ECONOMETRICS

BASICS OF ECONOMETRICS AND ITS SCOPE.

Introduction

WHAT IS ECONOMETRICS?

Econometrics refers to the application of economic theory and statistical techniques for the purpose of testing hypothesis and estimating and forecasting economic phenomenon. Literally interpreted, econometrics means "economic measurement." Although measurement is an important part of econometrics, the scope of econometrics is much broader, as can be seen from the following quotations: Econometrics, the result of a certain outlook on the role of economics, consists of the application of mathematical statistics to economic data to lend empirical support to the models constructed by mathematical economics and to obtain numerical results. econometrics may be defined as the quantitative analysis of actual economic phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference. Econometrics may be defined as the social science in which the tools of economic theory, mathematics, and statistical inference are applied to the analysis economic phenomena. Econometrics is concerned with the empirical determination of economic laws.

BASIC ECONOMETRICS

The art of the econometrician consists in finding the set of assumptions that are both sufficiently specific and sufficiently realistic to allow him to take the best possible advantage of the data available to him. Econometricians are a positive help in trying to dispel the poor public image of economics (quantitative or otherwise) as a subject in which empty boxes are opened by assuming the existence of can-openers to reveal contents which any ten economists will interpret in 11 ways. The method of econometric research aims, essentially, at a conjunction of economic theory and actual measurements, using the theory and technique of statistical inference as a bridge pier.

Objectives

1. plications of economic theory need a responsible understanding of economic relationships and econometrics method.

2. The econometrics theory thus becomes a very powerful tool for understanding of the applied economic relationships and for meaningful research in economics.

3. In this unit we learn basic theory of econometrics and relevant application of the method.

Methodology of Econometrics:

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

1. Statement of theory or hypothesis.

2. Specification of the mathematical model of the theory

3. Specification of the statistical, or econometric, model

4. Obtaining the data

5. Estimation of the parameters of the econometric model

6. Hypothesis testing

7. Forecasting or prediction

8. Using the model for control or policy purposes.

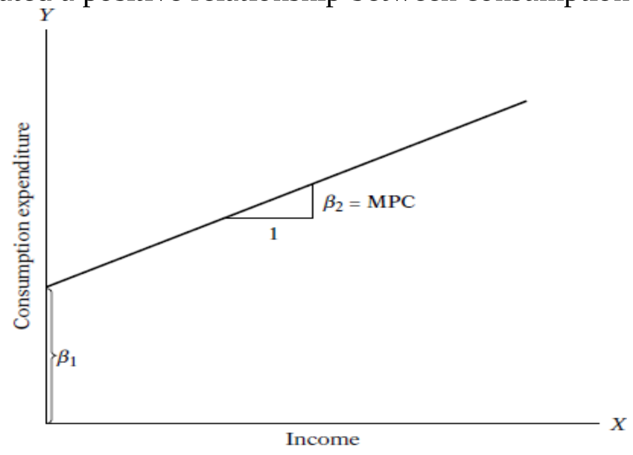
To illustrate the preceding steps, let us consider the well-known Keynesian theory of consumption:

1. Statement of theory or Hypothesis

Keynes postulated that Marginal propensity to consume (MPC), the rate of change of consumption for a unit, change in income, is greater than zero but less than one. i.e., $0 < MPC < 1$

2. Specification of the Mathematical Model of Consumption

Keynes postulated a positive relationship between consumption and income.



Keynesian consumption function.

The slope of the coefficient β_2 measures the MPC.

Keynesian consumption function

$$Y = \beta_1 + \beta_2 \mathbf{x} \quad O < \beta_2 < 1$$

Y = Consumption expenditure

X = Income

$\beta_1 \mathbf{x} \beta_2$ are knows as the parameters of the model and are respective, the interest and slope of coefficient.

Shows exact and determined relationship between consumption and income.

The slope of the coefficient β_2, measures the MPC.

Equation states that consumption is linearly related to income (Example of a mathematical model of the relationship between consumption and income that is called consumption function in economic).

Single or one equation is known as single equation model and more than one equation is known as multiple equation model.

