19BCE226 - Ansh Ray
2HSOE52 - Introduction to Econometrics

M T W T F S S | Page No.: 1 | Date: | YOUVA

Innovative Assignment.

Q1 Discuss assumptions of classical linear regression model.

Ans ⇒ Alike many statistical analyses, Ordinary Least Squares (OLS) regression has many underlying assumptions. When these assumptions for linear regression are true, OLS produces the best estimates.

→ The assumptions are mentioned below :-

1) The regression model is linear in parameters

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

→ In this equation, the betas are the parameters that OLS estimates.

→ The defining characteristic of linear regression is this functional form of the parameters rather than ability to model curvature.

→ The equations may/may not be linear in variables and can be non-linear variables by including polynomials & transforming exponential functions, but this equation has to be linear in β_1 & β_2.

2) X values are fixed in repeated sampling.

→ Values taken by the regressor X are considered fixed in repeated samples.

→ More technically, X is assumed to be nonstochastic.

→ Also our regression analysis is conditional regression analysis, i.e. conditional on the given values of the regressor(s) X.

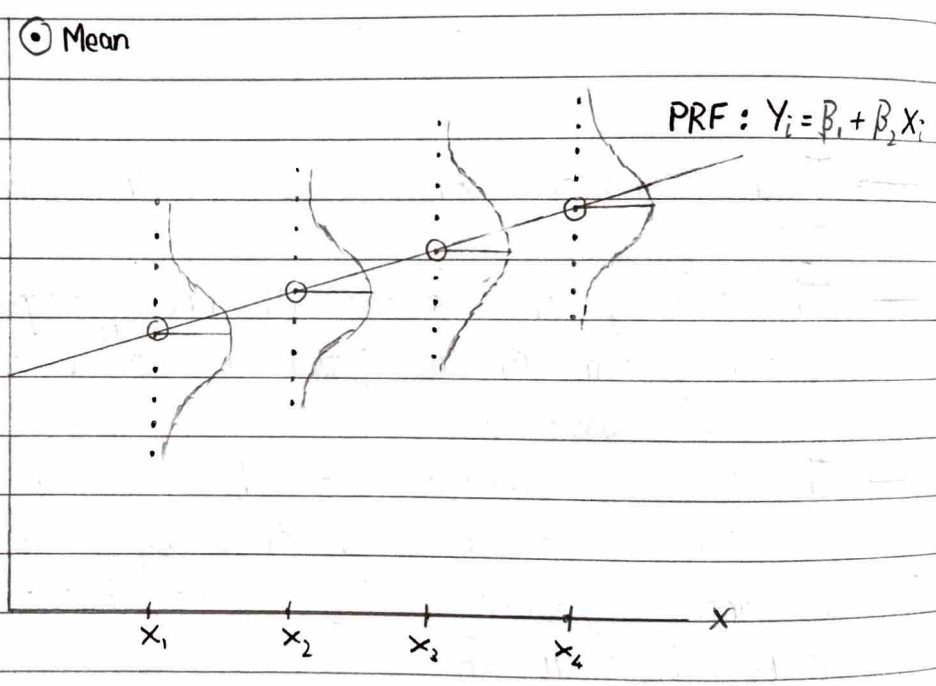
3) Zero mean value of disturbance u_i.

→ If given the value of X, the mean, or expected, value of the random disturbance term u_i is zero.
→ Technically, the conditional mean value of u_i is zero.

→ Symbolically, it is represented as

$$E(u_i / X_i) = 0$$

→ Geometrically this assumption can be presented as,



→ This figure shows a few values of variable X & the Y populations associated with each of them.
→ As shown, each Y population corresponding to a given X is distributed around its mean value with some Y values above the mean & some below it.
→ All that this assumption says is that the factors not explicitly included in the model, and therefore subsumed in u_i, do not systematically affect the mean value of Y; so, the positive u_i values cancel out the

negative u_i values so that their average or mean effect on Y is zero.

4) Homoscedasticity or equal variance of u_i.

→ Given the value of X, the variance of u_i is the same for all observations. That is, the conditional variances of u_i are identical.

