VARIABLE

Unit Structure:

1.0 OBJECTIVES

After going to this module you will be able :

- To understand concept of random variable
- To understand various types of random variable
- To understand the meaning of covariance and correlation

1.1 INTRODUCTION

A random variable is a variable whose value is not known or a function that assigns values to each of an experiment's outcome. Random variables are oflenly used in econometric or regression analysis to determine statistical relationships between two or more variables.

Random variables are associated with random processes where a random process is an event or experiment that has a random outcome. For e.g. rolling a die, tossing a coin, choosing a card or any one of the other possibilities. It is something which we would guess but cannot predict the exact outcomes. So we have to calculate the probability of a particular outcome.

Random variables are denoted by capital letters for e.g. 'X', 'Y' where it usually refers to the probability of getting a certain outcome. Random variables give numbers to outcomes of random events. It means though an event is random but its outcome is

mumotes.in

quantifiable. For e.g. rolling a die. Let's say we wanted to know how many sixes we will get if we roll a die for a certain number of times. In this case random variable X could be equal to 1 if we get a six & O if we get any other number.

Let us discuss another example of a random variable i.e. the outcome of a coin toss. Let us assume that the probability distribution in which the outcomes of a random event are not equally likely to happen. If random variable, Y, is the number of heads we get from tossing two coins, then Y could be 0, 1 or 2. This means that we could have no heads, one head or both heads on a two - coin toss. However, the two coins land in four different ways. TT, HT, TH and HH. Therefore, the $P(Y=0)=\frac{1}{4}$.

Since we have one chance of getting no heads (i.e. two fails (TT) when the coins are tossed. Similarly, the probability of getting two heads (HH) is also 1/4 . So getting one head has a likelihood of occurring two times : HT and TH. In this case, $P(Y=1)=2/4=1/2$.

1.2 TYPES OF RANDOM VARIABLES

There are two types of random variables :

A) Discrete random variables
B) Continuous random variables

Discrete random variables take into account a countable number of distinct values. For e.g. an experiment of a coin tossed for three times. If X represents the number of times that the outcome will come up heads, then X is a discrete random variable that can only have values 0, 1, 2, 3 (from no heads in thee successive coin tosses to all heads). No other value is possible for X.

Continuous random variables can represent any value within a specified range or interval and can take on an infinite number of possible values. E.g. an experiment that involves measuring the amount of rainfall in a city over a year or overage height or weight of a random group of 100 people.

1.3 MEAN OF A RANDOM VARIABLE

The mean of a discrete random variable X is a weighted average of the possible values that the random variable can take. The mean of a random variable weights each outcome x_i according to its probability, pi. Therefore expected value of X is μ and formula is

mynotes.in

$$\mu = x_1 p_1 + x_2 p_2 + \ldots\ldots\ldots\ldots + x_k p_k$$
$$= \sum x_i p_i$$

The mean of a random variable provides the long-run average of the variable, or the expected average outcome over many observations.

For a continuous random variable, the mean is defined by the density curve of the distribution. For a symmetric density curve, such as normal distribution, the mean lies at the center of the curve.

1.4 VARIANCE OF A RANDOM VARIABLE

The variance of a discrete random variable X measures the spread, or variability of the distribution and is defined by

$$\sigma_X^2 = \sum (x_i - \mu_x)^2 p_i$$

The standard deviation σ is the square root of the variance.

1.5 BASIC LAWS OF PROBABILITY

Probability is defined as a number between 0 and 1 representing the likelihood of an event happening. A probability of 0 indicates no chance of that event occurring, whereas a probability of 1 means the event will occur.

Basic Properties of Probability Rules:

- Every probability is between 1 and 1. In other words, if AA is an event, then $0 \le P(A) \le 10 \le P(A) \le 1$.
- The sum of the probabilities of all the outcomes is one. For e.g. if all the outcomes in the sample space are denoted by $A_i A_i$ then $\sum A_i = \left|\sum A_i = 1\right.$.
- Impossible events have probability zero. If event AA is impossible, then $P(A) = OP(A) = 0$.
- Certain events have probability 1. If event AA is certain to occur, then $P(A) = IP(A) = 1$.

The Probability Rules:

1) Rule 1 : Whenever an event is the union of two other events, the Addition Rule will apply. If AA & BB are two events, then

$$P(A \cup B) = P(A) + P(B) - P(A \text{ and } B)\, P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

munotes.in

If can be written as :
$P(A \cup B) = P(A) + P(B) - P(A \cap B) P(A \cup B) = P(A) + P(B) - P(A \cap B)$

2) Rule 2 : Whenever an event is complement of another event, the complementary rule will apply. If AA is an event then we have the following rule.

$$\boxed{P(not\ A) = 1 - P(A) P(not\ A) = 1 - P(A)}$$

This is also written as
$P(A^{-}) = 1 - P(A) P(A^{-}) = 1 - P(A)$

3) Rule 3 : Whenever partial knowledge of an event is available, then the condition rule will be applied. If event AA is already known to have occurred and probability of event BB is desired, then we will have the following rule.

$$\boxed{P(B\ given\ A) = P(A\ and\ B) P(A) P(B\ given\ A) = P(A\ and\ B) P(A)}$$

Where it is further written as :

$P(B/A) = P(A \cap B) P(A) P(B/A) = P(A \cap B) P(A)$

4) Rule 4 : Whenever an event is the intersection of two other events, the Multiplication rule will apply. If events AA and BB need to occur simultaneously then, we have the following rule.

$$\boxed{P(A\ and\ B) = P(A) P(B\ given\ A) P(A\ and\ B) = P(A) P(B\ given\ A)}$$

It is also written as :
$P(A \cap B) = P(A) P(B/A) P(A \cap B) = P(A) P(B/A)$

Let us discuss these rules with the help of an example of rolling a dice. Suppose we roll two dice.

1) The probability that both dice are 5 is:
P (both are 5) = P(first is a 5 and second is a 5) P(both are 5) =
P (first is a 5 and second is a 5) 12

= P(first is a 5) P(second is a 5, given first is a 5) = 1.6.16 = 136 =
P(first is a 5) P(second is a 5, given first is a 5) = 16.16 = 136

Here the word 'both' indicates two events had to happen at the same time, i.e. the first event and the second event. We used

mynotes.in

the multiplication rule because of the key word 'and'. The first factor resulted from the Basic Rule on a single die.

2) The probability that at least one die is a 5 is :
P(at least one is a 5) = P(first is a 5 or second is a 5) P(at least one is a 5) = P(first is a 5 or second is a 5) 12

= P(first is a 5) + P(second is a 5) - P(first is a 5 and second is a 5) = 16 + 16 - 136 = 1136 = P(first is a 5) + P(second is a 5) - P(first is a 5 and second is a 5) = 16 + 16 - 136 = 1136

"First we had to recognize that the event at least one" could be fulfilled by one or the other of two separate cases. We used 'Addition rule' because of the word 'or'. The first two terms come from the Basic Rule on a single die, while the third term resulted from only one outcome where both dice will be 5.

3) The probability that neither die is a 5 is :
P(neither is a 5) = 1 - P(at least one is a 5) = 1 - 1136 = 2536 P(neither is a 5) = 1 - P(at least one is a 5) = 1 - 1136 = 2536.

In this case, the word "neither" is complementary to the word "at least one" so we used the Complementary rule.

4) Given that at least one of the dice is a 5, the probability that the other is a 5 is:
P(other is a 5 / at least one is a 5) = P(both are 5) P(at least one is a 5) = 1361136 = 111P(other is a 5 / at least one is a 5) = P(both are 5) P(at least one is a 5) = 1361136 = 111.

The partial knowledge required the conditional rule.

1.6 TYPES OF DISCRETE RANDOM VARIABLES

When solving problems. we should be able to recognize a random variable which fits one of the formats of Discrete random variables.

1) Bernoulli Random Variable: Is the simple kind of random variable. It can take on two values 1 and 0. If an experiment with probability P resulted in success then it takes on a 1 and 0 if the result is failed. For e.g. If the shooter hits the target, we call it a 'success' and he misses it then we call it a 'failure'. Let us assume that whether the shooter hits or misses the particular target on any particular attempt has nothing to do with his success or failure on any other attempts. In this case we are ruling out the possibility of improvement by the shooter with practise. Assuming probability of a success is P and that of failure is 1- p, where p is a constant

mbnotes.in

between values 1 and 0. A random variable that take value 1 in case of success and 0 in case of failure is called Bernoulli random variable.

The Bernoulli distribution with parameter P if its probability mass function (pmf) is $P(x) = P^x X (1-p)^{1-x}, x = 0,1$

Where, $x = 0, P(x) = 1 - p$ and

if $x = 1, P(x) = p$.

Conditions for Bernoulli trials

1) A finite number of trials.
2) Each trial should have exactly two outcomes success or failure.
3) Trials should be independent.
4) The probability of success or failure should be the same in each trial.

For e.g. - Tossing a coin. Suppose, for a Bernoulli random variable, $p = 0.4$. Then

$$p(0) = 0.6, p(1) = 0.4.$$

Suppose the coin is tossed for four times. The event that the outcome will be Head on the first trial, and Tail on the next two and again Head on the last can be represented as :

$$S = (1,0,0,1)$$

The probability with which the outcome is Head is P, whereas the probability with which Tail will occur is 1 - p. The event 'H' or 'T' on each trial are independent events, in the sense that whether the outcomes is H or T on any trial is independent of the chance of 'Head' or 'Tail' on any previous or subsequent trials. If A and B are independent events, the probability of observing A and B equals the probability of A multiplied by the probability of B. Therefore, the probability of observing 1,0,0,1 together is :

$$p \times (1-) \times (1-p) \times p = p^2 \times (1-p)$$

2) The Binomial Random Variable:

A binomial distribution can be thought of as simply the probability of a success or failure outcome in an experiment or survey that is repeated multiple times. It has only two possible outcome (the prefix "bi" means two) for e.g. a coin toss has only two outcomes heads or fails or taking a test could have two outcomes pass or fail.

Kunotes.in

A binomial random variable is the number of successes X in n repeated trials of a binomial experiment. The probability distribution of a binomial random variable is called a binomial distribution.

For a variable to be classified as a binomial random variable, the following conditions must be satisfied:

1) There must be a fixed sample size (a certain number of trials)
2) For each trial, the success must either happen or it must not.
3) The probability for each event must be exactly the same.
4) Each trial must be an independent event.

The binomial probability refers to the probability that a binomial experiment results in exactly X successes. Given $X, n \& p$, we can compute the binomial probability based on the binomial formula :

Suppose a binomial consists of n trials & results in X successes and if the probability of success on an individual trial is P, then the binomial probability is :

$$b(X', n, p) = {}_nC_r XP^x X(1-p)^{n-x}$$

OR

$$b(X', n, p) = \{{}_n!/[X!(n-x)!]\} \times P^x X(1-p)^{n-x}$$

Where

$X \rightarrow$ The number of successes that result from the binomial experiment.

$n \rightarrow$ The number of trials

$p \rightarrow$ The probability of success on an individual trial

$Q \rightarrow$ The probability of failure on an individual trial $C = (1-p)$

$n! \rightarrow$ The factoral of n.

$b(X', n, p) \rightarrow$ binomial probability

${}_nC_r \rightarrow$ the number of combinations of n things, taken r at a time.

For e.g. Suppose a die is tossed or 5 times. What is the probability of getting exactly 2 fours?

Solution :

This is a binomial experiment in which the number of trials is equal to 5, the number of successes is equal to 2 and the probability of success on a single trial is 1/6 or about 0.167. Therefore, the binomial probability is :

nunotes.in

$$b(2;5,0.167) = {}_5C_2 \times (0.167)^2 \times (0.833)^3$$
$$b(2;5,0.167) = 0.161$$

3) The Poisson Distribution :

A poisson distribution is the discrete probability distribution that results from a Poisson experiment. It has the following properties.

→ The experiment results in outcomes as successes or failures.
→ The average number of successes (μ) that occurs in a specified known region.
→ The probability that a success will occur is proportional to the size of the region.
→ The probability that the success will occur in an extremely small region is virtually zero.

For e.g. A certain restaurant gets an average of 4 customers per minute for takeaway orders. In this case, a poisson distribution can be used to analyze the probability of various events, regarding total number of customers visiting for takeaway orders. It helps a manager of the restaurant to plan for such events with staffing & scheduling.

Likewise the poisson distribution can also be applied in subjects like biology, disaster management, finance where the events are time dependent.

A Poisson random variable is the number of success that result from a Poisson experiment. The probability distribution of a Poisson random variable is called a Poisson distribution.

Suppose the average number of successes within a given region is μ, then the Poisson probability is :

$$P(X'', \mu) = \frac{\left(e^{-\mu}\right) \times \left(\mu^x\right)}{X!}$$

Where e : a constant equal to approximately 2.71828 (e is the base of the natural logarithm system)
μ : the mean number of successes that occur in a specified region.
X : the actual number of successes that occur in a specified region.
$P(X'', \mu)$: The Poisson probability

mumtest.in

For e.g.

The average number of high end cars are sold by the dealer of a Luxury Motor Company is 2 cars per day. What is the probability that exactly 3 high end cars will be sold tomorrow?

Solution : We have the values of

μ = 2, the average number of high end car are sold per day

X = 3, probability that 3 high end cars will be sold tomorrow

e = 2.71828, a constant

By using the Poisson Formula we get :

$$P(X'', \mu) = \frac{\left(e^{-\mu}\right) \times \left(\mu^{x}\right)}{X!}$$

$$P(3;2) = \frac{\left(2.71828^{-2}\right) \times \left(2^{3}\right)}{3!}$$

$$P(3;2) = \frac{(0.13534) \times (8)}{6}$$

$$P(3;2) = 0.180$$

Thus, the probability of selling 3 high end cars by tomorrow is 0.180.

1.7 CONTINUOUS DISTRIBUTION : THE NORMAL DISTRIBUTION

The normal distribution refers to a family of continuous probability distribution. It is also known as the Goussian distribution & the bell curve. It is a probability function that describes how the values of a variable are distributed. It is a symmetric distribution where most of the observations cluster around the central peak and the probabilities for values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely.

The normal equation for the normal distribution when the value of the random variable X is :

$$f(X) = \left(1/\sqrt{2\pi\sigma}\right) \exp\left(-(x-\mu)^2 / 2\sigma^2\right)$$

$X \rightarrow$ normal random variable, $-\infty < X < \infty$

$\mu \rightarrow$ mean

$\sigma \rightarrow$ standard deviation

$\pi \rightarrow$ approximately 3.14159

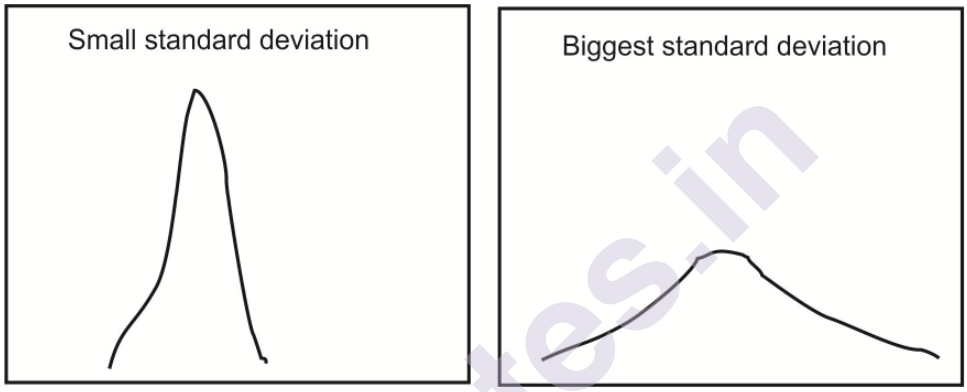
mumnotes.in

$e \rightarrow$ approximately 2.71828

The normal equation is the probability density function for the normal distribution.

The Normal curve: The normal distribution depends on two factors - the mean and the standard deviation. The mean of the distribution determines the location of the centre of the graph, and the standard deviation determines the height and width of the graph. All normal distributions look like a symmetric, bell-shaped curve as show below:

Figure No. 1.1



When the standard deviation is small, the curve is tall and narrow and when the standard deviation is big, the curve is short and wide.

Probability and Normal curve

The normal distribution is a continuous probability distribution, where

$\rightarrow$ the total area under the normal curve is equal to 1

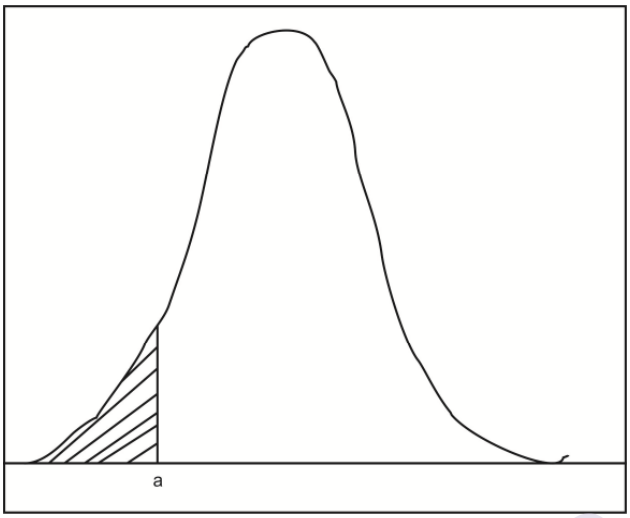
$\rightarrow$ the probability that a normal random variable X equals any particular value is 0

$\rightarrow$ the probability that X is greater than a equals the area under the normal curve bounded by a and plus infinity (indicated by non-shaded area in the figure below)

$\rightarrow$ the probability that X is less than a equals the area under the normal curve bounded by a and minus infinity (indicated by the shaded area in the figure below)

munotes.in

Figure No. 1.2



There are some important features of the normal distribution as follows:

1. The distribution is symmetrical about the mean, which equals the median and the mode.
2. About 68% of the area under the curve falls within 1 standard deviation of the mean.
3. About 95% of the area under the curve falls within 2 standard deviation of the mean.
4. About 99.7% of the area under the curve falls within 3 standard deviations of the mean.

These last 3 points are collectively known as the empirical rule or the 68 - 95 - 99.7 rules. Let us discuss it with an example of an express food delivery by a Restaurant. Assuming that a mean delivery time of 30 minutes and a standard deviation of 5 minutes. Using the Empirical Rule. We can determine that 68% of the delivery times are between 25-35 minutes (30 + / - 5), 95% are between 20 - 40 minutes (30 + / - 2 x 5), 99.7% are between 15 - 45 minutes (30 + / - 3 x 5).

Suppose, an average tubelight manufactured by ABC Corporation lasts 300 days with a standard deviation of 50 days. Assuming that tubelight life is normally distributed, what is the probability that ABC corporation's tubelight will last at most 365 days?

Solution : Given a mean score of 300 days & a standard deviation of 50 days, we want to find the cumulative probability that tubelight life is less than or equal to 365 days. Thus,

munotes.in

$\rightarrow$ the value of the normal random variable is 365 days.
$\rightarrow$ the mean is equal to 300 days.
$\rightarrow$ the standard deviation is equal to 50 days.

By entering these values to find out cumulative probability we get

$$P(X \leq 365) = 0.90$$

Hence, there is a 90% chance that a tubelight will burn out within 365 days.

1.8 REFERENCE

- S-Shyamala and Navdeep Kaur, 'Introduce too y Econometrics'.
- Neeraj R, Hatekar, 'Principles of Econometrics : An Introduction us in, R'

❖❖❖❖

munotes.in

2

COVARIANCE AND CORRELATION

Unit Structure:

2.0 OBJECTIVES

After going to this module you will be able :

- To understand the meaning of covariance and correlation.
- To understand the method of studying correlation.
- To understand the law of large numbers.

2.1 INTRODUCTION

Covariance is a measure used to determine how much two random variables differ by its respective mean and Correlation is a statistical method which helps in analysing the relationship between two or more variables. The value of the covariance coefficient lies between $-\infty$ and $+\infty$ and the value of correlation coefficient lies between -1 and +1.

2.2 COVARIANCE

Covariance is a measure used to determine how much two random variables differ by its respective mean. In other words, the prefix 'Co' refers to a joint action and variance refers to the change. In covariance, two variables are related based on how these variables change in relation with each other. The value of the covariance coefficient lies between $-\infty$ and $+\infty$.

For population,

$$COV(X,Y) = \frac{\sum_{i=1}^{n}\left((X_i - \bar{X})(Y_i - \bar{Y})\right)}{n}$$

munotes.in

Where $X, Y \rightarrow$ two random variables

$\bar{X} \rightarrow$ mean of random variable X

$\bar{Y} \rightarrow$ mean of random variable Y

$n \rightarrow$ length of random variable X, Y

For sample

$$COV(X,Y) = \frac{\sum_{i=1}^{n}\left((X_i - \bar{X})(Y_i - \bar{Y})\right)}{n-1}$$

$X \,\&\, Y \rightarrow$ mean of given sample set

$n \rightarrow$ total number of sample

$X_i \; and \; Y_i$ individual sample of set

2.3 CORRELATION ANALYSIS

Correlation is a statistical method which helps in analyzing the relationship between two or more variables. The study of correlation is useful due to following reasons:

1) Since most of the variables have some kind of relationship, quantification of it is necessary to learn more about them.

2) Correlation is a first step towards estimation or prediction of unknown values of the variables.