3. Specification of the econometric model of consumption.

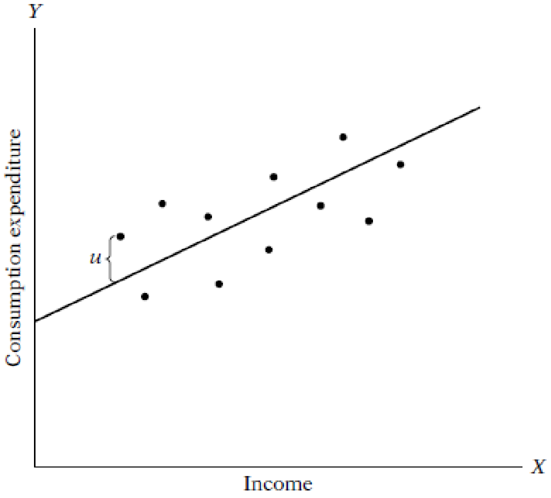
The inexact relationship between economic variables, the econometrician would modify the deterministic consumption function as.

$$Y=\beta_1+\beta_2\mathbf{x}+U$$

This equation is an example of the econometric model. More technically, it is an ex. of linear regression model.

This you may be well represent all those factors that affect consumption but are not taken into account explicitly.

The econometric consumption function hypothesizes that the dependent variable Y (consumption) is linearly related to the explanatory variable X (Income) but that is the relationship between. The two is not exact, it is subject to individual variation.



Econometric model of the Keynesian consumption function.

Q: Why inexact (not exact) relationship exits?

A: Because in addition to income, other variables affect consumption expenditure. For ex. are of family, ages of members of family, religion etc are likely to exert some influence on consumption.

4. Original Data

To obtain the numerical values of $\beta_1 \& \beta_2$ we need data.

{PCE Personal consumption expenditure)

Y variable in this table is the aggregate PCE &xis GD a measure of aggregate income.

Note: MPC: Average change in consumption over to change in real income.

5. Estimation of the Econometric Model

The statistical technique of regression analysis is the main tool used to obtain the estimates.

The estimated consumption function

$$\hat{Y}=\hat{\beta}_1+\hat{\beta}_2 x_i$$

$\hat{Y}=$ **Estimate of Y** The estimated consumption function (i.e., regress line).

Regression Analysis is used to obtain estimates.

6. Hypothesis Testing:

Keynes expected the MPC is positive but less than 1.

Confirmation or refulation of economic theories on the basis of sample evidence is based on a branch of statistical theory known as statistical inference (hypothesis testing)

7. Forecasting or Prediction

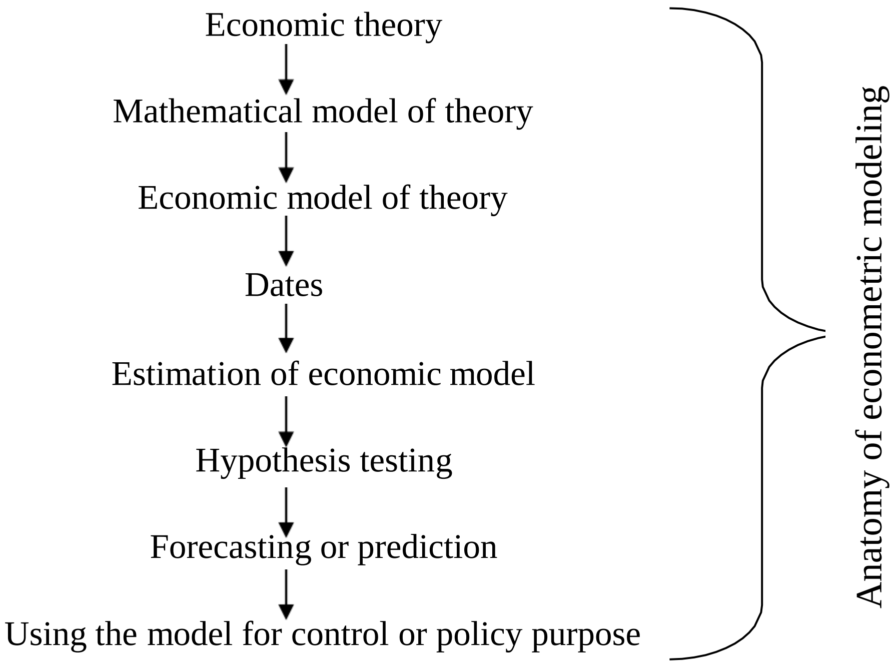
If the chosen model does refute the hypothesis or theory under consideration, we may use it to predict the future value(s) of the dependent, or forecast, variable Y on the basis of known or expected future value(s) of the explanatory, or predictor variable X.