→ Symbolically, we have

$$\text{var}(u_i/X_i) = E[u_i - E(u_i/X_i)]^2$$

$$= E(u_i^2/X_i) \quad \text{due to assumption 3.}$$

$$= \sigma^2 \quad \text{, where var stands for variance.}$$

5) No autocorrelation between the disturbances.

→ Given any two X values, X_i & X_j $(i \neq j)$, the correlation between any two u_i & u_j $(i \neq j)$ is zero.

→ Symbolically,

$$\text{cov}(u_i, u_j/X_i, X_j) = E\{[u_i - E(u_i)]/X_i\}\{[u_j - E(u_j)]/X_j\}$$

$$= E(u_i/X_i)(u_j/X_j)$$

$$= 0$$

where i & j are two different observations & where cov means covariance.

→ This assumption means that, given X_i, the deviations of any two Y-values from their mean value do not exhibit patterns.

6) Zero covariance between u_i & X_i or $E(u_i X_i) = 0$. Formally,

$$Cov(u_i, X_i) = E[u_i - E(u_i)][X_i - E(X_i)]$$

$$= E[u_i(x_i - E(x_i))], \text{ since } E(u_i) = 0$$

$$= E(u_i X_i) - E(x_i)E(u_i), \text{ since } E(x_i) \text{ is nonstochastic}$$

$$= E(u_i X_i), \text{ since } E(u_i) = 0$$

$$= 0 \quad \text{by assumption.}$$

→ This assumption states that the disturbance u & explanatory variable X are uncorrelated.

7) The number of observations n must be greater than the number of parameters to be estimated.

→ Alternatively, the number of observations n must be greater than the number of explanatory variables.

8) Variability in X values.

→ The X values in a given sample must not all be the same.
→ Technically, var(X) must be a finite positive number.
→ If all the X values are identical, then $X_i = \bar{X}$ & the denominator of that equation will be zero, making it impossible to estimate β_2 & consequtively β_1.

9) The regression model is correctly specified.

→ Alternatively, there is no specification bias or error in the model used in empirical analysis.

→ Questions such as by omitting important variables from the models, or by choosing wrong functional form, or by making worse stochastic assumptions about variables of the model, the validity of interpreting the estimated regression will be highly questionable.

→ Unfortunately, in practice one rarely knows which are the correct variables & which one to include in model.

→ Therefore, in practice one has to use some judgement in choosing number of variables entering the model & functional form of the model & has to make some assumptions about the stochastic nature of variables included in the model.

→ This assumption is there to remind us that our regression analysis & therefore the results based on the analysis are conditional upon the chosen model & to warn us that we should give very careful thought in formulating econometrics model especially theories trying to explain economic phenomenon such as inflation rate, or demand for money etc.

10) There is no perfect multicollinearity.

→ There are no perfect relationships among the explanatory variables

→ Multicollinearity can be checked by making a correlation matrix.

→ Almost a sure indication of the presence of multi-collinearity is when you get opposite signs for your regression coefficients.

→ It is highly likely that regression suffers from multi-collinearity

→ If the variable is not important then dropping that variable

or any of the correlated variables can fix the problem.

Q2 What is hypothesis & explain the procedure for hypothesis testing.

Soln A Hypothesis is a tentative statement about the relationship between 2 or more variables.

→ Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter.

→ A research hypothesis refers to a tentative solution to a problem which is framed in advance before the collection & analysis of data for the given objectives.

→ Once the researcher identifies & defines the research problem in a precise manner, he can make a guess as to the possible answers.

→ These guesses, which are assumed by the researcher for solving the problem or using as guide for further investigation, are called hypotheses.

⟹ Types of Hypothesis

→ Six forms of hypothesis are

i) Simple Hypothesis
ii) Complex Hypothesis
iii) Directional hypothesis
iv) Non - Directional hypothesis
v) Null hypothesis
vi) Associative & causal hypothesis.

⟹ Essentials of a good hypothesis

A good usable hypothesis is the one which satisfies the following criteria.

i) A hypothesis should be empirically testable & able to deduce logical inferences.
ii) Hypothesis should be closest to the things observable & it should enable a researcher to reach at correct decision
iii) It should be conceptually clear so as to explain the concept & leaving no scope for ambiguity.
iv) The hypothesis must be specific but not in general terms.