3) An understanding of the degree and nature of correlation between two or more variables helps in reducing uncertainties about the economic behaviour of important variables like price level and money supply, interest rate and investment, taxation and willingness to work, etc.

Correlation is classified into three ways:

1) **Positive and Negative correlation (Depends upon the direction of change) :** When both the variables change in the same direction, (i.e. they increase or decrease together) it is positive correlation. For example when price rises, supply also increases, when income falls, consumption also declines. When increase in one variable is accompanied by a fall in other, it is negative correlation. For example, increase in price leads to fall in demand; increase in interest rate is accompanied by a fall in investment.

2) Simple and Multiple correlation (Depends upon number of variables under study) : Simple correlation is the relationship

munotes.in

between two variables like height and weight of a person, or wage rate and employment in the economy. Multiple correlations, on the other hand, examines relationship between three or more variables. For example a relationship between production of rice per acre, rainfall and use of fertilizers is multiple in nature.

3) Linear and non-linear (Depends on the ratio of change between two variables) : When a change in one variable is in constant ratio with a change in other, it is linear relationship. For example doubling the amount of fertilizers used exactly doubles the yield per acre, it is linear relationship. Non-linear relationship exists when a change in one variable is not in constant ratio with a change in other. In this case doubling the amount of fertilizers may not exactly double the output per acre.

2.4 METHODS OF STUDYING CORRELATION

Following important method of studying correlation between two variable will be discussed in this unit.

- Scatter diagram method.
- Karl Pearson□s Coefficient of Correlation.
- Rank Correlation Coefficient.

2.4.1 Scatter diagram

It is the simplest method of studying correlation, by using graphical method. Under this method, a given data about two variables is plotted in terms of dots. By looking at the spread or scatter of these dots, a quick idea about the degree and nature of correlation between the two variables can be obtained. Greater the spread of the plotted points, lesser is an association between two variables. That is, if the two variable are closely related, the scatter of the points representing them will be less and vice versa.

Following are different scatter diagrams explaining the correlation of different degrees and directions.

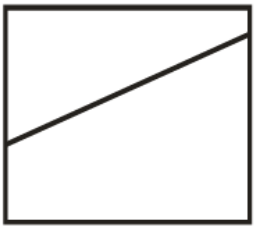


Fig. 1

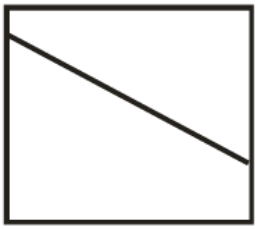


Fig. 2

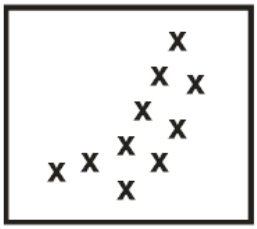


Fig. 3

mynotes.in

Fig. 4

Fig. 5

Fig. 6

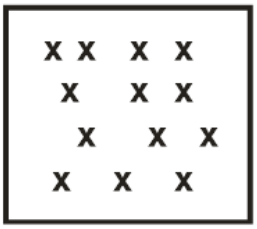


Fig. 7

1) Figure 1 represents positive perfect correlation where coefficient of correlation (r) = 1.
2) Figure 2 represents perfect negative correlation where coefficient of correlation (r) = -1
3) Figure 3 indicates high degree positive correlation where r = + 0.5 or more.
4) Figure 4 indicates high degree negative correlation where r = - 0.5 or more.
5) Figure 5 represents low degree positive correlation where the scatter of the points is more.
6) Figure 6 represents low degree negative correlation where the scatter for the points is more in negative direction.
7) Figure 7 indicates that there is no correlation between two variables. Here r = 0.

Thus, the closeness and direction of points representing the values of two variables determine the correlation between the same.

Advantages and Limitations of this method.

- It is a simple method giving very quick idea about the nature of correlation.
- It does not involve any mathematical calculations.
- It is not influenced by the extreme values of variables.
- This method, however, does not give exact value of

muinotes.in

- coefficient of correlation and hence is less useful for further
- statistical treatment.

2.4.2 Karl Pearson's Coefficient of Correlation (r) :

This is the most widely used method of studying a bi-variate correlation. Under this method, value of r can be obtained by using any of the following three ways.

I) Direct Method of finding correlation coefficient

$$r = \frac{N\sum XY - \sum X \quad \sum Y}{\sqrt{N\sum X^2 - \sum X^{\,2}}\sqrt{N\sum Y^2 - \sum Y^{\,2}}}$$

Where N = No. of observations

II) Taking deviations from actual mean

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$$

Where $x = X - \overline{X}$, $y = Y - \overline{Y}$

III) Taking deviations from assumed mean

$$r = \frac{N\sum dx\,dy - \sum d_x \times \sum d_y}{\sqrt{N\sum d_x^2 - \sum d_x^{\,2}}\sqrt{N\sum d_y^2 - \sum d_y^{\,2}}}$$

Ex.1 Calculate Karl Pearson□s coefficient of correlation using direct method.

X	Y	X^2	Y^2	XY
9	15	81	225	135
8	16	64	256	128
7	14	49	196	98
6	13	36	169	78
5	11	25	121	55
4	12	16	144	48
3	10	9	100	30
2	8	4	64	16
1	9	1	81	9
$\sum X = 45$	$\sum Y = 108$	$\sum X^2 = 285$	$\sum Y^2 = 1356$	$\sum XY = 597$

munotes.in

$$r = \frac{N\sum XY - \sum X \quad \sum Y}{\sqrt{N\sum X^2 - \sum X^{\,2}}\ \sqrt{N\sum Y^2 - \sum Y^{\,2}}}$$ N – No. of observations

As per the table,

$N = 9, \sum X = 45, \sum Y = 108, \sum XY = 597, \sum X^2 = 285$ and $\sum Y^2 = 1356$

By substituting the values in the formula,

$$r = \frac{9 \times 597 - \ 45 \quad 108}{\sqrt{9 \times 285 - \ 45^{\ 2}}\ \sqrt{9 \times 1356 - \ 108^{\ 2}}}$$

$$= \frac{5373 - 4860}{\sqrt{2565 - 2025}\ \sqrt{12204 - 11664}}$$

$$= \frac{513}{\sqrt{540}\sqrt{540}} \quad = \frac{513}{540} = +0.95$$

Since r = 0.95, there is high degree positive correlation between x and y.

Ex. 2 Calculate Karl Pearson□s coefficient of correlation by taking deviations from actual mean.

x	y	$x = x - \bar{x}$	$y = y - \bar{y}$	xy	x^2	y^2
6	9	6-6 = 0	9-8 = 1	0	0	1
2	11	2-6 = -4	11-8 = 3	-12	16	9
10	5	10-6 = 4	5-8 = -3	-12	16	9
4	8	4-6 = -2	8-8 = 0	0	4	0
8	7	8-6 = 2	7-8 = -1	-2	4	1
$\sum x = 30$	$\sum y = 40$			$\sum xy = -26$	$\sum x^2 = 40$	$\sum y^2 = 20$

Formula for this method:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$$

Since $x = x - \bar{x}$ and $y = y - \bar{y}$, it is necessary to find arithmetic mean $\bar{x}$ and $\bar{y}$.

$$\bar{x} = \frac{\sum x}{N} \qquad \bar{y} = \frac{\sum y}{N}$$

$$= \frac{30}{5} \qquad = \frac{40}{5}$$

$$= 6 \qquad = 8$$

Where N – number of observations

mynotes.in

From the table, it is clear that

$\sum xy = -26,\ \sum x^2 = 40,\ \sum y^2 = 20$

By substituting the values in the formula

$$r = \frac{-26}{\sqrt{40 \times 20}} = \frac{-26}{\sqrt{800}} = \frac{-26}{28 \cdot 28}$$

$$= -0.92$$

Since r = - 0.92, the correlation between x and y is high degree negative.

Ex.3 Compute Karl Pearson□s coefficient of correlation by taking deviations from assumed mean.

(This method is used when the actual means are in fractions)

X	Y	$dx = X - A$	$dy = Y - A$	dx^2	dy^2	dxdy
2	25	2-9 = -7	25-29 = -4	49	16	28
5	27	5-9 = -4	27-29 = -2	16	4	8
7	26	7-9 = -2	26-29 = -3	4	9	6
9	29	9-9 = 0	29-29 = 0	0	0	0
19	34	19-9 = 10	34-29 = 5	100	25	50
16	39	16-9 = 7	39-29 = 10	49	100	70
		$\sum dx = 4$	$\sum dy = 6$	$\sum dx^2 = 218$	$\sum dy^2 = 154$	$\sum dxdy = 162$

For the above data, actual means $\bar{X}$ and $\bar{Y}$ will be in fraction. So we can take assumed means for both the variables and then find the deviations dx and dy .

Let assumed means for X = 9
Let assumed mean for Y = 29

$$r = \frac{N\sum dx\,dy - \sum dx \times \sum dy}{\sqrt{N\sum dx^2 - \left(\sum dx\right)^2}\sqrt{N\sum dy^2 - \left(\sum dy\right)^2}}$$

munotes.in

From the table,

$N = 6, \sum dx\, dy = 162,\ \sum dx = 4,\ \sum dy = 6, \sum dx^2 = 218, \sum dy^2 = 154$

By substituting these values in the formula,

$$r = \frac{6 \times 162 - 4 \times 6}{\sqrt{6 \times 218 - 4^2}\sqrt{6 \times 154 - 6^2}}$$

$$= \frac{972 - 36}{\sqrt{1308 - 16}\sqrt{924 - 36}}$$

$$= \frac{948}{\sqrt{1292}\sqrt{888}} = \frac{948}{\sqrt{1147296}} = \frac{948}{1071 \cdot 12}$$

$$= 0.89$$

Since r = 0.89, there is high degree positive correlation between Xand Y.

Check your progress

1) Find correlation coefficient for the following data.

X	10	6	9	10	12	13	11	9
Y	9	4	6	9	11	13	8	4

Ans : $\boxed{r = 0.896}$

2)

X	45	70	65	30	90	40	50	75	85	60
Y	35	90	70	40	95	40	60	80	80	50

Ans: $\boxed{r = 0.903}$

3)

X	15	12	16	15	17	14	18
Y	17	14	20	25	20	24	22

Ans: $\boxed{r = 0.214}$

2.4.3 Rank Correlation:

For certain categories like beauty, honestry, etc quantitative measurement is not possible. Also sometimes the population under study may not be normally distributed. In such cases, instead of Karl Pearson□s co-efficient of correlation, Spearman□s Rank correlation coefficient is calculated. This method is used to the determine the level of agreement or disagreement between two judges. The calculations involved in this method are much simpler

munotes.in

than the earlier method. Rank correlation is calculated using the following formula.

$$R = 1 - \frac{6\sum D^2}{N\ N^2 - 1}$$

When D- difference between rank 1 and rank 2. N- No. of observations

Rank correlation is computed in following two ways:
1) When ranks are given.
2) When ranks are not given.

Rank correlation when ranks are given:
Ex.4 Following are the ranks given by two judges in a beauty contest. Find rank correlation coefficient.

Ranks by judge 1 R_1	Ranks by judge 2 R_2	$D = R_1 - R_2$	D^2
1	4	-3	9
2	5	-3	9
3	6	-3	9
4	7	-3	9
5	8	-3	9
6	2	4	16
7	3	4	16
8	1	7	49
$N = 8$			$\sum D^2 = 126$

$$R = 1 - \frac{6 \times \sum D^2}{N\ N^2 - 1}$$

By substituting the values from the table

$$= 1 - \frac{6 \times 126}{8\ 8^2 - 1}$$

$$= 1 - \frac{6 \times 126}{8 \times 63} = 1 - \frac{756}{504}$$

$$= 1 - 1.5 = -0.5$$

Since rank correlation co-efficient is -0.5, there is a moderate negative correlation between the ranking by two judges.

Calculation of rank correlation co-efficient, when the ranks are not given:

Ex.4 Calculate rank correlation for the following data.

munotes.in

X	Y	R_1 Rank to X	R_2 Rank to Y	$D = R_1 - R_2$	D^2
67	78	2	2	0	0
42	80	8	1	7	49
53	77	7	3	4	16
66	73	3	6	-3	9
62	75	4	4	0	0
60	68	5	7	-2	4
54	63	6	8	-2	4
68	74	1	5	-4	16
					$\sum D^2 = 98$

When the ranks are not given, we have to assign ranks to the given data. The ranks can be assigned in ascending (Rank 1 to the lowest value) or descending (Rank 1 to the highest value) order.

In this example, ranks are given in descending order.
The highest value gets rank 1 and so one.

$$R = 1 - \frac{6\sum D^2}{N\ N^2 - 1} = 1 - \frac{6 \times 98}{8 \times\ 8^2 - 1}$$

$$= 1 - \frac{6 \times 98}{8 \times 63} = 1 - \frac{588}{504} = 1 - 1$$

$$R = -0.167$$

Since rank correlation coefficient is -0.167, the relationship between X and Y is low degree negative.

Check your progress
Find rank correlation coefficient for the following data

1)

Rank by Judge A	7	6	5	8	3	1	2	4
Rank by Judge B	6	8	3	7	1	2	4	5

Ans: 0.76

2)

X	75	88	95	70	60	80	81	50
Y	120	134	150	115	110	140	142	100

Ans: 0.929

munotes.in

2.5 THE LAW OF LARGE NUMBERS

The law of large numbers is one of the most important theorems in probability theory. It stales that, as a probabilistic process is repeated a large number of times, the relative frequencies of its possible outcomes will get closer and closer to their respective probabilities. The law demonstrates and proves the fundamental relationship between the concepts of probability and frequency.

In 1713, Swiss mathematician Jakob Bernoulli proved this theorem in this book. It was later refined by other noted mathematicians, such as Pafnuty Chebyshev.

The law of large numbers shows that if you take an unpredictable experiment & repeat it enough times, you will end up with its average. In technical terms, if you have repeated, independent trials, with a probability of success P for each trial, the percentage of successes that differ from P converge to 0 as the number of trials n tends to infinity. In more simple words, if you repeated an experiments many times you will start to see a pattern and you will be able to figure out probabilities.

For e.g. throw a die and then we will get a random number (1, 2, 3, 4, 5, 6). If we throw if for 100,000 times and we will get an average of 3.5 - which is the expected value.

Another example is of tossing a coin 1, 2, 4, 10, etc. times, the relative frequency of heads can easily happen to be away from the expected 50%. That is because 1, 2, 4, 10,… are al small number. On the other hand, if we tossed a coin for 1000 or 100000 times, then the relative frequency will be very close to 50% since 1000 and 100000 are large numbers.

Weak Law of large numbers :

The Law of Large number is sometimes called the Weak Law of Large numbers to distinguish it from the Strong Law of Large numbers. The two versions of the Law are different depending on the mode of convergence. The weak law is weaker than the sample mean converges to the expected mean in mean square and in probability. The strong law of large numbers is where the sample mean M converges to the expected mean μ with probability.

2.6 REFERENCE

- S-Shyamala and Navdeep Kaur, ‘Introduce too y Econometrics’.
- Neeraj R, Hatekar, ‘Principles of Econometrics : An Introduction us in, R’

❖❖❖❖

MODULE 2

3

TEST OF HYPOTHESIS: BASIC CONCEPTS AND PROCEDURE

Unit Structure:

3.0 OBJECTIVES

- To understand the meaning of hypothesis testing.
- To understand the basic concepts of hypothesis testing.
- To understand the procession and procedure of hypothesis testing.

3.1 INTRODUCTION

Hypothesis is the proposed assumption explanation, supposition or solution to be proved or disproved. It is considered as main instrument in research. It stands for the midpoint in the research. If hypothesis is not formulated researcher cannot progress effectively. The main task in research is to test its record with facts. If hypothesis is proved the solution can be formed and if it is not proved then alternative hypotheses needs to be formulated and tested.

So, with hypothesis formulated it will help up to decide the type of data require to be collected.

The important function in research is formulation of hypothesis. The entire research activity is directed towards making of hypothesis. Research can begin with well formulated hypothesis or if may be the end product in research work. Hypothesis gives us guidelines for an investigation to the basis of previous available information. In absence of this research will called underquired data

munotes.in

and may eliminate required one. Thus hypothesis is an assumption which can be put to test to decide its validity.

3.2 HYPOTHESIS TESTING

In business research and social science research, different approaches are used to study variety issues. This type of research may be format or informal, all research begins with generalized idea in form of hypothesis. A research question is usually there. In the beginning research effort are made for area of study or it may take form of question abut relationship between two or more variable. For example do good working conditions improve employee productivity or another question might be now working conditions influence the employees work.

3.3 BASIC CONCEPTS IN HYPOTHESIS TESTING

Basic concepts in the context of testing of hypotheses need to be explained. Those are:

3.3.1 Null and Alternative hypotheses:

In the context of statistical analysis, we often talk about null hypothesis and alternative hypothesis. If we are to compare method A with method B about its superiority and if we proceed on the assumption that both methods are equally good, then this assumption is termed as the null hypothesis. As against this, we may think that the method A is superior or the method B is inferior, we are then stating what is termed as alternative hypothesis. The null hypothesis is generally symbolized as H0 and the alternative hypothesis as Ha. Suppose we want to test the hypothesis that the population mean (μ) is equal to the hypothesized mean (μH0) = 100. Then we would say that the null hypothesis is that the population mean is equal to the hypothesized mean 100 and symbolically we can express as:

100 : 0 0 H H

If our sample results do not support this null hypothesis; we should conclude that something else is true. What we conclude rejecting the null hypothesis is known as alternative hypothesis. In other words, the set of alternatives to the null hypothesis is referred to as the alternative hypothesis. If we accept H0, then we are rejecting Ha and if we reject H0, then we are accepting Ha. For 100: 0 0 H H , we may consider three possible alternative hypotheses as follows:

If a hypothesis is of the type 0 H , then we call such a hypothesis as simple (for specific) hypothesis but if it is of the type

0 H or 0 H or 0 H then we call it a composite (or nonspecific) hypothesis.

Alternative hypothesis	To be read as follows
$Ha : \mu \neq \mu H_0 \neq 100$	The alternative hypothesis is that the population mean is equal to 100 i.e., it may be more or less than 100.
$H_a : \mu > \mu H_0$	The alternative hypothesis is that the population mean is greater than 100.
$H_a : \mu < \mu H_0$	The alternative hypothesis is that the population mean is less than 100.

The null hypothesis and the alternative hypothesis are chose before the sample is drawn (the researcher must avoid the error of deriving hypotheses from the data that he collects and then testing the hypotheses from the same data.) In the choice of null hypothesis, the following considerations are usually kept in view:
1) Alternative hypothesis is usually the one which one wishes to prove and the null hypothesis is the one which one wishes to disprove. Thus, a null hypothesis represents the hypothesis we are trying to reject and alternative hypothesis represents all other possibilities.

2) If the rejection of a certain hypothesis when it is actually true involves great risk, it is taken as null hypothesis because then the probability of rejecting it when it is true is a (the level of significance) which is chosen very small.

3) Null hypothesis should always be specific hypothesis i.e., it should not state about or approximately a certain value.

Generally, in hypothesis testing we proceed on the basis of null hypothesis, keeping the alternative hypothesis in view. Why so? The answer is that on the assumption that null hypothesis is true, one can assign the probabilities to different possible sample results, but this cannot be done if we proceed with the alternative hypothesis. Hence, the use of null hypothesis (at times also known as statistical hypothesis) is quite frequent.

3.3.2 Parameter and Statistic:

The main objective of sampling is to draw inference about the characteristics of the population on the basis of a study made on the units of a sample. The statistical measures calculated from the numerical data obtained from population units are known as Parameters. Thus, a parameter may be defined as a characteristic of a population based on all the units of the population. While the statistical measures calculated from the numerical data obtained

from sample units are known as Statistics. Thus a statistic may be defined as a statistical measure of sample observation and as such it is a function of sample observations. If the sample observations are denoted by x1, x2, x3, ……, xn. Then, a statistic T may be expressed as T = f (x1, x2, x3, ……, xn).

Measure	Mean	Variance	Proportion	Unit
Parameter	μ	σ^2	P	Population
Statistics	$\overline{X}$	SD^2	P	Sample

3.3.3 Type I and Type II errors:

In the context of testing of hypothesis, there are basically two types of errors we can make. We may reject H0 when H0 is true and we may accept H0 when in fact H0 is not true. The former is known as Type I error and the latter as Type II error. In other words, Type I error means rejection of hypothesis which should have been accepted and Type II error means accepting the hypothesis which should have been rejected. Type I error is denoted by (alpha) known as error, also called the level of significance of test; and Type II error is denoted by (beta) known as error. In a tabular form the said two errors can be presented as follows:

	Decision	
	Accept H_0	**Reject H_0**
H_0 (True)	Correct decision	Type I error (α error)
H_0 (False)	Type II error (β error)	Correct decision

The probability of Type I error is usually determined in advance and is understood as the level of significance of testing the hypothesis. If type I error is fixed at 5 per cent, it means that there are about 5 chances in 100 that we will reject H0 when H0 is true.

We can control Type I error just by fixing it at a lower level. For instance, if we fix it at 1 per cent, we will say that the maximum probability of committing Type I error would only be 0.01.