Macroeconomic theory shows, the change in income following change in investment expenditure is given by the income multiplier M.

$$M=\frac{1}{1-MP}$$

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

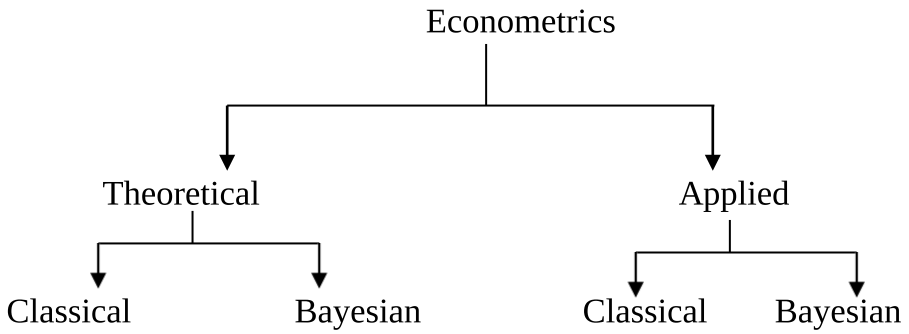
8. Use of the Model for control or Policy purpose


Economic theory
Mathematical model of theory
Economic model of theory
Dates
Estimation of economic model
Hypothesis testing
Forecasting or prediction
Using the model for control or policy purpose
Anatomy of econometric modeling

Note:

- Milton Friedmen has developed a model of consumption theory permanent income hypothesis.
- Robert Hall has developed a model of consumption as life cycle permanent income hypothesis

Types of Econometrics



- Theoretical econ is concerned with the development of appropriate methods of measuring economic relationship specified by economic models.
- Applied econ uses the tool of theoretical econ to study some special fields of eco and business, such as production function etc.

SUMMARY AND CONCLUSIONS:

Econometrics is an amalgam of economic theory, mathematical economics, economic statistics, and mathematical statistics. Yet the subject deserves to be studied in its own right for the following reasons.

Economic theory makes statements or hypotheses that are mostly qualitative in nature. For example, microeconomic theory states that, other things remaining the same, a reduction in the price of a commodity is expected to increase the quantity demanded of that commodity. Thus, economic theory postulates a negative or inverse relationship between the price and quantity demanded of a commodity. But the theory itself does not provide any numerical measure of the relationship between the two; that is, it does not tell by how much the quantity will go up or down as a result of a certain change in the price of the commodity. It is the job of the econometrician to provide such numerical estimates. Stated differently, econometrics gives empirical content to most economic theory.

The main concern of mathematical economics is to express economic theory in mathematical form (equations) without regard to measurability or empirical verification of the theory. Econometrics, as noted previously, is mainly interested in the empirical verification of economic theory. As we shall see, then econometrician often uses the mathematical equations proposed by the mathematical economist but puts these equations in such a form that they lend themselves to empirical testing. And this conversion of mathematical into econometric equations requires a great deal of ingenuity and practical skill.

Introduction:

The term *regression* was introduced by Francis Galton. In a famous paper, Galton found that, although there was a tendency for tall parents to have tall children and for short parents to have short children, the average height of children born of parents of a given height tended to move or "regress" toward the average height in the population as a whole.1 In other words, the height of the children of unusually tall or unusually short parents tends to move toward the average height of the population. Galton's *law of universal regression* was confirmed by his friend Karl Pearson, who collected more than a thousand records of heights of members of family groups.2 He found that the average height of sons of a group of tall fathers was less than their fathers' height and the average height of sons of a group of short fathers was greater than their fathers' height, thus "regressing" tall and short sons alike toward the average height of all men. In the words of Galton, this was "regression to mediocrity."

THE MODERN INTERPRETATION OF REGRESSION

The modern interpretation of regression is, however, quite different. Broadly speaking, we may say Regression analysis is concerned with the study of the dependence of one variable,the *dependent variable,* on one or more other variables, the *explanatory variables*,with a view to estimating and/or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling)
values of the latter.

Objectives:

1. The key objective behind regression analysis is the statistical dependence of one variable, the dependent variable, on one or more other variables, the explanatory variables.

2. The objective of such analysis is to estimate and/or predict the mean or average value of the dependent variable on the basis of the known or fixed values of the explanatory variables.

3. In practice the success of regression analysis depends on the availability of the appropriate data.

4. In any research, the researcher should clearly state the sources of the data used in the analysis, their definitions, their methods of collection, and any gaps or omissions in the data as well as any revisions in the data.

5. The data used by the researcher are properly gathered and that the computations and analysis are correct.

WHAT IS REGRESSION ANALYSIS:

Under single regression model one variable, called the dependent variable is expressed as a linear function of one or more other variable, called explanatory variable.