⇒ Procedure for testing a hypothesis.

→ After having completed collection, processing & analysis of data a test procedure has to be followed for determining if the null hypothesis is to be accepted or rejected.
→ The test procedure or the rule is based upon a test statistic & a rejection region.
→ The procedure of testing hypothesis is briefly described below.

i) Setting up a hypothesis

→ Generally there are 2 forms of hypothesis which must be constructed; and if one is accepted, the other one is rejected.

i) Null hypothesis

→ Any hypothesis concerned to a population is called

statistical hypothesis.

→ In the process of statistical test, the rejection or acceptance of hypothesis depends on sample drawn from population

→ The simple hypothesis states that the statistical measures of sample & those of the population under study do not differ significantly.

→ Similarly it may assume no relationship or association between 2 variables or attributes.

→ For example if we want to find out whether extra coaching has benefitted the students or not, the null hypothesis would be

H_0 :- The extra coaching class has not ~~bef~~ beniffited the students.

ii) Alternative Hypothesis :-

→ As against the null hypothesis, the alternative hypothesis specifies those values that the researcher believes to hold true, & he hopes that the sample data lead to acceptance of this hypothesis as true.

→ Rejection of Null hypothesis H_0 leads to the acceptance of alternative hypothesis which is denoted by H_1.

→ For same example we can write alternative hypothesis as

H_1 :- The extra coaching class has beniffited the students.

→ Null & Alternative hypothesis can also be written as

$H_0 : (\sigma_1 - \sigma_2 = 0)$, $H_0 : (\mu_1 - \mu_2 = 0)$
$H_1 : (\sigma_1 - \sigma_2 \neq 0)$, $H_1 : (\mu_1 - \mu_2 \neq 0)$

⇒ Type I & II errors

→ While we make a decision on the basis of data analysis & testing of the significant difference, it may lead to wrong conclusions in 2 ways.

i) Rejecting a true null hypothesis.
ii) Accepting a false Hypothesis.

	Decision based on Sample	
	Accepted H_0	Rejected H_0
H_0 true (H_a false)	Correct decision	Wrong decision (Type I error) $= \alpha$
H_0 false (H_a true)	Wrong decision (Type II error) $\beta = 1 - \alpha$	Correct decision

2) Setting up a suitable significance level

→ The max^m possibility of committing type I error, which we use to specify test is known as level of significance.
→ Generally 5% level of significance is fixed in statistical tests.
→ This implies we can have 95% confidence in accepting a hypothesis or could be wrong 5% in taking the decision.
→ The range of variation has 2 regions, acceptance region & rejection region or critical region.

→ The critical region under a normal curve can be divided into 2 ways

a) 2 sides under a curve (2-tailed test)
b) 1 side under a curve (1-tailed test) either right or left tail.

3) Setting a test Criterion :-

→ This involves selecting an appropriate probability fun^n for the particular test, that is, a probability distribution which can be properly applied.
→ Some probability distributions that are commonly used in testing procedures are Z, t, F & χ^2.

4) Computation

→ This step includes computing various measures from a random sample of size n, which are necessary for applying the test.
→ These calculations include the test statistic & standard error of the test statistic.

5) Making a decision or conclusion.

→ The decision is based on the computed value of test statistic.
→ If the computed value of the test statistic falls in the acceptance region, the null hypothesis is accepted.
→ On the contrary, if the computed value of test statistic is greater than critical value, then the computed value falls in rejection region & the null hypothesis is rejected.

Q3 What is Gauss - Markov theorem?

Sol^n The Gauss - Markov theorem states that if your linear regression model satisfies the classical assumptions, then OLS regression produces unbiased estimates that have the smallest variance of all possible linear estimators.

→ To comprehend this theorem, let us ~~understand~~ consider the best - linear unbiased property of an estimator.

→ We know that an estimator (Lets say $\hat{\beta}_2$) is said to be a best linear unbiased estimator of that particular OLS estimator if the following hold :-

i) It is linear, i.e., a linear function of a random variable, such as the dependent variable Y in the regression model

ii) It is unbiased, i.e. its average or expected value, $E(\hat{\beta}_2)$ is equal to the true value, β_2.