But with a fixed sample size, n, when we try to reduce Type I error, the probability of committing Type II error increases. Both types of errors cannot be reduced simultaneously. There is a trade-off between these two types of errors which means that the probability of making one type of error can only be reduced if we are willing to increase the probability of making the other type of error. To deal with this trade-off in business situations, decision makers decide the appropriate level of Type I error by examining

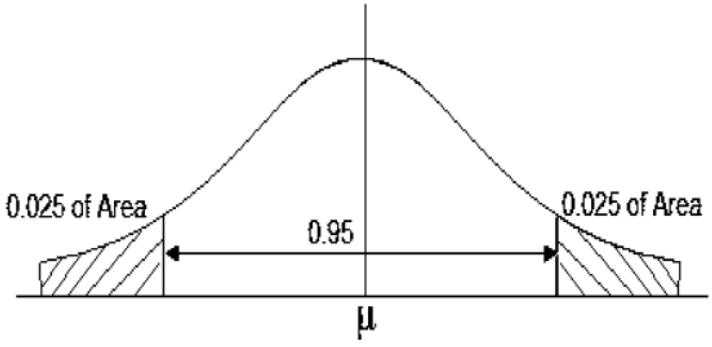
the costs or penalties attached to both types of errors. If Type I error involves the time and trouble of reworking a batch of chemicals that should have been accepted, whereas Type II error means taking a chance that an entire group of users of this chemical compound will be poisoned, then in such a situation one should prefer a Type I error to a Type II error. As a result one must set very high level for Type I error in one's testing technique of a given hypothesis. Hence, in the testing of hypothesis, one must make all possible effort to strike an adequate balance between Type I and Type II errors.

3.3.4 The level of significance:

It is a very important concept in the context of hypothesis testing. We reject a null hypothesis on the basis of the results obtained from the sample. When is such a rejection justifiable?

Obviously, when it is not a chance outcome. Statisticians generally consider that an event is improbable, only if it is among the extreme 5 per cent or 1 per cent of the possible outcomes. To illustrate, supposing we are studying the problem of non attendance in lecture among college students. Then, the entire number of college students is our population and the number is very large. The study is conducted by selecting a sample from this population and it gives some result (outcome). Now, it is possible to draw a large number of different samples of a given size from this population and each sample will give some result called statistic. These statistics have a probability distribution if the sampling is based on probability. The distribution of statistic is called a 'sampling distribution'. This distribution is normal, if the population is normal and sample size is large i.e. greater than 30. When we reject a null hypothesis at say 5 per cent level, it implies that only 5 per cent of sample values are extreme or highly improbable and our results are probable to the extent of 95 per cent (i.e. 1 – .05 = 0.95).

Figure No. 3.1



For example, above Figure shows a normal probability curve. The total area under this curve is one. The shaded areas at both extremes show the improbable outcomes. This area together

is 0.05 or 5 per cent. It is called the region of rejection. The other area is the acceptance region. The percentage that divides the entire area into region of rejection and region of acceptance is called the level of significance. The acceptance region, which is 0.95 or 95 per cent of the total area, is called the level of confidence. These are probability levels. The level indicates the confidence with which the null hypothesis is rejected. It is common to use 1 per cent or 5 per cent levels of significance. Thus, the decision rule is specified in terms of a specific level of significance. If the sample result falls within the specified region of rejection, the null hypothesis is rejected at that level of significance. It implies that there is only a specified chance or probability (say, 1 per cent or 5 per cent) that we are rejecting H0, even when it is true. i.e. a researcher is taking the risk of rejecting a true hypothesis with a probability 0.05 or 0.01 only. The level of significance is usually determined in advance of testing the hypothesis.

3.3.5 Critical region:

As shown in the above figure, the shaded areas at both extremes called the Critical Region, because this is the region of rejection of the null hypothesis H0, according to the testing procedure specified.

Check your progress:

1. Which basic concepts regarding hypothesis testing have you studied?
2. Define:
   i. Null Hypothesis
   ii. Alternative Hypothesis
3. What do you mean by parameter and statistic?
4. What are the Type I and Type II errors?
5. What are level of significance and level of confidence?
6. What is Critical Region?

3.4 PROCESSION OF HYPOTHESES TESTING:

Hypotheses testing is a systematic method. It is used to evaluate the data collected. This serve as aid in the process of decision making, the testing of hypotheses conducted through several steps which are given below.

mynotes.in

a. State the hypotheses of interest
b. Determine the appropriate test statistic
c. Specify the level of statistical significance.
d. Determine the decision rule for rejecting or not rejecting null hypotheses.
e. Collect the data and perform the needed calculations.
f. Decide to reject or not to reject the null hypotheses.

In order to provide more details on the above steps in the process of hypotheses testing each of test will be explained here with suitable example to make steps easy to understand.

1. Stating the Hypotheses

In statistical analysis of any research study if includes at least two hypotheses one is null hypotheses and another is alternative hypotheses.

The hypotheses being tested is referred as the null hypotheses and it is designated as HO. It is also referred as hypotheses of difference. It should include a statement which has to be proved wrong.

The alternative hypotheses present the alternative to null hypotheses. It includes the statement of inequality. The null hypotheses are and alternative hypotheses are complimentary.

The null hypothesis is the statement that is believed to be correct through analysis which is based on this null hypotheses. For example, the null hypotheses might state the average are for entering management institute is 20 years. So average age for institute entry = 20 years

2. Determining Appropriate Test Statistic

The appropriate test statistic which is to be used in statistic, which is to be used in statistical hypotheses testing, is based on various characteristics of sample population of interest including sample size and distribution.

The test statistic can assume many numerical values. As the value of test statistic has significant on decision one must use the appropriate statistic in order to obtain meaningful results. The formula to be used while testing population means is.

Z - test statistic, x - mean of sample μ - mean of population, σ - standard deviation, n – number of sample.

3. The Significance Level

As already explain, null hypothesis can be rejected or fail to reject null hypotheses. A null hypothesis that is rejected may in really be true or false.

A null hypothesis that fails to be rejected may in reality be true or false. The outcome that a researcher desires is to reject false null hypotheses or fail to reject true null hypotheses. However there is always possibility of rejecting a true hypotheses or failing to reject false hypotheses.

Type I and Type II Errors

Type I: error is rejecting a null hypotheses that is true

Type II: Error is failing to rejected a false null hypotheses

Table

	DECISION	
	Accept Ho	Reject Ho
Ho (True)	Correct Decision	Type I Error
Ho (True)	Type II Error	Correct Decision

The probability of committing a type I error is termed as a A and Type II error is termed as B.

4. Decision Rule

Before collection and analyses of data it is necessary to decide under which conditions the null hy7potheses will be rejected or fail to he rejected. The decision rule can be stated in terms of computed test statistics or in probabilistic terms. The same decision will he applicable any of the method so selected.

5. Data Collection and Calculation Performance

In research process at early stage method of data collection is decided. Once the research problem is decided that immediately decision in respect of type and sources of data should be taken. It must clear that fact that, which type of data will be needed for the purpose of the study and now researcher has a plan to collect required data.

The decision will provide base for processing and analysing of data. It is advisable to make use of approved methods of research for collecting and analysing of data.

6. Decision on Null Hypotheses

The decision regarding null hypotheses in an important step in the process of the decision rule.

Under the said decision rule one has to reject or fail to reject the null hypotheses. If null hypotheses is rejected than alternative hypotheses can be accepted. If one fails to reject null hypotheses one can only suggest that null hypotheses may be true.

7. Two Failed and One Failed Tests

In the case of testing of hypotheses, above referred both terms are quite important and they must be clearly understood. A two failed test rejects the null hypotheses.

a. if sample mean is significantly
b. higher or lower than the
c. hypothesized value of mean of the population
d. such a test is appropriate, when the null hypotheses is some specified value and the alternate hypotheses is a value not equal to the specified value and the alternative hypotheses is value not equal to the specified value of null hypotheses.

3.5 PROCEDURE FOR TESTING OF HYPOTHESES:

Testing of hypotheses mean to decide the validity of the hypotheses on the basis of the data collected by researcher. In testing procedure we have to decide weather null hypotheses is accepted or not accepted.

This requirement conducted through several steps between the cause of two action i.e. relation or acceptance of null hypothesis. The steps involved in testing of hypotheses are given below.

1. Setting up of Hypotheses

This step consist of hypotheses setting. In this step format statement in relation to hypotheses in made. In traditional practice instead of one, two hypotheses are set. In case if one hypotheses is rejected than other hypotheses is accepted. Hypotheses should be clearly stated in respect of the nature of the research problem.

There are hypotheses are.
a. Null hypotheses and
b. Alternative hypotheses.

Acceptance or rejection of hypotheses is based on the sampling information. Any sample which we draw from the population will vary from it therefore it is necessary to judge whether there difference are statistically significant or insignificant.

The formulation of hypotheses is an important step which must be accomplished and necessary care should be taken as per the requirement and object of the research problem under construction.

This should also specify the whether one failed or two failed test will be used.

mynotes.in

2. Selecting Statistical Technique

In this stage we will make selection of statistical technique which are going to he used. There are various statistical test which are being used in testing of hypotheses. There tests are

Z – Test
T – Test
F – Test
X2

It is the job of the researcher to make proper selection of the test.

Z- Test is used when hypotheses is related to a large sample. (30 or more)

T- Test is used when hypotheses is related to small sample (Less than 30)

The selection of test will be dependent on various consideration like, variable involved, sample size, type of data and whether samples are related or independent.

3. Selecting Level of Significance

This stage consists of making selection of desired level of significance. The researcher should specify level of significance because testing of hypotheses is based on pre-determined level of significance. The rejection or retention of hypothesis by the researcher is also based on the significance level.

The level of significance is generally expressed in percentage from such as 5% or 1%, 5% level of significance is accepted by the researcher, it means he will be making wrong decision about 5% of time. In case if hypotheses is reject at this level of 5% he will be entering risk hypotheses rejection ???out of 100 occasions.

The following factors may affect the level of significance.
- The magnitude difference between sample mean
- The size of sample
- The validity of measurement

4. Determining Sampling Distribution

The next step after deciding significance level in testing of hypothesis is to determine the appropriate sampling distribution. It is, normal distribution and ‘t’ – distribution in which choice can be excised.

nunotes.in

5. Selecting Sample and Value

In this step random sample is selected and appropriate value is computed from the sample data relating to the test statistic by utilizing the relevant distribution.

6. Performance Computation

In this step calculation of performance is done. The calculation includes testing statistics and standard error.

A hypothesis is tested for the following four possibilities, that the hypotheses is

a- True, but test lead to its rejection
b- False, but test lead to its acceptance
c- True, but test lead to its acceptance
d- False, but test lead to its rejection

Out of the above four possibilities a and b lends to wrong decision. In this case a lends to Type I error and, b lends to Type II error.

7. Statistical Decision

Thus is the step in which we have to draw statistical decision involving the acceptance or rejection of hypotheses.

This will be dependent on whether the calculated value of the test falls in the region of acceptance or in the region of rejection at given significance level.

If hypotheses is tested at 5% level and observed set recorded the possibilities less than 5% level than we considered difference between hypothetical parameter and sample statistics is significant.

3.6 REFERENCE

- S-Shyamala and Navdeep Kaur, 'Introduce too y Econometrics'.
- Neeraj R, Hatekar, 'Principles of Econometrics : An Introduction us in, R'

4

TEST OF HYPOTHESIS: VARIOUS DISTRIBUTION TEST

Unit Structure:

4.0 OBJECTIVES

- To understand the various distribution tests of hypothesis testing.
- To understand the uses of t – test.
- To understand the uses of F test and Chi-square test.

4.1 INTRODUCTION

The test of significance used for hypothesis testing is of two types the parametric and non-parametric test.

The parametric test is more powerful, but they depend on the parameters or characteristics of the population. They are based on the following assumptions.

1. The observations or values must be independent.
2. The population from which the sample is drawn on a random basis should be normally distributed.
3. The population should have equal variances.
4. The data should be measured at least at interval level so that arithmetic operations can be used.

mumotes.in

4.2 TESTING OF HYPOTHESIS USING VARIOUS DISTRIBUTION TEST

A. The Parametric Tests:

a) The Z – Test

Prof. R.A. fisher has develop the Z Test. It is based on the normal distribution. It is widely used for testing the significance of several statistics such as mean, median, mode, coefficient of correlation and others. This test is used even when binominal distribution or t distribution is applicable on the presumption that such a distribution lends to approximate normal distribution as the sample size (n) become larger.

b) The T – Test

The T – Test was developed by W.S. Gossel around 1915 since he published his finding under a bon name 'student', it is known as student's t – test. It is suitable for testing the significance of a sample man or for judging the significance of difference between the mams of two samples, when the samples are less than 30 in number and when the population variance is not known. When two samples are related, the paired t – test is used. The t – test can also be used for testing the significance of the coefficient of simple and partial correlation.

In determining whether the mean of a sample drawn from a normal population deviates significantly from a stated value when variance of population is unknown, we calculate the statistic.

$$t = \frac{\bar{x} - \mu}{s} \times \sqrt{\eta}$$

Where,

$\bar{x}$ = the mean of sample

μ = the actually or hypothetical mean of population

n = the sample size

s = standard deviation of the samples

$$s = \sqrt{\frac{\Sigma\left(x - \bar{x}\right)^2}{\eta - 1}}$$

Example

Ten oil tins are taken at random from an automatic filling machine the mean weight of the 10 tins is 15.8 kg and standard deviation 0.5 kg. Does the sample mean differ significantly from the intended weight of 16 kg?
(given for ν, to 0.05 – 2.26)

Solution:

Let us make the hypothesis tat the sample mean does not differ significantly from the intended weight of 16 kg applying t – test .

$$t = \frac{|\bar{x} - \mu|}{s}$$

$$\bar{x} = 15.8 \quad \mu = 6.5 = 0.05 \quad \eta = 10$$

$$t = \frac{|15.0 - 16|}{0.5} \times \sqrt{10}$$

$$= \frac{0.2 \times 3.162}{0.5}$$

$$= 1.27$$

$$\upsilon = \eta - 1 = 10 - 1 = 9$$

$$\nu = 9$$

t 0.05=2.26

For

The calculated value of t is less than the table value. The hypothesis is accepted.

3. The f- test

The f – test is based on f – distribution (which is a distribution skewed to the right, and tends to be more symmetrical, as the number of degrees of freedom in the numerator and denominator increases)

The f- test is used to compare the variances of two independent sample means at a time. It is also used for judging the significance of multiple correlation coefficients.

B The Non-parametric Tests

The non-parametric tests are population free tests, as they are not based on the characteristics of population. They do not specify normally distributed population or equal variances. They are easy to understand and to use.

The important non parametric tests are:

- The chi-square test
- The median test
- The Mann-Whitney U test
- The sign test
- The Wilcoxin matched –Paris test
- The Kolmogorow Smornov test.

Mumtes.in

The Chi-Square Test (x^2)

The Chi-Square test is the most popular non-parametric test of significance in social science research. It is used to make comparisons between two or more nominal variables. Unlike the other test of significance, the chi-square is used to make comparisons between frequencies rather than between mean scares. This test evaluated whether the difference between the observed frequencies and the expected frequencies under the null hypothesis can be attributed to chance or actual population differences. A chi-square value is obtained by formula.

$$\kappa^2 = \sum \left[\frac{(\text{difference of actual and expected frequencies})^2}{\text{expected frequencies}} \right]$$

$$\kappa^2 = \sum \frac{(f_A - f_e)^2}{f_r}$$

Where,

$k^2 =$ chi-square

$f_A =$ observed or actual frequency

$f_e =$ expected frequency

$k^2 =$ can also determined with the help of the following formula.

$$\kappa^2 = \sum \left(\frac{f_A}{f_e} \right) - N$$

N = total of frequencies

Example,

Weight of 7 persons is given as below:

Persons	Weight
1	40
2	45
3	50
4	60
5	55
6	52
7	48

In this information we can say, variance of distribution of sample of 7 persons was drawn is equal to weight of 30 kg.

Test this at 5% of 1% level of significance.

munotes.in

Solution:

Above information we will workout variance of sample data.

S. No.	X1	w/kg	(x,-x)	$(x,-x)^2$
1	40	-16	100	
2	45	-05	25	
3	50	+0	0	
4	60	+10	100	
5	55	+05	25	
6	52	+02	4	
7	48	-02	4	

$n = 7,\ EX = 350,\ \Sigma\left(x_1 - \bar{x}\right)^2 = 258$

$$\bar{x} = \frac{\Sigma x1}{n} = \frac{350}{7} = 50$$

$$\delta = \sqrt{\frac{\Sigma\left(x_1 - x^2\right)^2}{x - 1}} = \frac{\sqrt{258}}{7 - 1} = 43$$

$$x^2 = \frac{43}{30}(7 - 1) = 8.6$$

Degree of freedom is (n-1) = (7-1) = 6

At 5% Ye level of significance x^2 =12.592
1% level = 16.812
Value are greater than x^2 =8.6

So we accept null hypotheses and variance at both 5 and 1 pe level is significant. So sample of 30 kg is taken from the population.

The standard normal distribution and its application:

Normal distributions do not necessarily have the same means and standard deviations. A normal distribution with a mean of 0 and a standard deviation of 1 is called a standard normal distribution. It is centred at zero and the degree to which a given measurement deviates from the mean is given by the standard deviation. This distribution is also known as the Z - distribution.

A value on the standard normal distribution is known as a standard deviations above or below the mean that specific observation falls. For example, a standard score of 1.5 indicates that the observation is 1.5 standard deviation above the mean. On the other hand, a negative score represents a value below the average. The mean has a Z-score of 0.

mnotes.in

4.3 STANDARDIZATION : CALCULATING Z - scores

The process of standardization allows to compare observations and calculate probabilities across different populations. i.e. it allows to take observations drawn from normally distributed populations that have different means and standard deviations and place then on a standard scale. To standardize the data, we need to convert the raw measurements into Z-scores.

To calculate the standard score for an observation, following formula can be used.

$$Z = \frac{X - \mu}{\sigma}$$

$X \rightarrow$ raw value of the measurement of interest

μ and $\sigma \rightarrow$ parameters for the population from which the observations is drawn.

Let us discuss it with an example of mangoes and Apples. Let's compare their weights. Mangoes weights 110 grams and an Apples weights 100 grams. By comparing mere their raw value we can observe that the mango weights more than the Apple. Now we will compare their standard scores. Assuming that the weights of mangoes and Apples follow a normal distribution with the following parameter values :

	Mangoes	Apples
Mean weight grams	100	140
Standard deviation	15	25

We will use these value to get Z - score :

Mangoes = (110 - 100) / 15 = 0.667
Apples = (100 - 140) / 25 = -1.6

The Z - score for the Mangoes is (0.667) positive which means that Mangoes weight more than the average Apple. It is not an extreme value by any means, but it is above average for mangoes. On the other hand the Apples has fairly negative Z - score (- 1.6). It is much below the mean weight for apples.

To find areas under the curve of a normal distribution for it, we will use Z - score table.

Let's take the Z-score for mango (0.667) and use it to determine its weight percentile. A percentile is a proportion of a population that falls below a specific value. To determine percentile,

we need to find the area that corresponds to the range of Z scores that are less than 0.667. The closet value in Z - score table to it is 0.65. The table value indicates that the are of the curve between -0.65 and +0.65 is 48.43%. But we want the area that is less than a Z-score of 0.65.

The two halves of the normal distribution are mirror images of each other. So if the area for the interval from -0.65 and +0.65 is 48.43%, then the range from 0 to +0.65 must be half of that

$$\frac{48.43}{2} = 24.215\%.$$

We also know that the area for all scores less than zero is half (50%) of the distribution.

Therefore the area for all scores upto 0.65, 0.65 = 50% + 24.215% = 74.215%

So, the Mango is at approximately the 74th percentile.

Students t distribution:
In case of large sample test Z - test is

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2 / n}} : \; N(0,1)$$

If $\sigma^2 \rightarrow$ population variance is unknown then sample variance S^2 is used and normal test is applied. But when sample is small, the unbiased estimate of the population variance is used i.e.

Unbiased Variance of sample $S^2 = \frac{\sum (X - \bar{X})^2}{n-1}$

Biased Variance $S^2 = \frac{\sum (X - \bar{X})^2}{n}$

In small samples σ^2 is replaced by S^2 and not by S^2.

Student t : If $x_1, x_2, ...x_n$ is a random sample of size n from a normal population with mean μ and variance σ^2 then students t statistic is given by

Where $\bar{X}$ = Sample mean
μ = Population mean

mynotes.in

$$t = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}}$$

$$S^2 = \frac{\sum (X - \bar{X})^2}{n-1}$$

4.4 USES OF T - TEST

1) t - test for single Mean:

It is used to test the hypothesis that the population mean μ has specified value of μ_0 when population standard deviation (σ) is not known and $n \leq 30$ we use t - Test.

$$t = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}}$$

If follows t - distribution with (n - 1) degree of freedom

$$S^2 = \frac{\sum (X - \bar{X})^2}{n-1}$$

Steps for applying t - Test :

a) Set up the null hypothesis $H_0 : \mu = \mu_0$

alternative hypothesis $H_1 : \mu \neq \mu_0$ (Two tailed test)

$H_1 : \mu > \mu_0$ or $\mu < \mu_0$ (one failed test)

b) Find $S^2 = \frac{\sum (X - \bar{X})^2}{n-1}$ or $\sum (X - \bar{X})^2 = (n-1)S^2$ where $S^2 \rightarrow$ unbiased variance.