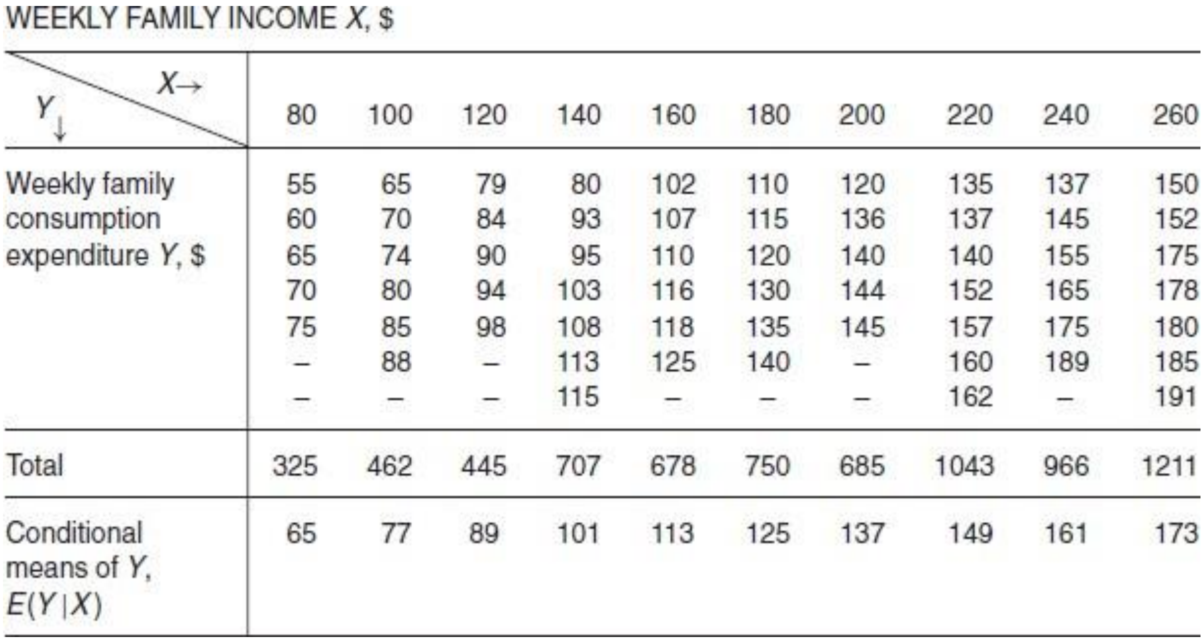
TWO VARIABLE REGRESSION MODEL ANALYSIS:

That means a function has only one dependent variable and only one independent variable.

Two variable or bivariate

Means regression in which the dependent variable (the regressand) is related to a single explanatory variable (the regression).

When mean values depend upon conditioning (variable X) is called conditional expected value. Regression analysis is largely concerned with estimating and/or predicting the (population) mean value of the dependent variable on the basis of the known or fixed values of the explanatory variable (s).



WEEKLY FAMILY INCOME X, $

$Y\downarrow$ $X\rightarrow$	80	100	120	140	160	180	200	220	240	260
Weekly family consumption expenditure Y, $	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	–	88	–	113	125	140	–	160	189	185
	–	–	–	115	–	–	–	162	–	191
Total	325	462	445	707	678	750	685	1043	966	1211
Conditional means of Y, $E(Y\mid X)$	65	77	89	101	113	125	137	149	161	173

To understand this, consider the data given in the below table. The data in the table refer to a total population of 60 families in a hypothetical community & their weekly income (X) and weekly consumption expenditure (Y), both in dollars. The 60 families are divided into 10 income groups (from $80 to $260) and the weekly expenditures of each family in the various groups are as shown in the table. Therefore, we have 10 fixed values of X and the corresponding Y values against each of the X values; and hence there are 10 Y subpopulations. There is considerable variation in weekly consumption expenditure in each income group, which can be seen clearly but the general picture that one gets is that, despite the variability of weekly consumption expenditure within each income bracket, on the average, weekly consumption expenditure increases as income increases. To see this clearly, in the given table we have given the mean, or average, weekly consumption expenditure corresponding to each of the 10 levels of income. Thus, corresponding to the weekly income level of $80, the mean consumption expenditure is $65, while corresponding to the income level of $200, it is $137. In all we have 10 mean values for the 10 subpopulations of Y. We call these mean values conditional expected values, as they depend on the given values of the (conditioning) variable X. Symbolically, we denote them as $E(Y \mid X)$, which is read as the expected value of Y given the value of X.