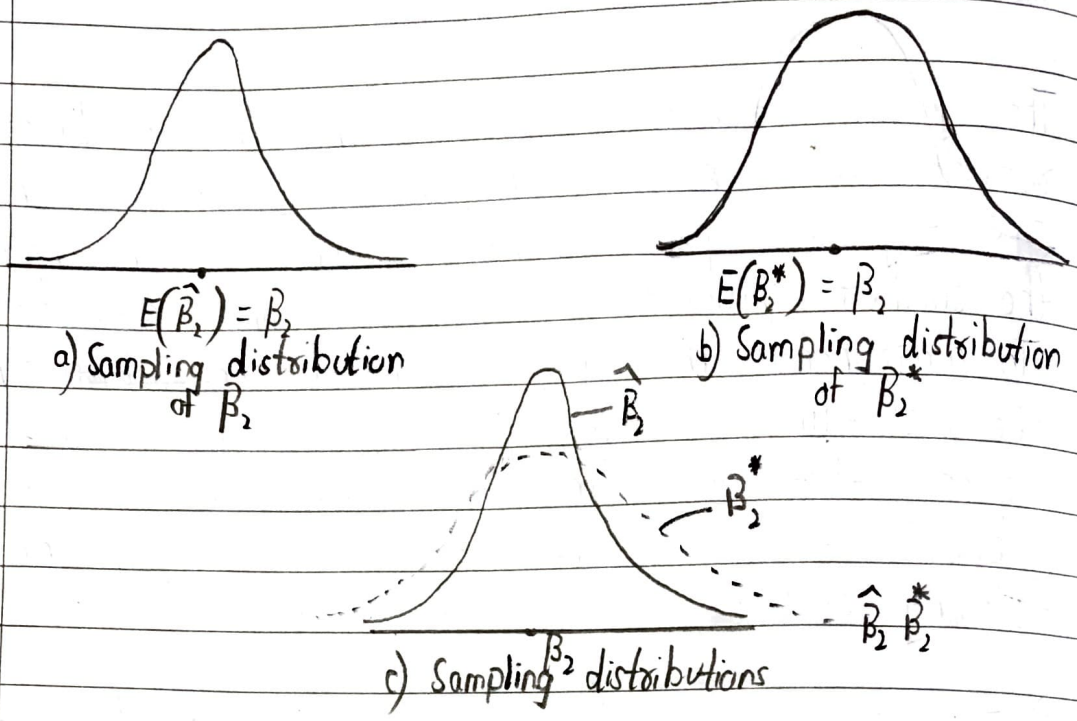
iii) It has minimum variance in the class of all such linear unbiased estimators; an unbiased with the least variance is known as an efficient estimator.

→ In the regression context it can be proved that the OLS estimators are BLUE.

[redacted]

[redacted]

→ It is sufficient to note here that the theorem has theoritical as well as practical importance.



a) Sampling distribution of $\hat{\beta}_2$

b) Sampling distribution of β_2^*

c) Sampling distributions

→ In figure (a), ~~we have shown~~ sampling distribution of the OLS estimator $\hat{\beta}_2$, i.e., the distribution of the values taken by $\hat{\beta}_2$ in repeated sampling experiments

→ Here for convenience we have assumed $\hat{\beta}_2$ to be distributed symetrically.

→ In figure (b), sampling distribution of β_2^*, an alternative estimator of β_2, obtained by using another method.

→ Here assume β_2^* is unbiased & further assume both $\hat{\beta}_2$ & β_2^* are linear estimators.

→ Now to resolve the dilemma of which estimator to choose, we have superimposed figure a) & b)

→ As β_2^* is more diffused than $\hat{\beta}_2$ in figure c) we can say that variance of β_2^* is greater than variance of $\hat{\beta}_2$.

→ So choosing estimator with smaller variance is more feasible & preferable. In short, we will choose BLUE estimator.

→ The Gauss - Markov theorem is remarkable, as it makes no assumptions about the probability distribution of the random variable u_i, & therefore of Y_i.

→ If one or more assumptions of the CLRM do not hold then the theorem is invalid.

→ As long as the assumptions of CLRM are satisfied, the theorem holds.