Biased Variance $S^2 = \frac{\sum (X - \bar{X})^2}{n}$ or $\sum (X - \bar{X})^2 = nS^2$

where $S^2 \rightarrow$ biased variance.

Since $\sum (X - \bar{X})^2 = (n-1)S^2 = nS^2$

or $\frac{S^2}{n} = \frac{S^2}{n-1}$ $\therefore S^2 = \frac{n}{n-1} \times S^2$

mnotes.in

c) Use the values in t - Test and compare calculated value with table value for $V = n - 1$ degree of freedom.

d) If calculated value is greater than table value accept H_1 and vice versa.

Suppose a group of 5 students has weight 42, 39, 48, 60 and 41 kg. Can it be said that this sample has come from the population whose mean weight is 48 kg?

Solution :

	Weight (X)	$(X - \bar{X})$
1	42	16 $(42 - 46)^2$
2	39	49 $(39 - 46)^2$
3	48	4 $(48 - 46)^2$
4	60	196 $(60 - 46)^2$
5	41	25 $(41 - 46)^2$
$n = 5$	$\sum X = 230$	$\sum (X - \bar{X})^2 = 290$

$$\bar{X} = \frac{\sum X}{n} = \frac{230}{5} = 46$$

$$S^2 = \frac{\sum (X - \bar{X})^2}{n - 1} = \frac{290}{5 - 1} = \frac{290}{4} = 72.5$$

Where $H_0 : \mu = 48$ (No significant difference between sample mean and population mean)

$H_1 : \mu \neq 48$ (Significant difference between sample and population mean)

$$t^* = \frac{|\bar{X} - \mu|}{\sqrt{\frac{S^2}{n}}} = \frac{|146 - 48|}{\sqrt{\frac{72.5}{5}}}$$

$$|t|^* = \frac{2}{\sqrt{14.5}} = \frac{2}{3.81} = 0.525$$

Table value of t at 5% level of significance for two tailed test for V = 5 - 1 = 4 is 2.776.

$t^* < \frac{0.05}{2}, V = 4$ we accept H_0 and conclude that the mean weight of the population is 48 kg.

ii) t - Test for difference of means:

Suppose two independent samples have been taken from two normal population having the same mean, the population variance are also equal & hypothesis $H_0 : \mu_x = \mu_y$ where two samples have come from the normal population with the same means.

$$t = \frac{\left|\bar{X} - \bar{Y}\right|}{S \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\bar{X} = \frac{\sum X}{n}, \bar{Y} = \frac{\sum Y}{n}, S^2 = \frac{\sum \left(X - \bar{X}\right)^2 + \sum \left(Y - \bar{Y}\right)^2}{n_1 + n_2 - 2}$$

Let us discuss this with the help of the following example.

In an examination 12 students in Class A had a mean score of 78 and standard deviation is 6 whereas 15 students in Class B had a mean score of 74 with standard deviation 8. Is the significant difference between the means of the two classes?

Solution :

$n_1 = 12 \bar{X} = 78 \quad S_x^2 = 6$

$n_2 = 15 \bar{Y} = 74 \quad S_y^2 = 8$

$H_0 : \mu_x = \mu_y$ (no significant difference between the means of the two classes)

$H_1 : \mu_x \neq \mu_y$ (Significant difference between the means of the two classes)

$$t^* = \frac{\left|\bar{X} - \bar{Y}\right|}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$S^2 = \frac{\sum \left(X - \bar{X}\right)^2 + \sum \left(Y - \bar{Y}\right)^2}{n_1 + n_2 - 2}$$

$$S_x^2 = \frac{\sum \left(X - \bar{X}\right)^2}{n_1} \; or \sum \left(X - \bar{X}\right)^2 = n_1 S_x^2$$

Similarly $\sum \left(Y - \bar{Y}\right)^2 = n_2 S_x^2$

munotes.in

$$S^2 = \frac{n_1 S_x^2 + n_2 S_y^2}{n_1 + n_1 - 2} = \frac{12(6)^2 + 15(8)^2}{12 + 15 - 2}$$

$$S^2 = \frac{432 + 960}{25} = \frac{1392}{25} = 55.68$$

$$S = 7.46$$

$$t = \frac{|78 - 74|}{7.46\sqrt{\frac{1}{12} + \frac{1}{15}}} = \frac{4}{7.46(0.15)}$$

$$t = \frac{4}{1.15} = 3.48$$

Table value of t for $V = n_1 + n_2 - 2 = 25$ at 5% level of significance for 2 tailed tests is 2.064.

$$t^7 > \frac{^t 0.05}{2} v = 25$$

i.e. $3.48 > 2.064$

Therefore Accept H_1 and conclude that there is significant difference between the sample mean.

iii) t - Test for difference of means with dependent samples (paired t - Test):

This test is applicable when two samples are dependent. Following are the conditions to apply this test :

$\Rightarrow$ Two samples should be of equal size $n_1 = n_2$

$\Rightarrow$ Sample observations of X and Y are dependent in pairs.

The formula for paired t - Test is

$$t^* = \frac{\bar{d}}{\sqrt{\frac{S^2}{n}}}$$

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

$d_i \to x - y$ (x & $y \to$ sample observations) i.e. difference between each matched pair.

Suppose, a test is conducted for 5 students in a coaching centre to know the subject knowledge of the students before and after tutoring for one month.

mumotes.in

Students	1	2	3	4	5
Results before test	110	120	123	132	125
Result after test	120	118	125	126	121

Is there any change in result after tutoring?

Solution :

X	Y_i	$d_i = X - Y$	d_i^2
110	120	- 10	100
120	118	2	4
123	125	- 2	4
132	136	- 4	16
125	121	4	16
		$\sum d_i = -10$	$\sum d_i^2 = 140$

$$\bar{d} = \frac{\sum d_i}{n} = \frac{-10}{5} = -2$$

$$S^2 = \frac{1}{n-1}\left[\sum d_i^2 - \left(\frac{\sum d_i}{n}\right)^2\right]$$

$$= \frac{1}{5-1}\left[140 - \frac{(-10)^2}{5}\right]$$

$$= 30$$

$H_0 : \mu_x = \mu_y$ (mean score before and after tutoring are same)

$H_1 : \mu_x \neq \mu_y$ (mean score before & after tutoring are not same)

$$t = \frac{|\bar{d}|}{\sqrt{\frac{S^2}{n}}} = \frac{2}{\sqrt{\frac{30}{5}}} = 0.816$$

Table value of t at 5% level of significance (2 tailed test) for $n-1 = v = 5-1 = 4$ is 2.776.

$$t^* < \frac{t\,0.05}{2}, v = 4$$

i.e. $0.816 < 2.776$

munotes.in

Therefore H_0 is accepted and conclude that there is no significant difference in score of the students after one month of tutoring.

iv) t - Test for significance of an observed sample correlation coefficient:

When r is a sample correlation & P is correlation for the population which is unknown, t - Test is applied to test the significance of correlation coefficient.

$$t^* = \frac{r}{S.E_r}$$

$$S.E_r = \sqrt{\frac{1-r^2}{n-2}}$$

Let us assume that a coefficient of correlation of sample of 27 pair of observation is 0.42. Is it likely that variables in the population are not correlated?

Solution :
In our example,

Let $H_0 : P = 0$ (the variables in the population are uncorrelated)
$H_1 : P \neq 0$ (variables in the population are correlated)

$$t^* = \frac{r}{S.E.r} = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

$$t^* = \frac{0.42}{\sqrt{1-(0.42)^2}}\sqrt{27-2}$$

$$t^* = \frac{0.42}{\sqrt{.8236}} \times \sqrt{25}$$

$$t^* = 2.315$$

$$V = n - 2 = 25$$

Table value of t for 25 degree of freedom is 2.06

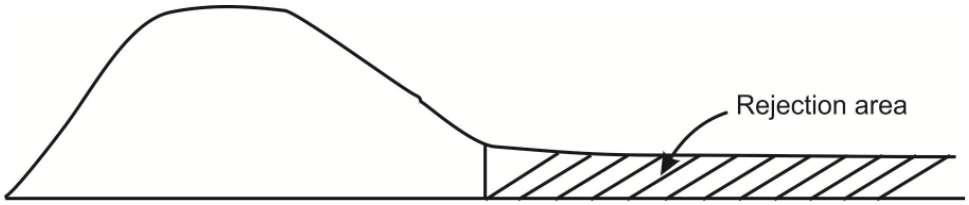
$$\therefore t^* > t0.25 \; for \; V = 25$$

Therefore H_1 is accepted & conclude that variables in the population are correlated.

munotes.in

4.5 F - TEST

F statistic is ratio of two independents chisquare variate divided by their respective degree of freedom. Critical values of F test are based on right tailed test which depends on V_1 (degree of freedom for numerator) and V_2 (degree of freedom for denominator)



F - Test is used to test the equality of population variances. Where

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$ (population variances are same)

$$F = \frac{S_1^2}{S_2^2} = \frac{\text{Larger estimate of population variance}}{\text{Smaller estimate of population variance}}$$

Where S_1^2 and S_2^2 are unbiased estimates of common population variance σ^2 and are gives by

$$S_1^2 = \frac{\sum (X - \bar{X})^2}{n_1 - 1} \text{ and } S_2^2 = \frac{\sum (Y - \bar{Y})^2}{n_2 - 1}$$

Where $V_1 = n_1 - 1$ and $V_2 = n_2 - 1$.

This test is also called variance ratio test

$$S_1^2 = \frac{\sum (X - \bar{X})^2}{n_1 - 1} \text{ and } S_2^2 = \frac{\sum (X - \bar{X})^2}{n}$$

$$S_1^2 (n_1 - 1) = \sum (X - \bar{X})^2$$

$$\left(n_1 - S_1^2\right) = \sum (X - \bar{X})^2$$

R.H.S. are equal. Hence L.H.S. are also equal $S_1^2 (n_1 - 1) = n_1 S_2^2$

Similarly we can find relation between S_2^2 and S_2^2.

Assumption of F - Test

$\Rightarrow$ Sample should be random

mgnotes.in

$\Rightarrow$ Sample observations should be independent
$\Rightarrow$ Sample should be taken from normal population

Let us discuss F-test with the help of the following example.

Suppose Two samples gave the following results.

Sample	Size	Mean	Sum of the squares of deviation from mean
1	10	15	90
2	12	14	108

Test the equality of sample variance.

Solution :

Let $H_0 : \sigma_1^2 = \sigma_2^2$ (Difference in variances of two samples is not significant)

Given $n_1 = 10, n_2 = 12 \sum \left(X - \bar{X}\right)^2 - 90 \sum \left(Y - \bar{Y}\right)^2 = 108$

$$S_1^2 \frac{\sum \left(X - \bar{X}\right)^2}{n_1 - 1} = \frac{90}{10 - 1} = \frac{90}{9} = 10$$

$$S_2^2 \frac{\sum \left(Y - \bar{Y}\right)^2}{n_2 - 1} = \frac{108}{12 - 1} = \frac{108}{11} = 9.82$$

Apply F - Test,

$$F^* = \frac{S_1^2}{S_2^2} = \frac{10}{9.82} = 1.02$$

For $V_1 = n_1 - 1 = 9$ and $V_2 = n_2 - 1 = 11$

$F_{0.05} = 2.90$

Since $F^* < F_{.05}$

There H_0 is accepted and conclude that there is no significant difference in the variance.

muotes.in

4.6 CHI-SQUARE TEST

Properties of x^2 distribution.

1) Moment Generating function x^2 distribution is $\mu_{x^2}(t) = (1-2t)^{-n/2}$ with parameters $\frac{1}{2}$ and $\frac{n}{2}$.
2) Mean of x^2 distribution is 'n'.
3) Variance of x^2 distribution is '2n'
4) Skewness of χ^2 distribution is $\gamma_1 = \sqrt{\frac{8}{n}} > 0$ i.e. χ^2 distribution is positively skewed. But as $n \to \infty, \gamma_1 \to 0$, the distribution becomes normal.
5) Kurtosis of χ^2 distribution is $\gamma^2 = \frac{12}{n} > 0$ i.e. χ^2 distribution is Leptokurtic.
   But as $n \to \infty, \gamma_2 \to 0$, the distribution tends to Mesokurtic.
6) χ^2 distribution tends to normal distribution $n \to \infty$.
7) The sum of independent Chi-square variate is also a chi-square variate.

Application of Chi-square distribution

i) Goodness to fit :
   This test is used to test if the experimental results support a particular hypothesis or theory.

Assuming Null hypothesis that there is no significant difference between the observed and expected frequencies. Chi-square distribution with $V = n-1$ degree of freedom.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where $O_i \to$ observed frequency

$E_i \to$ expected or theoretical frequency.

Steps to compute χ^2 test : -

$\Rightarrow$ Consider null hypothesis H_0 that the theory fits the data well.

$\Rightarrow$ Compute the expected frequencies (E_i) corresponding to the observed frequencies (O_i) under the considered hypothesis.

$\Rightarrow$ Compute $(O_i - E_i)^2$

mnotes.in

$\Rightarrow$ Divide the square of the deviation $(O_i - E_i)^2$ by the corresponding expected frequencies i.e. $(O_i - E_i)^2 / E_i$

$\Rightarrow$ Add the values obtained in the above step to

Calculate:

$$\chi^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

$\Rightarrow$ Calculate degree of freedom i.e. $V = n - 1$

$\Rightarrow$ Find the table value of χ^2 for $n-1$ degree of freedom at certain level of significance.

$\Rightarrow$ Compare the calculated value of χ^2 to the table value, if $\chi^2 < t_{0.05}$ then accept the null hypothesis and conclude that there is good fit between theory and experiment.

$\Rightarrow$ If calculated value of $\chi^2 > t_{0.05}$ then reject the null hypothesis & conclude that the experiment does not support the theory.

Chi-square test can be used under following condition :

1) The sample observations should be independent.
2) $\sum O_i = \sum E_i = N$
3) The total frequency N should be greater than 50 i.e. N > 50
4) No expected frequency should be less than 5. If any expected cell frequency is less than 5 then we cannot use χ^2 test. In that case, we use pooling techniques where we add the frequencies which are less than 5 with succeeding or preceding frequency so that sum process more than 5 and adjust - degree of freedom accordingly.
5) The given distribution should not be replaced by relative frequencies or proportions but the data should be given in original units.

Let us discuss this with the help of an example.

A sample analysis of examination results of 450 final year degree students was made. It is found in the analysis that 200 students have failed, 160 have got pass class, 75 got second class and only 15 students have got first class. Find out whether these figures are consistent with the general final year degree examination result which is in the ratio of 4:2:2:1 for the above mentioned categories respectively.

Solution :

Assuming null hypothesis H_0 that the figure are consistent with the general examination result.

Category	Observed frequency (O_i)	Expected frequencies (E_i)	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
Fail	200	180	400	2.22
Pass	160	135	625	4.63
Second	75	90	225	2.5
First	15	45	900	20
	$\sum O_i = 450$	$\sum E_i = 450$		29.35

Expected frequencies :
Failed : $4/10 \times 450 = 180$
Pass : $3/10 \times 450 = 135$
Second : $2/10 \times 450 = 90$
First : $1/10 \times 450 = 45$

$$\chi^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right] = 29.35$$

d.f. = 4 - 1 =3

Table value of χ^2 at 5% level of significance for df 3 = 7.815. Since calculated χ^2 value is greater than the table value i.e.

$\chi^2 > t_{0.05}$

29.35 > 7.815

H_0 is rejected at 5% level of significance and conclude that the figures are not consistent with the general final year degree examination result.

ii) Chi - square test for independence of Attributes suppose the given population has N items, divided into 'p' mutually disjoint and exhaustive classes. $A_1, A_2, ...A_P$ with respect to the attribute A. So that randomly selected item belongs to one and only one of the attributes $A_1, A_2, ...A_P$. Similarly suppose the population is divided into 'q' mutually disjoint and exhaustive B. So that randomly selected items posseses one and only one of the attributes $B_1, B_2, ...B_q$. The frequency distribution of the items belonging to

munotes.in

the classes $A_1, A_2, ...A_P$ and $B_1, B_2, ...B_q$ can be represented as $(p \times q)$.

Steps for the test -

⇒ Consider null hypothesis that two attributes A and B are independent.

⇒ Compute the expected frequencies (E_i) Corresponding to the observed frequencies (O_i) Expected frequency for $(A_i B_j)$

$$F(A_i B_j) = \frac{(A_i)(B_j)}{n} \text{ where } \begin{bmatrix} i = 1, 2, ..., p \\ j = 1, 2, ..., q \end{bmatrix}$$

⇒ Computer $(O_i .....E_i)^2$

⇒ Divide the square of the deviations $(O_i - E_i)^2$ by the corresponding expected frequency i.e. $(O_i - E_i)^2 / E_i$

⇒ Add the values obtained in the above step to calculate

$$\chi^2 = \sum \left[\frac{(O_i .....E_i)^2}{E_i} \right]$$

⇒ Calculate degree of freedom $= (r-1)(C-1)$

$r =$ No. of rows C = No. of columns

⇒ Compute the calculated value χ^2 with the table value for $(r-1)(C-1)$ degree of freedom at certain level of significance. If the calculated value of χ^2 is greater than the table value of χ^2 the null hypothesis is accepted and vice versa.

Let us discuss this with the help of following example.

The following data on vaccination is collected in a government hospital to find out whether vaccination reduces the severity of attack of influenza.

	Degree of Severity		
	Very Severe	**Severe**	**Mild**
Vaccinated	10	150	240
Not Vaccinated	60	30	10

Use χ^2 - test, to test the association between the attributes.

munotes.in

Solution :

	Observed frequencies			
	Very Severe	**Severe**	**Mild**	**Total**
Vaccinated	10	150	240	400
Not Vaccinated	60	30	10	100
Total	70	180	250	N = 500

Assume the null hypothesis that the two attributes are independent i.e. Vaccine is not effective in controlling the severity of attack of influenza. The expected frequencies are as follows :

Expected Frequencies

	Degree of Severity			
	Very Severe	**Severe**	**Mild**	**Total**
Vaccinated	$\frac{70\times400}{500}$ $=56$	$\frac{180\times400}{500}$ $=144$	$\frac{250\times400}{500}$ $=200$	400
Not Vaccinated	70 - 56 = 14	180 - 144 = 36	250 - 200 = 50	100
Total	70	180	250	N = 500

Compution of Chi square

O_i	E_i	$(O_i-E_i)^2$	$(O_i-E_i)^2/E_i$
10	56	2116	37.786
60	14	2116	151.143
150	144	36	0.25
30	36	36	1
240	200	1600	8
10	50	1600	32
$\sum O_i = 500$	$\sum E_i = 500$	500	230.179

$$d.f = (r-1)(c-1)$$
$$= (2-1)(3-1) = 1\times2 = 2$$

munotes.in

Table value of χ^2 for 2 d.f. at 5% level of significance is 5.99

Computed value of χ^2 is greater than the table value of χ^2, it is highly significant and hence the null hypothesis the rejected. Hence we conclude that both attributes are not independent ant vaccination helps to reduce the severity of attack of influenza.

iii) χ^2 - test for the population variance

To test if the given normal population has a specified variance $\sigma^2 - \sigma^2$, we assume the null hypothesis.

$$H_0 : \sigma^2 = \sigma_0^2$$

If $X_1, X_2, X_3, ..., X_n$ is a random sample of size 'n' from the given population, then under the null hypothesis H_0, the statistic

$\chi^2 = \frac{ns^2}{\sigma_0^2} = \frac{\sum (X - X)^2}{\sigma_0^2}$ follows χ^2 distribution with $(n-1)$ d.f. where $S^2 = \frac{\sum_{i=1}^{n} (X - \bar{X})^2}{n}$ denotes the sample variance.

By comparing the calculated value of χ^2 with the table value for $(n-1)$ d.f. at certain level of significance null hypothesis can be accepted or rejected.

Let us discuss this with the help of following example.

Weight in kgs. Of 10 members in a Gym are given below :
36, 40, 45, 55, 47, 44, 56, 48, 53, 46

Can it be said that population variance is 20 square kg?

Solution :

Assume null hypothesis $H_0 : \sigma^2 = 20$ against the alternative hypothesis $H_1 : \sigma^2 > 20$

Weight (in kg) (X_1)	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
36	- 11	121
40	- 7	49

munotes.in

45	- 2	4
55	8	64
47	- 0	0
44	- 3	9
56	9	81
48	1	1
53	6	36
46	- 1	1
$\sum X = 470$		$\sum (X_i - \bar{X})^2 = 366$

$$\bar{X} = \frac{\sum X}{n} = \frac{470}{10} = 47$$

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{m}$$

$$nS^2 = \sum (X_i - \bar{X})^2 = 366$$

$$\chi^2 = \frac{nS^2}{\sigma_0^2} = \frac{366}{20} = 18.3$$

Degree of freedom $= n - 1 = 10 - 1 = 9$

Table value of χ^2 for 9 .d.f. at 5% level of significance is 16.92.

Since calculated χ^2 is greater than table value of χ^2 at 5% level of significance, null hypothesis is rejected and conclude that the population variance is not 20 sq.km.

4.7 REFERENCE

- S-Shyamala and Navdeep Kaur, 'Introduce too y Econometrics'.
- Neeraj R, Hatekar, 'Principles of Econometrics : An Introduction us in, R'

munotes.in

MODULE 3

5

ESTIMATED LINEAR REGRESSION EQUATION AND PROPERTIES OF ESTIMATORS

Unit Structure:

5.0 OBJECTIVES

- To understand the concepts of simple linear regression model.
- To understand the various test in regression.