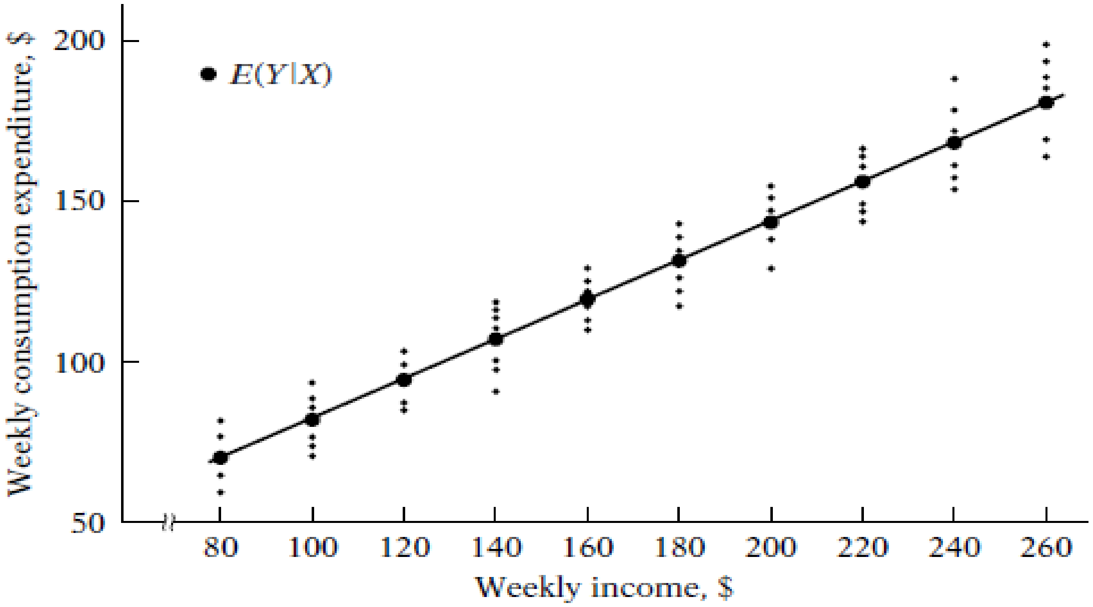


fig.: Conditional distribution of expenditure for various levels of income

It is important to distinguish these conditional expected values from the unconditional expected value of weekly consumption expenditure, $E(Y)$. If we add the weekly consumption expenditures for all the 60 families in the population and divide this number by 60, we get the number $121.20 ($7272/60), which is the unconditional mean, or expected, value of weekly consumption expenditure, $E(Y)$; it is unconditional in the sense that in arriving at this number we have disregarded the income levels of the various families. Obviously, the various conditional expected values of Y given in given table are different from the unconditional expected value of Y of $121.20. When we ask the question, "What is the expected value of weekly consumption expenditure of a family," we get the answer $121.20 (the unconditional mean). But if we ask the question, "What is the expected value of weekly consumption expenditure of a family whose monthly income is, differently, if we ask the question, "What is the best (mean) prediction of weekly expenditure of families with a weekly income of $140," the answer would be $101. Thus the knowledge of the income level may enable us to better predict the mean value of consumption expenditure than if we do not have that knowledge. This probably is the essence of regression analysis, as we shall discover throughout this text.

The dark circled points in figure show the conditional mean values of Y against the various X values. If we join these conditional mean values, we obtain what is known as the population regression line (PRL), or more generally, the population regression curve**.** More simply, it is the regression of Y on X. The adjective "population" comes from the fact that we are dealing in this example with the entire population of 60 families. Of course, in reality a population may have many families.

Geometrically, then, a population regression curve is simply the locus of the conditional means of the dependent variable for the fixed values of the explanatory variable(s). More simply, it is the curve connecting the means of the subpopulations of Y corresponding to the given values of the regressor X. It can be depicted as in figure.