5.1 INTRODUCTION

Linear regression models are used to predict the relationship between two variables. The factors which is being predicted is called the dependent variable and the factors which are used to predict the value of the dependent variable are called the independent variables. So in this simple linear regression model, a straight line approximates the relationship between the dependent variable and the independent variable.

Assuming the two factors that are involved in simple linear regression analysis are X and Y then the equation that describes how Y is related to X is represented in the following formula for a simple Linear Regression Model.

$$Y = \beta_0 + \beta X + \mu$$

Where, β_0 and β_1, are parameters

This equation contains an error term which is represented by μ. It is used to account for the variability in Y that cannot be explained by the linear relationship between X and Y.

knunotes.in

For e.g. In economic theory, Consumption (C) is determined by income (Y)

$$\therefore C = f(Y) = \beta_0 + \beta_0 Y$$

Here we assume that consumption depends only on income (other determinants of consumption taken to be constant). But in real world such exact relationship between C and Y never exists.

Therefore we add 'μ' an error term in the equation where μ is a random variable called residual error. The error arises from the measurement errors in Y or imperfections in the specification of the function $f(Y)$.

So the standard form of the simple linear regression model is

$Y_i = f(X_i) + \mu_i$ ..... (1)

$Y_i = \beta_0 + \beta_1 X_i + \mu_i$ .....(2)

where $Y_i \rightarrow$ dependent variable

$X_i \rightarrow$ explanatory or independent variable

$\beta_1 \rightarrow$ slope parameter

$\beta_0 \rightarrow$ intercept

It is based on the assumption that

a) the relationship between X & Y is linear
b) Assumption about the random disturbance (μ).

A regression line can show a positive linear relationship, a negative linear relationship and no relationship.

i) No relationship - The line in the graph in a simple linear regression is flat (not sloped). There is no relationship between the two variables.

ii) Positive relationship - Exists when the regression line slopes upward with the lower end of the line at the y-intercept (axis) of the graph and the upper end of the line extending upward into the graph, away from the X-intercept (axis). There is a positive linear relationship between the two variables representing that as the value of one variable increases, the value of the other also increases.

iii) Negative relationship - The regression line slopes downwards with the upper end of the line at the y-intercept (axis) of the graph and the lower end of the line extending downward into the graph field, toward the X intercept (axis). There is a negative

munotes.in

linear relationship between the two variables i.e. as the value of one variable increases, the value of the other decreases.

5.2 THE ESTIMATED LINEAR REGRESSION EQUATION

If the parameters of the population were unknown, the simple linear regression equation could be used to compute the mean value of y for a known value of X.

$$E(Y) = \beta_0 + \beta_1 X + \mu$$

In practice, however parameter values are generally unknown so they must be estimated by using data from a sample of the population. The population parameters are estimated by using sample statistics. They are represented by β_0 and β_1 when these sample statistics are substituted for the population parameters, the estimated regression equation is formed as following.

$$\left(\hat{Y}\right) = \beta_0 + \beta_1 X + \mu$$

(note $\left(\hat{Y}\right)$ is pronounced y hat)

The graph of the estimated simple regression equation is called the estimated regression line.

where $\beta_0 \rightarrow$ y-intercept of the regression line.

$\beta_1 \rightarrow$ slope

$\left(\hat{Y}\right) \rightarrow$ estimated value of y for a given value of X.

5.3 PROPERTIES OF ESTIMATORS

There are different econometric methods with the help of which estimates of the parameters are obtained. We have to choose a good estimator which is close to the population parameter. This closeness is determined on the basis of following properties.

A) Estimator Properties for small sample are :

i) Unbiased :

The bias of an estimator is defined as the difference between its expected value and the true parameter.

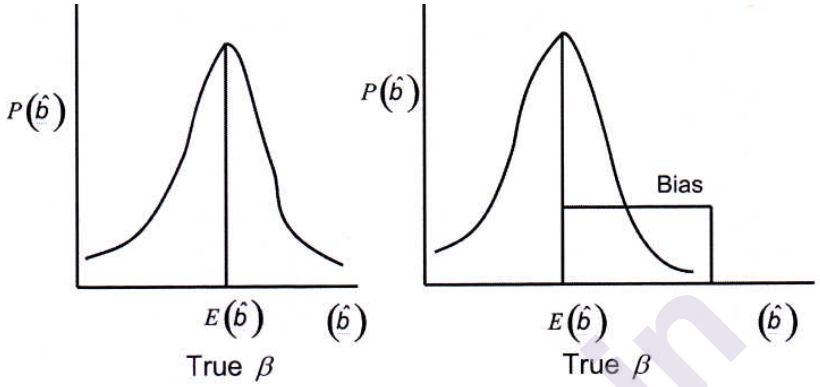
Bias = $E\left(\hat{\beta}\right) - \beta$

munotes.in

If bias is O, an estimator is said to be unbiased i.e. $E\left(\hat{\beta}\right) = \beta$

A biased and an unbiased estimator of the true β is explained in the following figure.

Figure No. 3.1



Unbiasedness is a desirable property and becomes important only when it is combined with a small variance.

ii) Least Variance:

An estimator is best when it has the smallest variance as compared with any other estimate obtained from other econometric methods. Symbolically, $\hat{\beta}$ is best if.

$$E\left[\hat{\beta} - E\left(\hat{\beta}\right)\right]^2 < E\left[\beta^* - E\left(\beta^*\right)\right]^2$$

$$\text{var}\left(\hat{\beta}\right) < \text{var}\left(\beta^*\right)$$

where $\beta^* \rightarrow$ any other estimate of the true parameter β.

iii) Efficiency :

An estimator is efficient when it possesses the various properties as compared with any other unbiased estimator.

$\hat{\beta}$ is efficient if

$E\left(\hat{\beta}\right) = \beta$ and $E\left[\hat{\beta} - E\left(\hat{\beta}\right)\right]^2 < E\left[\beta^* - E\left(\beta^*\right)\right]^2$

iv) Best, Linear Unbiased Estinator (BLUE) :

An estimator $\hat{\beta}$ is BLU if it is linear, unbiased and has smallest variance as compared with all the other linear unbiased estimator of the true β.

munotes.in

v) Least mean square Error estimator (LMSE):

An estimator is a minimum / least MSE if it has the smallest mean square error defined as the expected value of the squared difference of the estimator around the true population parameter β.

$$\text{MSE}\left(\hat{\beta}\right) = E\left(\hat{\beta} - \beta\right)^2$$

vi)Sufficiency:

An estimator is said to be sufficient estimator that utilise all the information a sample contains about the true parameter. It must use all the observations of the sample. Arithmetic mean (A.M.) is sufficient estimator because it give more information than any other measures.

B) Estimator Properties for Large Sample:

They are required when the sample is infinitely large. These properties therefore are also called as asymptotic

i) Asymptotic Unbaisedness :

An estimator is an asymptotically unbiased estimator of the true population parameter β, if the asymptotic mean of $\hat{\beta}$ is equal to β.

$$\lim_{n\to\infty} E\left(\hat{\beta}\right) = \beta$$

Asymptotic bias is an estimator is the difference between its asymptotic mean and true parameter.

(Asymptotic bias of $\hat{\beta}$) = $\left(\lim_{n\to\infty} E\left(\hat{\beta}\right)\right) - \beta$

If an estimator is unbiased in small samples it is also asymptotically unbiased.

ii) Consistency :

An estimator $\hat{\beta}$ is said to be consistent estimator of the true population of β if it satisfies two conditions.

a) $\hat{\beta}$ must be asymptotically unbiased

$$\lim_{n\to\infty} E\left(\hat{\beta}\right) = \beta$$

b) the variance of $\hat{\beta}$ must approach zero as n tends to infinity.

$$\lim_{n\to\infty} \left(Variance\hat{\beta}\right) = 0$$

munotes.in

If the variance is zero, the distribution collapses on the value of the true population parameter β. Both the bias and variance should decrease as n increases.

iii) Asymptotic Efficiency:

An estimator $\hat{\beta}$ is said to be asymptotically efficient estimator of the true population parameter, if :

a) $\hat{\beta}$ is consistent and

b) $\hat{\beta}$ has smaller asymptotic variance as compared with any other consistent estimator.

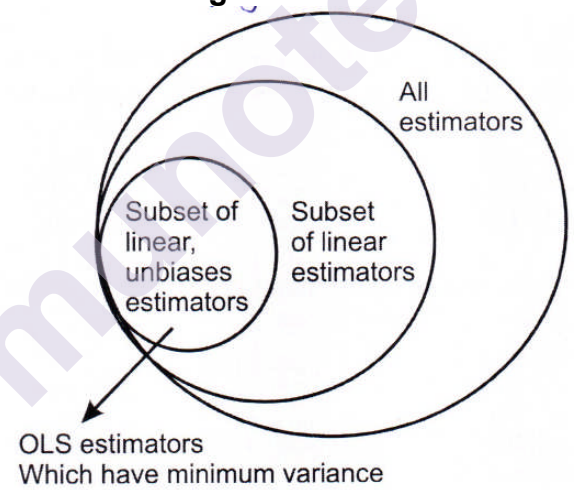
Statistical properties of Least Square Estimators :

Least square estimators are BLUE i.e. Best, Linear and Unbiased estimator provided error term U_i satisfies some assumption. The BLU properties of OLS (Ordinary Least Square) estimators are also called Gauss Markov.

Theorem :

The BLU properties are shown in the following diagram.

Figure No. 3.2



The properties of the OLS estimates of simple linear regression of the equation $Y_i = \beta_0 + \beta_1 X_i + U_i$ is based on the following assumptions :

1) U_i is a random real variable.

2) The mean value of U in any particular period is zero. i.e. $U_i = O$ i.e. $E(U_i) = O$

3) Assumption of Homo - scedasficity : i.e. the probability distribution of U remains the same over all observations of X. i.e. Variance of U_i is constant i.e. $E(U_i^2) = \sigma_U^2 =$ constant.

4) The random terms of different observation of U_i are independent. i.e. $E\left(U_i, U_j\right) = O$

5) X's are assumed to be fixed.

In a group of linear, unbiased estimators the OLS estimator. $\hat{\beta}_1$ has smallest variance i.e. they are best.

1) Linearity : The OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, are linear functions of the observed values of Y_i. Given the assumption that X's appear always with same values in repeated sampling process.

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

Where, x & y are in deviation form i.e. $x = X - \overline{X} y_i = Y - \overline{Y}$

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

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

Put the value of $y_i = Y_i - \overline{Y}$

$$\hat{\beta}_1 = \sum k_i \left(Y_i - \overline{Y}\right) = \sum k_i Y_i - \overline{Y} \sum k_i \text{ ..... (1)}$$

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

Q $\sum \left(X_i - \overline{X}\right) = O$

Put the value of $\sum k_i$ in equation (1) we get

$$\hat{\beta}_1 = \sum k_i Y_i - O = \sum k_i Y_i \left(Y_i - \overline{Y}\right) = \sum k_i Y_i - \overline{Y} \sum k_i \text{ ..... (2)}$$

Where $k_i Y_i = k_1 Y_1 + k_2 Y_2 + - - - - - + k_n Y_n$

This implies that $\hat{\beta}_1$ is a linear function of Y_i. Because k_i depends upon X_s^1 and X_s^1 are assumed to be fixed.

Similarly $\hat{\beta}_0 = \overline{Y} - \overline{\beta}, \overline{X}$

mynotes.in

Putting the value of $\hat{\beta}_1$ from equation (2)

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

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

$$\hat{\beta}_0 = \sum \left[\frac{1}{n} - \overline{X} k_i \right] Y_i \ldots\ldots\ldots \text{(3)}$$

Thus both $\hat{\beta}_0$ and $\hat{\beta}_1$ are the linear functions of the Y_s^1.

2) Unbiased: Both $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators.

i.e. $\sum\left(\hat{\beta}_1\right) = \beta_1$ and $\sum\left(\hat{\beta}_0\right) = \beta_0$

Proof : $\hat{\beta}_1 = \sum k_i Y_i \ldots\ldots\ldots\ldots$ (From 2)

$$= \sum k_i \left(\beta_0 + \beta_1 X_i + U_i \right)$$

$$= = \beta_0 \sum k_i + \beta_1 \sum k_i X_i + \sum k_i U_i \ \ldots\ldots\ldots\ldots \text{(4)}$$

$$\text{Q } k_i = \frac{x_i}{\sum X_i^2}$$

$$\sum k_i = \frac{\sum x_i}{\sum x_i^2} \ \text{Q}\left[\sum x_i = \sum \left(X - \overline{X} \right) = 0 \right]$$

$$\sum k_i = \frac{0}{\sum X_i^2}$$

$$\sum k_i = 0$$

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

Putting the value of $X_i = x_i + \overline{X}$

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

$$= 1 + \frac{\overline{X}(0)}{\sum x_i^2}$$

$$\therefore \sum x_i = 0$$

$$\sum k_i X_i = 1$$

Substituting the value $\sum k_i = 0, \sum k_i X_i = 1$ in equation (4)

munotes.in

$$\hat{\beta}_1 = \beta_0(0) + \beta_1(1) + \sum k_i U_i$$

$$\hat{\beta}_1 = \beta_1 + \sum k_i U_i \text{ ..... (5)}$$

Take expectations on both sides.

$$E\left(\hat{\beta}_1\right) = \beta_1 + \sum k_i E\left(U_i\right)$$

$$\text{Q } E\left(U_i\right) = 0$$

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

This is known as unbiasedness of the estimated parameter. Thus $\hat{\beta}_1$ is an unbiased estimator of β_1.

It is known that $\hat{\beta}_0$ from OLS is

$$\hat{\beta}_0 = \sum\left[\frac{1}{n} - \overline{X}k_i\right]Y_i \text{ ..... (from 3)}$$

$$= \sum\left[\frac{1}{n} - \overline{X}k_i\right]\left(\beta_0 + \beta_1 X_i + U_i\right)$$

$$= \beta_0 + \beta_1\frac{\sum X_i}{n} + \frac{\sum U_i}{n} - \overline{X}\sum k_i\beta_0 - \beta_1\overline{X}\sum k_i X_i - \overline{X}\sum k_i U_i \text{ ..... (6)}$$

It is proved that $\sum k_i = 0, \sum k_i X_i = 1$

By substituting these values in equation (6)

$$\hat{\beta}_0 = \beta_0 + \frac{\sum U_i}{n} - \overline{X}\sum k_i U_i$$

Taking expectation on both sides

$$E\left(\hat{\beta}_0\right) = \beta_0 + \frac{\sum E\left(U_i\right)}{n} - \overline{X}\sum k_i E\left(U_i\right)$$

$$\text{Q } E\left(U_i\right) = 0$$

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

This implies that $\hat{\beta}_0$ is an unbiased estimator of β_0.

3) Minimum variance property:

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

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

$$= E\left(\sum k_i U_i\right)^2 \text{ since } \hat{\beta}_1 = \beta_1 + = \sum k_i U_i$$

$$= E\left(k_1 U_1 + k_2 U_2 + ... + k_n U_n\right)^2 \text{ ..... (see equation 5)}$$

munotes.in

$$= E\left(k_1^2U_1^2 + k_2^2U_2^2 + ... + k_n^2U_n^2 + 2k_1k_2U_1U_2 + ... + 2k_{n-1}kU_{n-1}U_n\right)$$

$$= E\left(\sum k_i^2U_i^2 + 2\sum_{i\neq j} k_ik_jU_iU_j\right)$$

$$= \sum k_i^2E\left(U_i^2\right) + 2\sum k_ik_jE\left(U_iU_j\right)$$

Since $E\left(U_i, U_j\right) = 0, E\left(U_i^2\right) = \sigma_u^2$ (Assumption)

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

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

Q $K_i^2 = \dfrac{x_i}{\sum x_i^2}$

$$\sum K_i^2 = \frac{\sum x_i^2}{\left(\sum x_i^2\right)^2} = \frac{1}{\sum x_i^2}$$

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

$$\therefore \text{Var}\left(\hat{\beta}_0\right) = E\left[\left(\hat{\beta}_0 - \beta_0\right)^2\right]$$

$$= E\left[\sum\left(\frac{1}{n} - \overline{X}K_i\right)^2 U_i^2\right]$$

$$= \sigma_u^2\sum\left(\frac{1}{n} - \overline{X}K_i\right)^2$$

$$= \sigma_u^2\sum\left(\frac{1}{n^2} - \frac{2}{n}\overline{X}K_i + \overline{X^2}K_i^2\right)$$

Since $\sum k_i = 0 \sum k_i^2 = \dfrac{1}{\sum x_i^2}$

$$\text{Var}\left(\hat{\beta}_0\right) = \sigma_u^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum x_i^2}\right) \text{..... (8)}$$

Now $\dfrac{1}{n} + \dfrac{\overline{X}^2}{\sum x_i^2} = \dfrac{\sum x_i^2 + n\overline{X}^2}{n\sum x_i^2} = \sigma_u^2\left[\dfrac{\sum\left(X_i - \overline{X}\right)^2 + n\overline{X}^2}{n\sum x_i^2}\right]$

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

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

munotes.in

Var $\left(\hat{\beta}_0\right) = \sigma_u^2 \dfrac{\sum x_i^2}{n\sum x_i^2}$

We are interested in the least square estimators which have the smallest variance.

Let β_1^* be another estimator of β_1.

$\beta^* = \sum W_i Y_i$ where constant $W_i \neq K_i$ but $W_i = K_i + C_i$

$$\therefore \beta_1^* = \sum W_i \left(\beta_0 + \beta_1 X_i + U_i\right)$$
$$= \beta_0 \sum W_i + \beta_1 \sum W_i X_i + \sum W_i U_i$$
$$\therefore E\left(\beta_1^*\right) = \beta_0 \sum W_i + \beta_1 \sum W_i X_i$$

$\left[\text{Q } E\left(U_i\right) = 0\right]$ Assumption.

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

$$\sum W_i = 0 \text{ and } \sum W_i X_i = 1$$
$$\sum W_i = \sum \left(K_i + C_i\right) = \sum K_i + \sum C_i = 0$$

But $\sum K_i = 0$

$$\sum W_i = 0 + \sum C_i = 0$$

Hence $\sum C_i = 0$ and $\sum W_i = 0$

$$\sum W_i X_i = \sum \left(K_i + C_i\right) X_i = 1$$
$$= \sum K_i X_i + \sum C_i X_i = 1$$

But $\sum K_i X_i$ =1

$$1 + \sum C_i X_i = 1$$
$$\sum C_i X_i = 1 - 1 = 0$$

Hence $\sum C_i = 0$ and $\sum C_i X_i = 0$

Var $\beta_1^* = E\left[\beta_1^* - E\left(\beta_1^*\right)\right]^2$

$$= E\left(\beta_1^* - \beta_1\right)^2$$

Var $\beta_1^* = E\sum \left(W_i U_i\right)^2$ Q $\beta_1^* = \beta_1 + \sum W_i U_i$

$$= E\left(W_1^2 U_1^2 + W_2^2 U_2^2 + ... + W_n^2 U_n^2 + 2\sum W_i W_j U_i U_j\right)^2$$

mrunotes.in

Var $\beta_1^* = E\left(W_i^2 U_i^2 + 2\sum W_i W_j U_i U_j\right)$

$= \sum W_i^2 E\left(U_i^2\right) + 2\sum W_i W_j E\left(U_i U_j\right)$

Since $EU_i U_j = 0, \in \left(U_1^2\right) = \sigma_u^2$ (Assumptions)

Var $\beta_1^* = \sigma_u^2 \sum w_i^2 + 0$

Putting the values of w_i^2

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

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

Var $\beta_1^* = \dfrac{\sigma_u^2}{\sum x_i^2} + \sigma_u^2 \sum C_i^2 + 0$ Q $\sum k_i C_i = 0$

Var $\beta_1^* =$ $\hat{\beta}_1$ + constant Q $\sum C_i^2 > 0$

Var $\beta_1^* > \hat{\beta}_1$

It implies that OLS estimator has the minimum variance.

Similarly, let us take a new estimator β_0^*, which is assumed to be a linear function of the Y_i and unbiased.

$w_i = k_i + C_i$

Let $\beta_0^* = \sum\left(\dfrac{1}{n} - \overline{X} w_i\right) Y_i$ where $w_i \neq k_i \in \left(\beta_0^*\right) = \beta_0^*$ only if $\sum w_i = 0$ and $\sum w_i X_i z$.

It implies that $\sum C_i = 0$ and $\sum C_i X_i = 0$

Var $\left(\beta_0^*\right) = \sigma_u^2 \sum\left(\dfrac{1}{n} - \overline{X} w_i\right)^2$

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

$= \sigma_u^2 \left[\dfrac{1}{n} + \overline{X}^2 \left(\sum k_i^2 + \sum C_i^2\right)\right]$ Q $\sum w = 0$

Since $\sum w_i^2 = \sum k_i^2 + \sum C_i^2 + 2k_i C_i$

But $\sum k_i C_i = 0 \sum w_i^2 = \sum K_1^2 + k C_1^2$

munotes.in

$$= \sigma_u^2 \left[\frac{1}{n} + \overline{X}^2 \left(\frac{1}{\sum x_i^2} + \sum C_i^2 \right) \right]$$

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

Var $\left(\beta_0^*\right)$ = Var $\left(\hat{\beta}_0\right)$ + a positive constants

$\therefore$ Var $\left(\beta_0^*\right)$ > Var $\left(\hat{\beta}_0\right)$

Thus it is proved that the OLS estimators are BLU.

The standard error test of the estimators β_0 and β_1.

The least square estimates are obtained from a sample of observations. So sampling errors are inevitable to occur in all estimates. Therefore to measure the size of the error it becomes necessary to apply test of significance. Let us discuss the standard error test. It helps us to decide whether the estimates are statistically reliable or not. To test the null hypothesis.

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

$H_1 : \beta_1 \neq 0$

where

$$S \in \left(\hat{\beta}_1\right) = \sqrt{Var\left(\hat{\beta}_1\right)} = \sqrt{\frac{\sigma_4^2}{\sum x_i^2}}$$

$$S \in \left(\hat{\beta}_0\right) = \sqrt{Var\left(\hat{\beta}_0\right)} = \sqrt{\frac{\sigma_4^2 \sum X_i^2}{n \sum x_i^2}}$$

When the standard error is less than half of the numerical value of the parameter estimate $\left[S \in \left(\hat{\beta}_1\right) < \frac{1}{2}\left(\hat{\beta}_1\right) \right]$, we conclude that the estimate is statistically significant. Therefore we reject the hull hypothesis & accept the alternative hypothesis i.e. the true population parameter β_1 is different from zero.

If the standard error is greater than half of the numerical value of the parameter estimate $\left[S \in \left(\hat{\beta}_1\right) > \frac{1}{2}\left(\hat{\beta}_1\right) \right]$, we conclude that the null hypothesis is accepted and the estimate is not statistically significant.

munotes.in

The acceptance of null hypothesis implies the explanatory variable to which the estimate relates does not effect the dependent variable. i.e. there is no relationship between Y and X variables.

5.4 REFERENCE

- S-Shyamala and Navdeep Kaur, 'Introduce too y Econometrics'.
- Neeraj R, Hatekar, 'Principles of Econometrics : An Introduction us in, R'

❖❖❖❖

munotes.in

6

TESTS IN REGRESSION AND INTERPRETING REGRESSION COEFFICIENTS

Unit Structure:

6.0 OBJECTIVES

- To understand the meaning of adjusted R squared.
- To use the F- test in regression.
- To interpret the regression coefficients.

6.1 INTRODUCTION

Regression coefficients are a statistical tool or measure of the average functional relationship between two or more than two variables. In the regression analysis, one variable is dependent and other variables are independent. In short, it measures the degree of dependence of one variable on another variable.

Regression coefficient was used first to estimate the relationship between the heights of father's and their sons. Regression coefficient denoted by b.

6.2 Z - TEST:

The Z test of the least squares estimates is based on standard normal distribution and is applicable when the population variance is known or the population variance is unknown if the sample is sufficiently large i.e. $n > 30$.

munotes.in

Assuming The null hypothesis $H_0 : \beta = 0$

Alternative hypothesis $H_1 : \beta \neq 0$ Then the least square estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ have the following normal distribution.

$$\hat{\beta}_0 : \ N\left[\beta_0, \sigma_{\hat{\beta}_0} = \sqrt{\sigma_u^2 \frac{\sum X^2}{n\sum x_i^2}}\right]$$

$$\hat{\beta}_1 : \ N\left[\beta_1, \sigma_{\hat{\beta}_1} = \sqrt{\sigma_u^2 \frac{1}{\sum x_i^2}}\right]$$

After transforming it into $Z : \ N(0,1)$

$$Z_i = \frac{X_i - \mu}{\sigma} : \ N(0,1)$$

$X_i \rightarrow$ value of the variable which is to be normalise
$\mu \rightarrow$ mean of the distribution
$\sigma \rightarrow$ standard deviation

$$Z^* = \frac{\hat{\beta}_0 - \beta_0}{\sigma_{\hat{\beta}_0}} = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sigma_u^2 \sum x_i / n\sum x_i^2}} : \ N(0,1)$$

$$Z^* = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma_u^2 / n\sum x_i^2}} : \ N(0,1)$$

Given the calculated value of Z^*, we select the level of significance to decide the acceptance or rejection of null hypothesis. Generally speaking, in econometrics we choose 5% or 1% level of significance. i.e. we tolerate / consider 5 times out of 100 to be wrong while making decisions.

We perform a two tail test i.e. critical region for both tails of standard normal distribution. For i.e. for 5% level of significance, each tail will include area 0.25 probability. The table value of Z corresponding to probability 0.25 at each end of the curve or both the tails is $Z_1 = -1.96$ and $Z_2 = 1.96$

To conclude we compare the observed value Z^* with the table value of Z. If it falls in the critical regions. i.e. if $Z^* > 1.96$ or $Z^* < -1.96$, we reject the hull hypothesis. In case if it is outside of the critical region, i.e. $-1.96 < Z^* < 1.96$, we accept the hull hypothesis.

nunotes.in

in econometrics, it is customarily to test the hypothesis that true population parameter is zero.

$H_0 : \beta_1 = 0$ and is tested against the alternative hypothesis.

$H_1 : \beta_1 \neq 0$.

To test the above null hypothesis, $\beta = 0$ in the Z transformed formula.

$$Z^* = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - 0}{\sigma_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\sigma_{\hat{\beta}_1}}$$

If $Z^* > |1.96|$ we accept H_1 and reject H_0.

Given the 5% level of significance the critical value of Z is 1.96 which is approximately equal to 2.0. In standard error test we reject null hypothesis if $\sigma_{\hat{\beta}_1} > \frac{\beta_1}{2}$. In case of 2 test it $Z^* > 2$ we reject null hypothesis. The two statements are identical because $Z^* => \frac{\hat{\beta}_1}{\sigma_{\hat{\beta}_1}} > 2$ (if we accept H_1) or $\sigma_{\hat{\beta}_1} > \frac{\hat{\beta}_1}{2}$.

Thus standard error test and 2 tests give the same result.

6.3 T TEST -

t Test includes the variance estimates S_X^2 instead of true variance σ_X^2. So the formula is as following :

$t = \frac{X_i - u}{S_X}$ with (n - 1) degrees of freedom

$u \rightarrow$ value of population mean

$S_X^2 \rightarrow$ sample estimate of the population variance

$S_X^2 \rightarrow \sum\left(X_i - \overline{X^2}\right)/(n-1)$ $n \rightarrow$ sample size.

The sampling distribution in $\overline{X}$: $N\left(u, S_X^2\right)$ and the transformation statistic is $\left(\overline{X} - u\right)/\sqrt{S_X^2/n}$ and has t distribution with $(n-1)$ degrees of freedom.

We have least square estimates as :

munotes.in

$\hat{\beta}_0 : N\left[\beta_0, \hat{\sigma}_u^2 \frac{\sum X_i^2}{n\sum X_i^2}\right]$ and $\hat{\beta}_1 : N\left[\beta_1, \hat{\sigma}_{\hat{\beta}_1}^2 = \hat{\sigma}_u^2 \frac{1}{\sum X_i^2}\right]$

From this the t statistic for $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained from a sample reduces to $t^* = \frac{\hat{\beta}_0 - \beta_0^*}{\sqrt{\hat{\sigma}\hat{\beta}_0}}$ and $t^* = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\hat{\sigma}\hat{\beta}_1}}$ with $n-k$ degrees of freedom.

$\hat{\beta}_0$ and $\hat{\beta}_1 \rightarrow$ least squares estimates of β_0 and β_1 respectively.

β_0^* and $\beta_1^* \rightarrow$ hypothesised value of β_0 and β.

$\hat{\sigma}_{\hat{\beta}_0}^2 \rightarrow$ estimated variance of β_0 (from the regression)
$\hat{\sigma}_{\hat{\beta}_1}^2 \rightarrow$ estimated variance of β_1
$n \rightarrow$ sample size
$K \rightarrow$ total number of estimated parameters
(in oure case of K = 2)

Assuming The null hypothesis is $H_0 : \beta_0 = 0$
The alternative hypothesis $H_1 : \beta_0 \neq 0$

$$t^* = \frac{\hat{\beta}_0}{S \in_{\hat{\beta}_0}}$$

Then the calculated t^* value is compared to the table values of t with n - K degrees of freedom.

If $t^* > t_{0.025}$, we reject the null hypothesis, i.e. we accept that the estimate $\hat{\beta}_0$ is statistically significant.

When $t^* > t_{0.025}$, we accept the null hypothesis, that is, the estimate $\hat{\beta}_0$ is not statistically significant at the 5% level of significance.

Similarly for the estimate $\hat{\beta}_1$.

Null hypothesis $H_0 : \beta_1 = 0$ and Alternative hypothesis $H_1 : \beta_1 = 0$

nunotes.in

$$t^* = \frac{\hat{\beta}_1}{S \in_{\hat{\beta}_1}}$$

If $t^* > t_{0.025}$ we reject the null hypothesis and we conclude that the estimate $\hat{\beta}_1$ is statistically significant at 5% level of significance.

If $t^* > t_{0.025}$ we accept the null hypothesis that is, we conclude that the estimate $\hat{\beta}_1$ is not statistically significant at 5% level of significance.

Confidence intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$

The t statistic for $\hat{\beta}_0$ is $t^* = \frac{\hat{\beta}_0 - \beta_0^*}{S \in_{(\hat{\beta}_0)}}$ with n - k degrees of freedom. First we choose the 95 percent confidence level or and find t values of $\pm t_{0.025}$ from t table with n - K degrees of freedom. This implies that the probability of t lying between $-t_{0.025}$ and $+t_{0.025}$ is 0.95.

Thus the 95 percent confident interval for β_0, small sample for its estimation is $\hat{\beta}_0 - t_{0.025} \; S \in_{(\hat{\beta}_0)} < \beta_0 < \hat{\beta}_0 + t_{0.025} S \in_{(\hat{\beta}_0)}$ with n - K degrees of freedom or $\beta_0 < \hat{\beta}_0 + t_{0.025} S \in_{(\hat{\beta}_0)}$ with n - K degrees of freedom.

Similarly, for the estimates of $\hat{\beta}_1$, $t^* = \frac{\hat{\beta} - \beta}{S \in (\hat{\beta})}$ with n - K degrees of freedom.

The confidence interval 95 percent level is $\hat{\beta}_1 - t_{0.025}$ $S \in_{(\hat{\beta}_1)} < \beta_1 < \hat{\beta}_1 + t_{0.025} S \in_{(\hat{\beta}_1)}$ with n - K degrees of freedom or $\beta_1 = \hat{\beta}_1 \pm t_{0.025} S \in_{(\hat{\beta}_1)}$ with n - k degrees of freedom.

6.4 GOODNESS OF FIT (R^2)

A measure of goodness of fit is the square of the correlation coefficient (R^2), which shows the percentage of the total variation of the dependent variable that can be explained by the independent variable (X).

munotes.in

Since,

TSS = RSS + ESS

TSS $\rightarrow$ Total sum of squares = $\sum y_i^2$

RSS $\rightarrow$ Residual sum of squares = $\sum e_i^2$

ESS $\rightarrow$ Explained sum of squares = $\hat{\beta}_1 \sum x_i^2$ and $y = Y_i - \overline{Y}$ and $x = X_i - \overline{X}$.

The decomposition of the total variations in Y leads to a measure of goodness of fit, also called the coefficient of determination which is represented by :

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

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

As $ESS = TSS - RSS$

$$R^2 = \frac{TSS - RSS}{TSS}$$

$$\therefore R^2 = \frac{\sum y_i^2 - \sum e_i^2}{\sum y_i^2}$$

$$= 1 - \frac{\sum e_i^2}{\sum y_i^2}$$

Properties of R^2

i) It is a non-negative quantity i.e. it is always positive $R^2 \geq 0$. It is calculated with the assumption that there is an intercept term in the regression equation of Y on X_1

ii) Its limits ranges from $O \leq R^2 \leq 1$ when $R^2 = 0$, it implies no relationship between dependent and explanatory variables.

When $R^2 = 1$, there is a perfect fit.

iii) $R^2 = r^2$

From definition, r can be written as

$$r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \sqrt{\sum y_i^2}}$$ where $x = X_i - \overline{X}$

munotes.in

$$R^2 = \hat{\beta}_1^2 \frac{\sum x_i^2}{\sum y_i^2} \text{ and } \hat{\beta}_1 = \frac{\sum x_i y_1}{\sum y_i^2}$$

$$\therefore R^2 = \left[\frac{\sum x_i y_i}{\sum x_i}\right]^2 \frac{\sum x_i^2}{\sum y_i^2} = \frac{\left(\sum x_i y_i\right)}{\sum x_i^2 \sum y_i^2}$$

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

$$R^2 = r^2$$

Correlation coefficient $r = \pm\sqrt{R^2}$

While R^2 varies between 0 and 1 i.e. $O \le R^2 \le 1$ r varies between - 1 and + 1 i.e. $-1 \le r \le +1$, indicating negative correlation and positive linear correlation respectively, at the two extreme values.

6.5 ADJUSTED R SQUARED

The R squared statistic suffers from a major drawback. No matter the number of variables we add to our regression model the value of R square never decreases.

If either remains same or increases with the new independent variable even though the variable is redundant. In reality, its result can not be accepted since the new independent variable might not be necessary to determine the target variable. So the adjusted R square deals with this problem.

Adjusted R squared measures the proportion of variation explained by only those independent variables which are really helpful in determining the dependent variable. It is represented with the help of the following formula

$$\text{Adjusted } R^2 = \left\{1 - \left[\frac{\left(1-R^2\right)(n-1)}{(n-k-1)}\right]\right\}$$

Where $n \rightarrow$ sample size

$k \rightarrow$ number of independent variable

$R \rightarrow$ R squared values determined by the model

To conclude the difference between R square and adjusted R square we may say that

i) When we add a new independent variable to a regression model, the R-squared increase, even though the new independent variable is not useful indeteming the dependent variable. Whereas

mumotes.in

adjusted R squared increases only when new independent variables is useful and affect the dependent variable.

ii) Adjusted R - squared can be negative when R-squared is close to zero.

iii) Adjusted R-squared value always be less than or equal to R-squared value.

6.6 THE F-TEST IN REGRESSION

F - test is a type of statistical test which is very flexible. It can be used in a wide variety of settings. In this unit we will discuss the F-test of overall significance. It indicates whether our regression modd provides a better fit to the data than a model that contains no independent variables. So here we will explain how the F-test of overall significance fits in with other regression statistics, such as R -square. R-square provides an estimate of the strength of the relationship between regression model and the response variable. It does not provide any formal hypothesis test for this relationship. Whereas the overall significance F-test determines whether this relationship is statistically significant or not. If the P value for the overall F-test is less than the level of significance, we conclude that the R-square value is significantly different from zero.

The overall F-test compares the model with the model with no independent variables such type of model is known as intercept only model. It has the following two hypothesis.

a) The null hypothesis - The fit of the intercept only model and our model are equal.

b) Alternative hypothesis - The fit of the intercept - only model is significantly reduced compared to our model.

We can find the overall F - test in the ANOVA table.

Table : ANOVA

Source	DF	Adj SS	Adj MS	F - Value	P - Value
Regression	3	12833.9	4278.0	57.87	0.000
East	1	226.3	226.3	3.06	0.092
South	1	2255.1	2255.1	30.51	0.000
North	1	12330.6	12330.6	166.80	0.000
Error	25	1848.1	73.9		
Total	28	14681.9			

In the above table, compare the p-value for the F-test our significance level. If the p-value is less than the significance level, our sample data provide sufficient evidence to conclude that our regression model fits the data better than the model with no independent variables.

6.7 INTERPRETING REGRESSION COEFFICIENTS

Regression coefficients are a statistical tool or measure of the average functional relationship between two or more than two variables. In the regression analysis, one variable is dependent and other variables are independent. In short, it measures the degree of dependence of one variable on another variable.

Regression coefficient was used first to estimate the relationship between the heights of father's and their sons. Regression coefficient denoted by b.

Basically, there are two types of regression coefficients, i.e. regression coefficient of regression y on X $\left(b_{yx}\right)$ and regression coefficients of regression X on Y $\left(b_{xy}\right)$.

Properties of Regression Coefficient:

Some important properties of regression coefficient are as follows:

1) The both regression coefficients have the same sign. If b_{yx} is positive, b_{xy} will be also positive and if b_{yx} is negative, b_{xy} will be also negative.

If, $b_{yx} > 0$, $b_{xy} > 0$

$b_{yx} < 0$, $b_{xy} < 0$

2) If a regression coefficient is more than unity, the other regression coefficient must be less than unity. If a regression coefficient is more than - 1, other regression coefficient must be less than - 1.

If, $b_{yx} > 1$, $b_{xy} < 1$

$b_{yx} > -1$, $b_{yx} < -1$

3) The geometric mean (GM) of two regression coefficients is equal to the correlation coefficient.

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

Where,

r = correlation coefficient

b_{yx} = Regression coefficient of regression y on x.

munotes.in

b_{xy} = Regression coefficient of regression x on y.

4) Correlation coefficient and regression coefficient have the same sign.

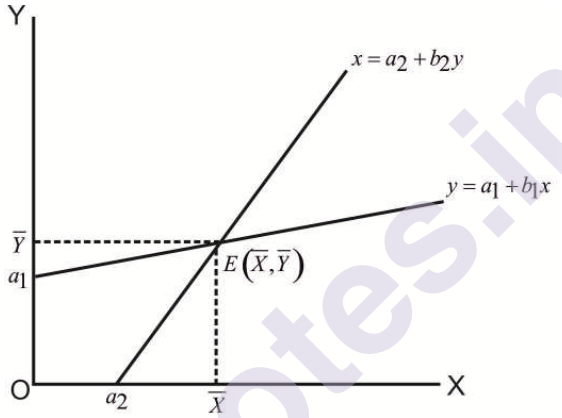
If, $r > 0, b_{yx} > 0 \& b_{xy} > 0$

$r < 0, b_{yx} < 0 \& b_{xy} < 0$

5) Arithmetic mean of two regression coefficients is equal to or greater than correlation coefficient.

$$\frac{\left(b_{yx} + b_{xy}\right)}{r} \leq r$$

6) Two regression lines intersects to each other on arithmetic means of these variables. $\left(\overline{X}, \overline{Y}\right)$



Computation of Regression Coefficients:

Regression coefficients can be calculated from following formulas.

$$b_{yx} = \frac{\sum xy - \left(\sum x . \sum y\right)}{\sum y^2 - \left(\sum y\right)^2}$$

$$b_{xy} = \frac{\sum xy - \left(\sum x . \sum y\right)}{\sum x^2 - \left(\sum x\right)^2}$$

Steps:

For the calculation of regression coefficients have to follow the following steps.

1) Take the sums of all observations of X and Y variables $\left(\sum x, \sum y\right)$.
2) Take the sums of squares of X and Y variables $\left(\sum x^2, \sum y^2\right)$
3) Take the sum of products of all observations of X and Y variables $\left(\sum xy\right)$.
4) Use the following formulas for calculating the regression coefficients.

rtunotes.in

$$b_{yx} = \frac{\sum xy - (\sum x - \sum y)}{\sum y^2 - (\sum y)^2}$$

$$b_{xy} = \frac{\sum xy - (\sum x - \sum y)}{\sum x^2 - (\sum x)^2}$$

Example :

X	2	4	1	5	6	7	8	1	0
Y	3	1	5	7	8	9	0	5	4

Calculate the b_{yx} and b_{xy} from above information.

Solution :

X	Y	XY	X^2	Y^2
2	3	6	4	9
4	1	4	16	1
1	5	5	1	25
5	7	35	25	49
6	8	48	36	64
7	9	63	49	81
8	0	0	64	0
1	5	5	1	25
0	4	0	0	16

First take the sums of all observations of X and Y variables $(\sum x \,\&\, \sum y)$

$$\sum x = 2+4+1+5+6+7+8+1+0$$

$$\boxed{\sum x = 34}$$

$$\sum y = 3+1+5+7+8+9+0+5+4$$

$$\boxed{\sum y = 42}$$

Then, take sums of squares of X and Y variables $\left(\sum x^2 \,\&\, \sum y^2\right)$

$$\sum x^2 = 4+16+1+25+36+49+64+1+0$$

$$\boxed{\sum x^2 = 196}$$

$$\sum y^2 = 9+1+25+49+64+81+0+25+16$$

$$\boxed{\sum y^2 = 270}$$

munotes.in

Now take the sum of products of all observations of X and Y variables $(\sum xy)$.

$$\sum xy = 6+4+5+35+48+63+0+5+0$$

$$\boxed{\sum xy = 166}$$

Now keep the above values in following equations and calculate the regression coefficients.

Regression coefficient of Regression Y on X -

$$b_{yx} = \frac{\sum xy - (\sum x . \sum y)}{\sum y^2 - (\sum y)^2}$$

$$= \frac{166 + (34 \times 42)}{270 - (42)^2}$$

$$= \frac{166 - (1428)}{270 - (1764)}$$

$$= \frac{-1262}{-1494}$$

$$= \frac{1262}{1494} ((-) \div (-) = +)$$

$$= 0.845$$

$$\boxed{b_{yx} = 0.85}$$

Regression coefficient of Regression X on Y -

$$b_{xy} = \frac{\sum xy - (\sum x . \sum y)}{\sum x^2 - (\sum x)^2}$$

$$= \frac{166 - (34 \times 42)}{196 - (34)^2}$$

$$= \frac{166 - 1428}{196 - (1156)}$$

$$= \frac{-1262}{-960}$$

$$= \frac{1262}{960}$$

$$\boxed{b_{xy} = 1.32}$$

So, $b_{yx} = 0.85$

$b_{xy} = 1.32$

mynotes.in

6.8 QUESTIONS

Q.1

X	2	4	6	5	3	9	10
Y	4	2	5	7	8	0	4

Calculate regression coefficients (b_{yx} and b_{xy})

Q.2

X	4	5	6	8	9	10	7	6
Y	4	1	5	4	10	12	7	8

Calculate regression coefficients (b_{yx} and b_{xy})

6.9 REFERENCE

- S-Shyamala and Navdeep Kaur, 'Introduce too y Econometrics'.
- Neeraj R, Hatekar, 'Principles of Econometrics : An Introduction us in, R'

❖❖❖❖

munotes.in

MODULE 4

7

PROBLEMS IN SIMPLE LINEAR REGRESSION MODEL: HETEROSCEDASTICITY

Unit Structure:

7.0 OBJECTIVES :

1. To understand the causes of Heteroscedasticity.
2. To understand the detection of Heteroscedasticity.
3. To understand the consequences of Heteroscedasticity.

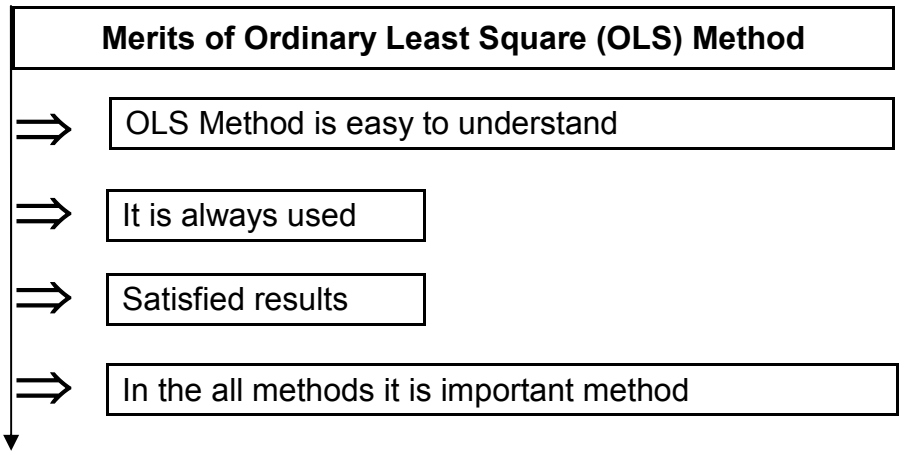
7.1 INTRODUCTION :

In the previous unit, you learnt about simple linear regression as meaning, estimation of simple linear regression model etc. In this unit you learn about the problems in simple linear regression model.

Simple regression model includes only two variables, so simple regression model is also known as 'Two Variables Regression Model. When we consider the linear relationship between two variables in the simple regression model, then it is called as simple linear regression model. There are two methods for the estimation of simple linear regression model which are namely ordinary least square method (OLS) and maximum likelihood principle. When OLS method is unable to use for the estimation of simple linear regression model, maximum likelihood principle is being used. But because of the following factors, OLS

mjnotes.in

method is appropriate for estimation of simple linear regression model.



Simple Linear Regression model has been written as follows:

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

Where, Y_i = Dependent Variable

β_1 = Intercept

β_2 = Slope

X_i = Independent Variable

u_i = Random Variable

For the estimation of above simple linear regression if we have to use the OLS method, then the study of assumptions of OLS method become necessary.

7.2 ASSUMPTIONS OF ORDINARY LEAST SQUARE (OLS) METHOD

Least Square principle is developed by German mathematician Gaurs.

There are ten assumptions of OLS method. In short, we discuss as below –

1. The regression model is linear in the parameters.

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

It is a simple linear regression model and this model is linear in both (X, Y) Variables and parameters (β_1, β_2). In short, linearly

in the parameters is crucial for the use or application of least square principle.

2. X values are fixed in repeated sampling:

Values taken by the regression X are assumed or considered to be fixed in repeated sampling:

$$Y_i = \beta_1 + \beta_1 X_i + u_i$$

Where X_i = Fixed / Constant

Y_i = Varies

Because of this assumption the regression analysis becomes the conditional regression analysis.

3. Zero mean value of disturbance u_i :

It means, expected value of the disturbance u_i is zero.

Given the values of X, the mean or expected value of the disturbance term (u_i) is zero.

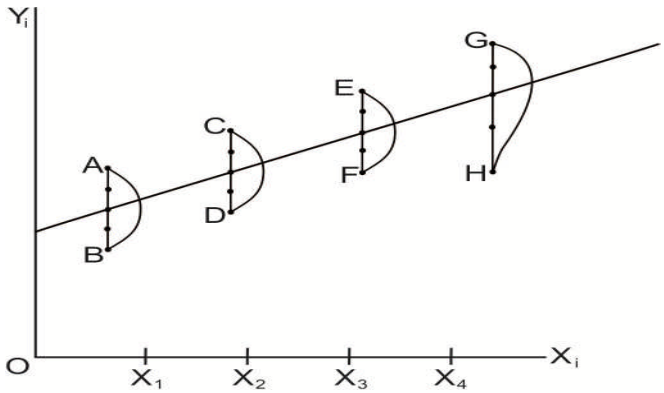
Symbolically,

$E(u_i) = O$ Or $E = (u_i / X_i) = O$

4. Homoscedasticity or equal variance of u_i : Homo means equal and scedasticity means spread. So Homoscedasticity means equal spread. Given the values of X, the variance of u_i is the same for all observations

Symbolically,

$$Var (u_i / X_i) = 6^2$$



munotes.in

In the above figure, AB is the spread of u_i for X_1, CD is the spread of u_i for X_2 and EF is the spread of u_i for X_3,

So,

AB = CD = EF

It means, u_i is Heteroscedastic – In this case, var $(u_i / X_1) \neq 6^2$

5. No autocorrelation between the disturbance terms :

Given any two X values, X_i and X_j , ($i \neq j$) the correlation between any two u_i and u_j ($i \neq j$) is zero.

Symbolically,

$$\text{Cov}\ (u_i u_j / X_i X_j) = E\left[(u_i - E(u_i)/X_i \ \ (u_j - E(u_j)/X_j)\right]$$
$$= E\left[(u_i/X_i \ \ (u_j X_j)\right]$$

Here, $E(u_i) = O$

$E(u_j) = O$

$= O$

Here, $E(u_i/X_i) = O$

$E(u_j/X_j) = O$

6. Zero covariance between u_i and X_i :

$$\text{Cov}\ (u_i , X_i) = E\left[(u_i - E(u_i)\ (X_i - E(X_i)\right]$$

Here, $E(u_i) = O$

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

Here, $E(u_i) = O$

$$= E\ (u_i X_i)$$

Here, X_i = non stochastic

$$= X_i\ E\ (u_i)$$

Here, $E\ (u_i)\ ) = O$

$\text{Cov}\ (u_i , X_i) = O$

munotes.in

7. The number of observation 'n' is greater than the number of parameters (to be estimated).

8. Variability in X values:
The X variable is a given sample must not all be the same.

9. The regression model is correctly specified.

10. There is no perfect multicolinearity: It means that there is no perfect linear relationship among the explanatory variables.

These are the ten important assumption of OLS method.

While using the OLS method for the estimation of simple linear regression model, if assumption no. 4, 5 and 10 do not fulfil, problems create in the simple linear regression model which are namely heteroscedasticity, autocorrelation and multicolinearity.

Check your progress:

1. What are the ten principles of ordinary least square (OLS) method?

7.3 HETEROSCEDASTICITY

The term Heteroscedasticity is the opposite term of homoscedasticity; heteroscedasticity means unequal variance of disturbance term (u_i).

$E(u_i^2) = 6^2 \rightarrow$ Homoscedasticity

$E(u_i^2) \neq 6^2 \rightarrow$ Heteroscedasticity

Given the values of X, the variance of u_i (Expected or mean value of u_i) that E (u_i), is the same for all observations. This as assumption of OLS, principle which is useful for the estimation of simple linear regression model.

$$E(u_i^2) = Var(u_i) = 6^2$$

If above assumption does not fulfil, then the problem of heteroscedasticity arises in the estimation of simple linear regression.

mynotes.in

Ex. If Income of individual increases, has saving increases but the variance of saving will be the same, it is known as homoscedasticity

$$Y^{\uparrow} \rightarrow S^{\uparrow} \rightarrow \text{Var (S) = same} \rightarrow \text{Homoscedasticity}$$

If the variance of saving will be variable, it is known as heteroscedasticity.

$$Y^{\uparrow} \rightarrow S^{\uparrow} \rightarrow \text{Var (S)} \neq \text{same} \rightarrow \text{Heteroscedasticity}$$

7.4 SOURCES OF HETEROSCEDASTICITY

The problem of heteroscedasticity in the simple linear regression model is arisen because of the following reasons.

1. The old technique of data collection:

While estimating the simple linear regression model by OLS method, the old technique has been used for collecting the data or information then the problem of heteroscedasticity creates in the simple linear regression model.

2. Presence of Outliners:

The problem of heteroscedasticity creates because of the presence of outliners. Because of it the variance of disturbance term does not fix on same.

3. Incorrect Specification of the model:

If the model (Simple linear regression model specified incorrect, the problem of heteroscedasticity arises in it.

7.5 DETECTION OF HETEROSCEDASTICITY

There are mainly five methods on tests of the detection of the problem of heteroscedasticity in the simple linear regression model. With the help of these detecting methods of heteroscedasticity, you will be able to find the problem of heteroscedasticity in the simple linear regression model.

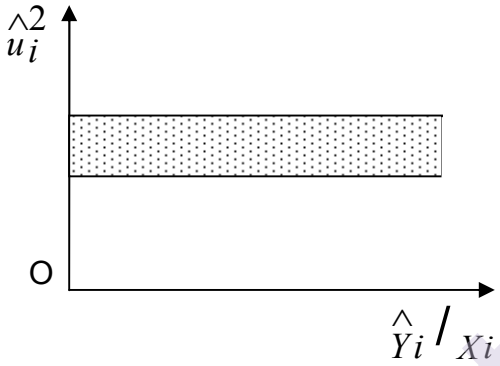
- Graphical method
- Park Test
- Glejser Test
- Spearman's Rank Correlation Test
- Goldfeld - Quandt Test.

1. GRAPHICAL METHOD:

For the detection of heteroscedasticity problem in the simple linear regression model, in this method squared residuals $(\hat{u}_i^2)$ are plotted against the estimated value of the independent variance $(\hat{Y}_i)$.

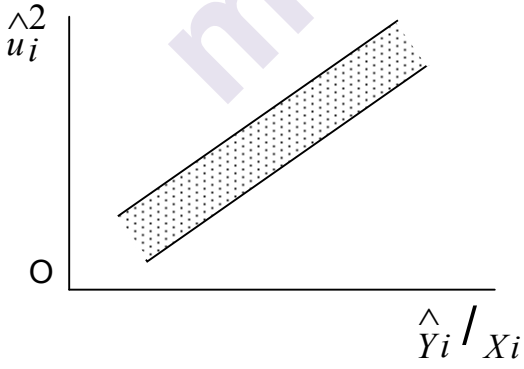
In the graphical method, there are mainly following four patterns.

i) No Systematic Pattern:



In the above graph, there is no systematic relationship between $\hat{Y}_i / X_i$ and $\hat{u}_i^2$ so, there is no heteroscedasticity.

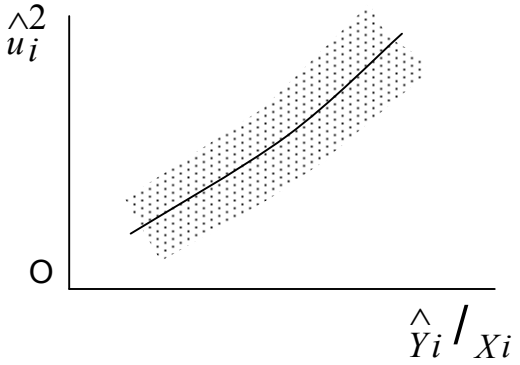
ii) Linear Pattern :



Above graph indicates the linear relationship between $\hat{Y}_i / X_i$ and $\hat{u}_i^2$ which showed the presence of the problem of heteroscedasticity.

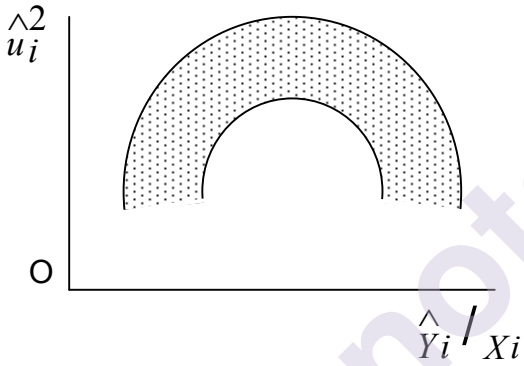
munotes.in

iii) Quadratic Pattern:



Above graph also shows, the presence of heteroscedasticity in simple linear regression model.

iv) Quadratic Pattern:



Above graph indicates that there is the present of problem of heteroscedasticity. In short, when there is the systematic relationship between $\hat{Y}i$ / Xi and $\hat{u}_i^2$ then there is the presence of heteroscedasticity.

munotes.in

2. PARK TEST:

R. E. Park developed the test for the detection of heteroscedasticity in the regression model which is known as Park Test. R. E. Park developed this test in Econometrica in article entitled 'Estimation with Heteroscedastic Error Terms' in 1976.

Park said that, $6i^2$ is the heteroscedastic variance of ui which varies and the relationship between heteroscedastic variance of residuals ($6i^2$) and explanatory variable (Xi).

$$6i^2 = 6^2 \, Xi^{\beta} \, e^{vi} \qquad - (1)$$

$$\text{In } 6i^2 = \text{In } 6^2 + \beta \text{In } Xi + Vi \qquad - (2)$$

Where,

$6i^2$ = Heteroscedastic Variance of ui

6^2 = Homoscedastic Variance of ui

X = explanatory variable

Vi = Stochastic term

If, $6i^2$ is unknownm Park suggeted $\hat{u}_i^2$ (squared regression residuals) instead of $6i^2$.

$$\ln \hat{u}_i^2 = \ln 6^2 + \beta \ln Xi + Vi \quad - (3)$$

where, $\ln 6^2 - \alpha$

$$\ln \hat{u}_i^2 = \alpha + \beta \ln Xi + Vi \quad - (4)$$

Criticisms on Park Test:

Goldfeld and Quandt criticized that Park used the, Vi Stochastic term in the process of detection of the problem of heteroscedasticity which is or can be already heteroscedastic.

But Park has shown, Vi is a stochastic term which is homoscedastic.

3. GLEJSER TEST:

H. Glejser developed the test for the detecting the heteroscedasticity in 1969 in the article entitled 'A New Test for Heteroscedasticity' in Journal of the American Statistical Association.

Glejser suggested that get the residuals value while regressing on the data and the regress on residual value, while regressing, Glejser used the following six types of functional form.

$$\left| \hat{u}_i \right| = \beta_1 + \beta_2 Xi + Vi \quad - (i)$$

$$\left| \hat{u}_i \right| = \beta_1 + \beta_2 \sqrt{Xi} + Vi \quad - (ii)$$

$$\left| \hat{u}_i \right| = \beta_1 + \beta_2 \frac{1}{Xi} + Vi \quad - (iii)$$

$$\left| \hat{u}_i \right| = \beta_1 + \beta_2 \frac{1}{\sqrt{Xi}} + Vi \quad - (iv)$$

$$\left| \hat{u}_i \right| = \sqrt{\beta_1 + \beta_2 X_i} + Vi \quad - (v)$$

$$\left| \hat{u}_i \right| = \sqrt{\beta_1 + \beta_2 X_i^2} + Vi \quad - (vi)$$

munotes.in

Above first 4 equations are linear in parameters and last 2 equations are non - linear in parameters.

Glejser suggested above 6 functional forms for testing the relationship between the stochastic term (Vi) and explanatory variable (X).

According Glejser, first four equations (1, 2, 3, 4) give the satisfied results because these are linear in parameter and last two equations (5, 6) give non - satisfied result, because these are non - linear in parameters.

Criticisms on Glejser Test :

Goldfeld and Quandt criticized on Glejser test as below –

1. Glejser suggested six functional forms, in the last two functional forms get the non – linear estimates while taking variance of ordinary least square (OLS) estimates.

2. Vi is a stochastic term which can be heteroscedastic and multicolinears and the expected value of Vi is non – zero.

$$E(Vi) \neq O$$

4. SPEARMAN'S RANK CORRELATION TEST:

This test is based on the rank correlation coefficient. That is why this test is known as Spearman's Rank Correlation Test.

Spearman's Rank Correlation Test indicates the absolute value of $\left| \hat{u}_i \right|$ and Xi. Spearman's Rank Correlation is denoted by r_s.

Symbolically,

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

Where,

r_s = Spearman's Rank Correlation Coefficient.

n = no. of pairs of observation ranked

d_i = the difference in the ranks assigned to two different characteristics of the 9^{th}.

For detecting the heteroscedasticity in the simple linear regression model, following steps has been suggested by spearman.

$$Yi = \beta_1 + \beta_2 Xi + u_i$$

This is the simple linear regression model
Steps :

i) Fit the regression to the data obtains residuals (u_i).

ii) Ignoring the sign of $\hat{u}_i$ rank $|\hat{u}_i|$ in ascending /descending form and compute.

$$t = \left[\frac{r_s - \sqrt{n-2}}{\sqrt{7 - rs^2}}\right]$$

df = n - 2

If computed value of t is greater than critical t value, there is the presence of heteroscedasticity in the simple linear regression model.
If the computed value of t is less than critical t value, there is the absence of heteroscedasticity in the simple linear regression model.

5. GOLDFELD - QUANDT TEST :

Goldfeld and Quandt developed a test to detect the problem of heteroscedasticity which is known as Goldfeld - Quandt test.

This test is depends on 'there is positive relationship between heteroscedasticity ($6i^2$) and explanatory variable (Xi).

Steps :

There are mainly following 4 steps for detecting the problem of heteroscedasticity.

1) Order or rank the observations according to the value of Xi beginning with the lowest X value.

Ex.

Yi	Xi
20	18
30	15
40	17
50	25
60	30

$\Rightarrow$

Yi	Xi
30	15
40	17
20	18
50	25
60	30

munotes.in

2) Omit central observations and divide the remaining (n - c) observations into two groups

Yi	Xi		Yi	Xi
30	15	A	30	15
40	17		40	17
20	18	⇒ ignored	50	25
50	25	B	60	30
60	30			

iii) Fit separate OLS regressions to the first observation and the last observation (B) and obtain the respective residual sums of squares RSS_1 and RSS_2.

iv) Compute the ration -

$$F = \frac{RSS_2/df}{RSS_1/df} = \frac{RSS_2}{RSS_1}$$

Calculated value of F ration at the given level of significance (α) is greater or more than given critical F value, the homoscedastic hypothesis is rejected sand heteroscedastic hypothesis is decepted.

Calculated F value	>	Critical F value	⇒	Presence of the Heteroscedasticity

6. Other Tests for detecting the problem of Heteroscedasticity

i) Breush - Pagan - Godfrey Test
ii) White's General Heteroscedasticity Test
iii) Koenker - Bassett (KB) Test.

7.6 CONSEQUENCES OF HETEROSCEDASTICITY

Consequences of using OLS for estimation of simple linear regression model in the presence of the problem of heteroscedasticity are as follows -

1. In the presence of heteroscedasticity, values of OLS estimators do not change, but it affect on variance of estimators.
2. The properties of OLS estimators which are Linearity and Unbiasedness do not change or vary in the presence of heteroscedasticity, but there is lack of minimum variance, that is why the estimators are not efficient.

munotes.in

3. Get the more confidence interval.
4. There is impossibility to test the statistics significant of parameter estimates because of the presence of heteroscedasticity.

7.7 QUESTIONS

1. Explain any two tests in detection of heteroscedasticity.
2. Explain the assumptions of OLS method of estimation of simple linear regression model.
3. What is heteroscedasticity? Explain the causes and consequences of heteroscedasticity.

7.8 REFERENCES

- Gujarati Damodar N, Porter Drawn C & Pal Manoranjan, 'Basic 5Ecometrics', Sixth Edition, Mc Graw Hill.
- Hatekar Neeraj R. 'Principles of Econometrics : An Introduction (Using R) SAGE Publications, 2010
- Kennedy P, 'A Guide to Econometrics', Sixth Edition, Wiley Blackwell Edition, 2008

munotes.in

8

PROBLEMS IN SIMPLE LINEAR REGRESSION MODEL: AUTOCORRELATION

Unit Structure:

8.0 OBJECTIVES :

1. To understand the causes of Autocorrelation.
2. To understand the detection of Autocorrelation.
3. To understand the consequences of Autocorrelation.

8.1 INTRODUCTION :

While using the OLS method for the estimation of simple linear regression model, if assumption 5 which is no autocorrelation between the disturbance terms does not fulfil, the problem of autocorrelation in the simple linear regression model arises.

8.2 AUTOCORRELATION

The autocorrelation may be defined as 'correlation between residuals disturbances (u_i, u_j).

The OLS method of estimation of linear regression model assumes that such autocorrelation does not exist in disturbances (u_i, u_j).

muinotes.in

Symbolically,

$$E(u_i, u_j) = 0$$

Here, $i \neq j$

In short, autocorrelation is a problem which creates while using the OLS method to estimate the simple linear regression model.

According to Tinmer 'autocorrelation is tag correlation between two different series.'

8.3 SOURCES OF AUTOCORRELATION

The problem of autocorrelation arises while estimating the simple linear regression model by OLS method because of the following reasons.

1. Time series that varies or changes slowly has a problem of autocorrelation.

2. If some important independent variables are omitted from the regression model, the problem of autocorrelation arises.

3. If the regression paradigm is framed in the wrong mathematical form, then the successive values of the residual become interdependent.

4. While taking averages of data, it becomes slow, that is why the disturbance term indicates the problem of autocorrelation.

5. It the calculation process is done while searching for the missing figure of the compound, this creates a problem of interdependence between them.

6. In the regression model, when is disturbance term is incorrectly arranged autocorrelation is formed.

mynotes.in

8.4 DETECTION OF AUTOCORRELATION

There are mainly three methods to detect the problem of autocorrelation as follows -

- Graphical Method
- The Runs Test
- Durbin - Watson & Test

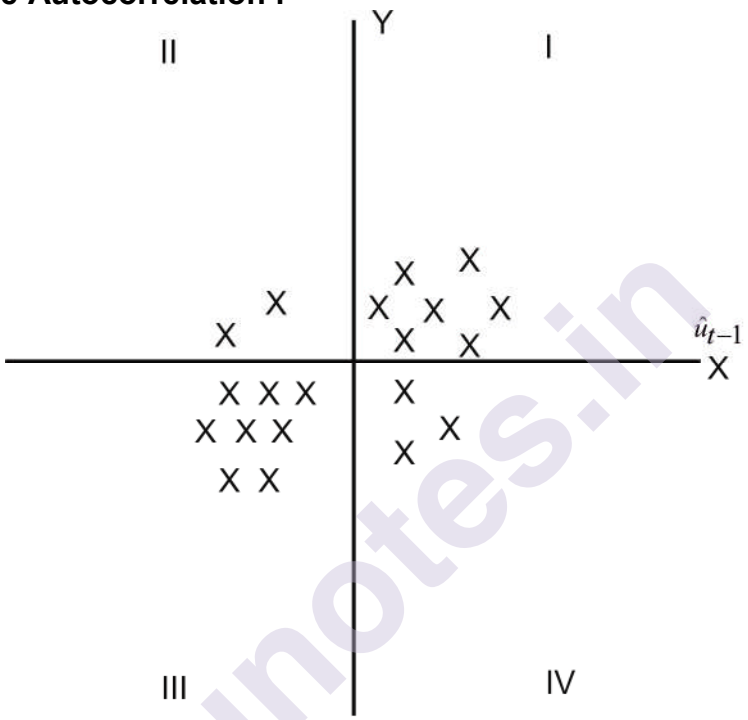
1. Graphical Method :

Whether is there the problem of autocorrelation? the answer of this question will be got by the examining the residuals.

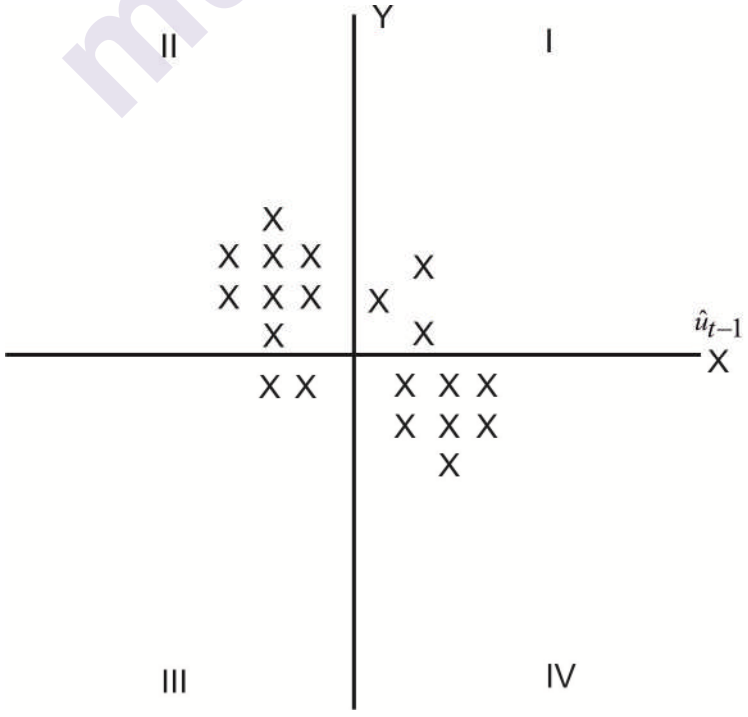
There are various ways of examining the residuals :

1) We can simply plot the residuals against time which known as the time sequence plot.
2) We can plot the standardized residuals against time and examine for detecting the problem of autocorrelation.
3) Alternatively, we can plot the residuals $\hat{u}_t$ against $\hat{u}_{t-1}$.

Positive Autocorrelation :

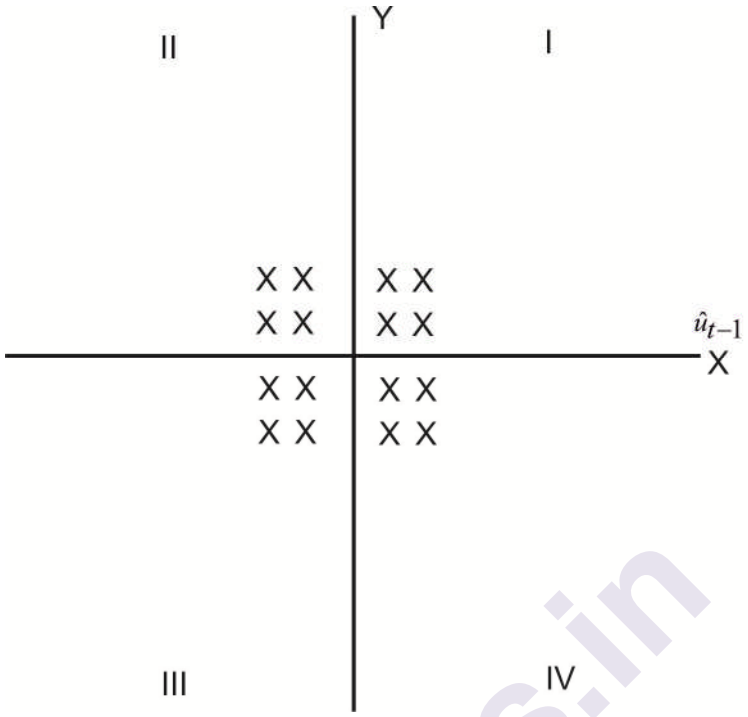


Negative autocorrelation :



munotes.in

No Autocorrelation :



- When, pairs of residuals are more in I and II quadrants, there is the presence of positive autocorrelation.
- When, pairs of residuals are more in II and IV quadrants, there is the presence of negative autocorrelation.
- When, pairs of residuals are equals in the all four quadrants, there is no presence of autocorrelation.

2. The Runs Test

The run test is developed by R. C. Geary IN 1970 in the article entitled 'Relative Efficiency of Count Sign changes for Asserting Residual Autoregression in least squares Regression' in Biometrica.

The run test is also known as Geary test and it is non - parametric test.

Suppose, there are 40 observations of residuals as follows -

(- - - - - - - - - -) (+ + + + + + + + + + +

+ + + + + + + + + +) (- - - - - - - - - -)

Thus, there are 9 negative residuals, followed by 21 positive residuals followed by 10 negative residuals, for a total of 40 observations.

First let we know the concept of run and length of run.

nuncies.in

Run:

Run is an uninterrupted sequence of one symbol of attribute such as + or - .

Length of Run:

Length of run is the number of elements in the series.

In the above series -

N = Number of total observations

N = $N_1 + N_2 = 40$

N_1 = Number of positive residuals = 21

N_2 = Number of negative residuals = 19

R = Number of Run = 3

Now taking the null hypothesis that the successive residuals are interdependent and assuming that both N_1 & N_2 ($N_1 > 10$, $N_2 > 10$) the number of runs (R) are follows normal distribution.

Mean: $E(f) = \frac{2N_1N_2}{N} + 1$

Variance: $6_R^2 = \frac{2N_1N_2\,(2N_1N_2 - N)}{(N)^2\,(N-1)}$

Now let we decide the confidence interval (CI) for R.

95% CI for R = $E(R) \pm 1.96\ 6_R$

99% CI for R = $E(R) \pm 2.56\ 6_R$

Take any confidence interval for R from above two.

Decision Rule -

If number of Runs (R) lies in the preceding confidence interval, the null hypothesis accepted.

If number of Runs (R) lies in the preceding confidence interval, the null hypothesis rejected.

When we reject the null hypothesis, it means that residuals exhibit autocorrelation and vicevarsa.

3. Durbin - Watson d Test:

The most celebrated test for detecting the autocorrelation or serial correlation which is developed by statisticians Durbin and Watson - in the article entitled 'Testing for social, correlation in least

munotes.in

squares regression in Priometrica in 1951. This test is popularly known as the Durbin - Watson d statistic test.

Durbin - Watson d statistic test as defined as -

$$d = \frac{\sum_{t=2}^{t=n} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{t=n} \hat{u}_t^{\,2}}$$

Where,

Numerator = Sum of squares of difference of continuing residuals

$$\left(\sum (\hat{u}_t - \hat{u}_{t-1})^2 \right)$$

Denominator = Sum of squared residuals

$$\left(\sum \hat{u}_t^{\,2} \right)$$

Thus, the Durbin - Watson d statistic is the ratio of sum of squares of difference between continuing two residuals $\left(\sum (\hat{u}_t - \hat{u}_{t-1})^2 \right)$ to the sum of squared residuals $\left(\sum \hat{u}_t^{\,2} \right)$.

Assumptions :

This test is based on the following some assumptions -

i) Regression model includes intercept term (β_1)

ii) Residuals follow the first order auto-regressive scheme.

$u_t = e\, u_{t-1} + v_t$

iii) This test assume that there is no tag value of dependent variable in the regression model.

$Y_t = \beta_1 + \beta_2 X_t + \beta_3 Y_{t-1} + u_t$

iv) All explanatory variables (X' s) are non - stochastic.

v) There is presence of all observations in the data.

$$d = \frac{\sum \hat{u}_t^{\,2} + \hat{u}_{t-1}^{\,2} - 2\sum \hat{u}_t\, \hat{u}_{t-1}}{\sum \hat{u}_t^{\,2}}$$

mumotes.in

Approximately, $\sum \hat{u}_t^2$ and $\sum \hat{u}_{t-1}^2$ are same.

$$d = \frac{2\sum \hat{u}_t^2 - 2\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t^2}$$

$$= 2\frac{\sum \hat{u}_t^2}{\sum \hat{u}_t^2} - \frac{2\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t^2}$$

$$= 2 - \frac{2\sum \hat{u}_t^2 \hat{u}_{t-1}}{\sum \hat{u}_t^2}$$

$$= 2 - \left(1 - \frac{\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t^2} \right)$$

Where,

$$= \frac{\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t^2} = \hat{e} = \frac{\sum XY}{\sum X^2}$$

$\therefore$ $d = 2(1 - \hat{e})$

$$-1 \leq \hat{e} \leq 1$$

The value of $\hat{e}$ is between -1 and 1 or equal to -1 and 1.

$$0 \leq d \leq 4$$

The value of d is between 0 and 4 or sometimes equal to 0 and 4.

When, $\hat{e} = 0$, $d = 2$ $\Rightarrow$ No Autocorrelation

When, $\hat{e} = 1$, $d = 0$ $\Rightarrow$ Perfect Positive Autocorrelation

When, $\hat{e} = -1$, $d = 4$ $\Rightarrow$ Perfect Negative Autocorrelation

munotes.in

How to apply this test –

1. Run the regression and obtain the residuals $(\hat{u}_t)$
2. Compute d (by using equation (1)).
3. For given sample size and given number of explanatory variables, find out the critical d_L and d_u value.
4. Then take decision about presence of autocorrelation by using following rules.

No.	If	Null Hypothesis	Decision
1.	$0 < d < d_L$	Reject	No positive autocorrelation
2.	$d_L \leq d \leq d_u$	No decision	No positive autocorrelation
3.	$4 - d_L < d < 4$	Reject	No negative autocorrelation
4.	$4 - d_u \leq d \leq 4 - d_L$	No decision	No negative autocorrelation
5.	$d_u \leq d \leq 4 - d_u$	Do no reject	No autocorrelation (Positive/Negative)

In the term no decision, Durbin – Watson test remains inconclusive. This is the limitation of this test.

8.5 CONSEQUENCES OF AUTOCORRELATION

1. When the problem of autocorrelation creates in the regression model, we get linear, unbiased and consistent parameter estimates; but we do not get minimum variance of parameter estimates.

2. In the presence of autocorrelation is regression model, we get inefficient parameter estimates.

3. Hypothesis testing becomes invalid in the case of presence of autocorrelation.

4. While estimating the regression model, variance of parameter estimates is not minimum confidence intervals are big in the presence of autocorrelation in regression model.

mynotes.in

5. If we ignore the presence of autocorrelation in the regression model, $\hat{6}^2$ becomes less identified and determination coefficient becomes over identified.

8.6 QUESTIONS

1. Explain the meaning and sources of autocorrelation.
2. Explain the detection of autocorrelation.
3. Explain the sources and consequences of autocorrelation.

8.7 REFERENCES

- Gujarati Damodar N, Porter Drawn C & Pal Manoranjan, 'Basic Ecometrics', Sixth Edition, Mc Graw Hill.
- Hatekar Neeraj R. 'Principles of Econometrics : An Introduction (Using R) SAGE Publications, 2010
- Kennedy P, 'A Guide to Econometrics', Sixth Edition, Wiley Blackwell Edition, 2008

munotes.in

9

PROBLEMS IN SIMPLE LINEAR REGRESSION MODEL: MULTICOLLINEARY

Unit Structure:

9.0 OBJECTIVES

1. To understand the causes of Autocorrelation.
2. To understand the detection of Autocorrelation.
3. To understand the consequences of Autocorrelation.

9.1 INTRODUCTION

While using the OLS method for the estimation of simple linear regression model, if assumption 10 which is no perfect multicolinearity does not fulfil, the problem of autocorrelation in the simple linear regression model arises.

9.2 MULTICOLINEARITY

You all studied the ten assumptions OLS (Ordinary Least Square) method which are also assumptions of Classical Linear Regression Model (CLRM). The tenth assumption of OLS method is that there is no perfect linear relationship among the explanatory variables (X's)

The multicolinearity is due to economist Ragner Frisch. The multicolinearity is a existence of a perfect linear relationship

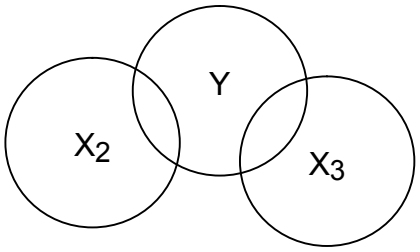
munotes.in

between the some or all explanatory variables of a regression model.

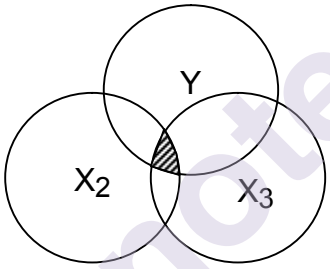
There are five types of degree of multicolinearity which have been shown in the following figures.

If we consider, there are two explanatory variables namely X_2, X_3 and Y is dependent.

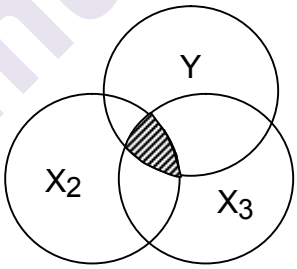
No Colinearity:



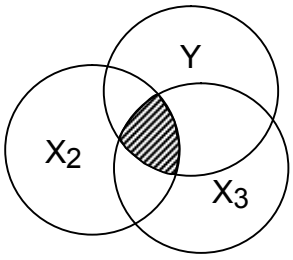
Low Colinearity:



Moderate Colinearity:

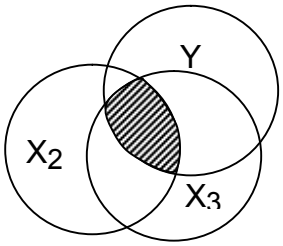


High Colinearity:



muNotes.in

Very High Colinearity:



Why the OLS method or classical linear regression model assumes that there is not existence of multicolinearity? The answer of this question is that, if the multicolinearity is perfect, the regression coefficients of the explanatory variable (X's), the regression coefficients of the explanatory variable (Xs) are indeterminate and the standard errors are infinite. And if the multicolinearity is less, the regression coefficients are determinate; possess large standard error which means that the coefficients cannot be estimated with accuracy.

9.3 SOURCES OF MULTICOLINEARITY

There are mainly four causes or sources of multicolinearity.

1. The data collection method is responsible to create the problem of s multicolinearity. For example, sampling of limited range of the values which taken by regressions in the population.

2. To constraints on the model which can be responsible to create the problem of multicolinearity.

3. Because of model specification, the problem of multicolinearity arises.

4. Because of over identified, multicolinearity arises.

These are the major causes of multicolinearity.

9.4 DETECTION OF MULTICOLINEARITY

There is no specific method available for detection of multicolinearity. Thus, following these rules are used to detect the problem of multicolinearity.

1. High R^2 but few significant – Ratio's.
2. High pair-wise correlations among regressions.
3. Examination of partial correlation.
4. Auxiliary Regression.

mumtes.in

1. High R^2 but few significant – Ratio's:

If R^2 (coefficient of determination) is high (more than 0.8), the f test in most cases will reject the hypothesis that the partial slope coefficients are simultaneously equal to zero, but individual t tests will indicate, vary few of the partial slope coefficients are statistically different from zero.

2. High pair-wise correlations among regressions:

If zero order correlation coefficient between two independent variables in the regression model is high, the nature of problem of multicolinearity is high. But high zero order correlation coefficient is not necessary condition, but it is complementary condition of the presence of multicolinearity in regression model. If there are only two explanatory variables regression model high zero order correlation coefficient is the useful method for identifying the presence of multicolinearity.

3. Examination of partial correlation:

The way or test or method of detecting the problem of multicolinearity that is examination of partial correlation has suggested Farror and Glauber. In this method, if we regress the y on X, overall coefficient determination is very high; but other partial **R^2** is comparatively small and at least one variable is unnecessary, that is the condition of the problem of multicolinearity.

4. Auxiliary Regression:

For identifying independent variables are correlated to which independent variables, we have to by regressing each independent variable $(X_i S)$. Then we have to consider the relation between F test, criterion (f_i) and coefficient of determination $\left(R_i^2\right)$ and for it following formula has been used.

$$f_i = \boxed{\frac{R_i^2/(K-2)}{\left(1-R_i^2\right)/(n-k+1)}}$$

Where,

R_i^2 = coefficient of determination for i^{th}

K = Number of explanatory variables

n = Sample size

mnotes.in

9.5 CONSEQUENCES OF MULTICOLLINEARITY

Consequences of the term multicolinearity are as follows:

1) OLS estimators show the BLUE properties, but variance & covariance are very high.
2) Confidence intervals are so wider because of the high variance and covariance. So, null hypothesis (H_O) does not accept easily.
3) t-ratio to one or more than one coefficients is not statistically significant because of high variance and co-variance.
4) If t-ratios for one or more than coefficients are not statistically significant, but we get very high value of R^2.
5) In the presence of multicolinearity, estimators and its standard errors can respond also to the small change or variation in the data.
6) There is exactly linear correlation in the explanatory variables in the model. So regression coefficients are indeterminate and standard errors are infinite.
7) If there is imperfect linear correlation between explanatory variable in the explanatory variables in the model and regression coefficient are determinate, but standard errors are so high.

9.6 SUMMARY

When we consider the linearity in simple regression model or two variable models, it is called as simple linear regression model.

There are two ways or methods for estimating the simple linear regression model. When we use the ordinary least square (OLS) method for the estimation of simple linear regression model; homoscedaticity or equal variance of u_i, no autocorrelation between the disturbance terms and no prefect multicolinearity these three assumption are unable to fulfil, sequentially the problem of heteroscedasticity, autocorrelation and multicolinearity raise which has been discussed in this unit.

9.7 QUESTIONS

1. Explain the meaning and sources of multicolinearity.
2. Explain the detection of multicolinearity.
3. Explain the sources and consequences of multicolinearity.

mjnotes.in

9.8 REFERENCES

- Gujarati Damodar N, Porter Drawn C & Pal Manoranjan, 'Basic Ecometrics', Sixth Edition, Mc Graw Hill.
- Hatekar Neeraj R. 'Principles of Econometrics : An Introduction (Using R) SAGE Publications, 2010
- Kennedy P, 'A Guide to Econometrics', Sixth Edition, Wiley Blackwell Edition, 2008

❖❖❖❖

munotes.in