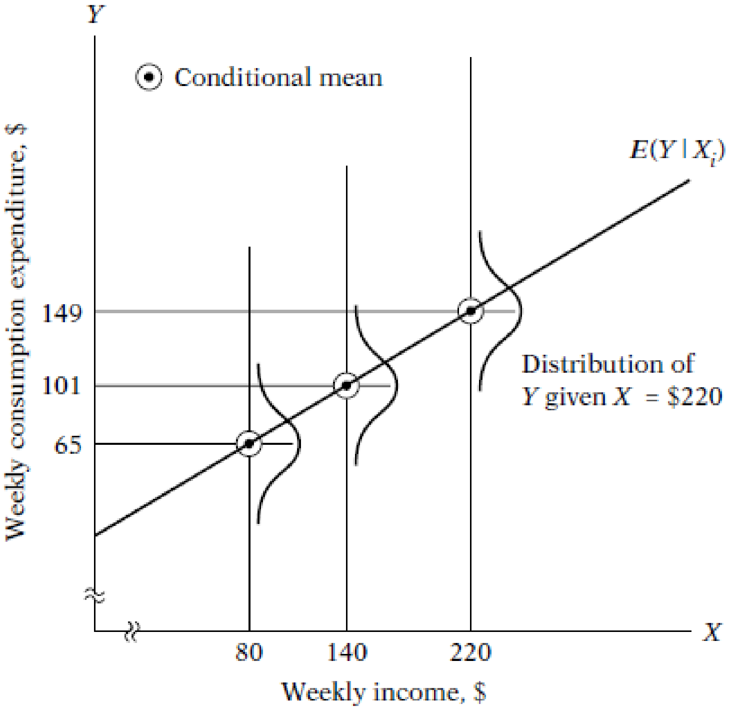


Fig.: Population Regression line.

This figure shows that for each X (i.e., income level) there is a population of Y values (weekly consumption expenditures) that are spread around the (conditional) mean of those Y values.

For simplicity, we are assuming that these *Y* values are distributed symmetrically around their respective (conditional) mean values. And the regression line (or curve) passes through these (conditional) mean values.

Concept of Population Regression function (PRF) Or Conditional Expectation function

$$\Sigma(Y/X_i)=f(x_i)$$

$f(X_i)$: Some function of the explanatory variable X

$\Sigma(Y/X_i)$: Linear function of X_i

$$\Sigma(Y/X_i)=\beta_1+\beta_2X_i$$

$\beta_1 \& \beta_2$ are unknown but fixed parameters known as the regression coefficients are also known as intercept and slope coefficient.

In regression analysis our interest is in estimating the PRFs.

ESTIMATION THROUGH OLS

Properties of OLS:

1) Our estimation are expressed solely in term of observatory can be easily complete.

2) They are point estimation.

3) Once OLS estimation is obtained from the sample data. The sample regression line can be easily obtained.

$$Y_i=(b_0+b_1x_{1i}+b_2x_{2i})+(u_i)$$

Assumptions of Model

1) Variable u is real random variable.

2) Homoscedasticity

$$E(u_1^2)=\sigma^2$$

3) Normality of u

$$u \sim N(O,\sigma_0^2)$$

4) Non auto correlation

$$E(u_i u_j)=u \quad i \neq j$$

5) Zero mean of u

$$E(u_i)=0$$

6) Independence of u_i and X_i.

$$E(u_i / x_{1i})=E(u_i X_{2i})=0$$

7) No perfect multicollinear X's

8) No error of measurement in the X's.

Estimation through OLS

$$\hat{Y}_i=\hat{\beta}_1+\hat{\beta}_2 X_i+\hat{u}$$

$$Y_i=\hat{Y}+\hat{u}$$

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

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

$$\therefore \quad \sum \hat{u}_i^2=Y_i-\hat{Y}$$

$$(Y_i-\hat{u}_i=\hat{Y}_i)$$

$$(\hat{\beta}_1+\hat{\beta}_2 X_i+\hat{u}_i-\hat{u}_i=\hat{Y})$$

$$(\hat{Y}_i=\hat{\beta}_1+\hat{\beta}_2 X_i)$$

Sq. them we get variation of deviation

$$\hat{u} = (Y_i - \hat{Y}_i)^2$$

$$\sum \hat{u}_i^2 = (Y_i - \hat{Y}_i)^2$$

$$\sum \hat{u}_i^2 = \sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

$$\frac{\delta \sum \hat{u}_i^2}{\delta \beta_1} = 2\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

$$\sum = \sum(\hat{\beta}_1 - \hat{\beta}_2 X_i)$$

$$\sum = n\hat{\beta}_1 - \hat{\beta}_2 \sum X_i$$ n= sample size

$$\frac{\delta \sum \hat{u}_i^2}{\delta \beta_2} = 2\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)(X_i) = 0$$

$$X_i \sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

$$\sum X_i = X_i \sum(\hat{\beta}_1 - \hat{\beta}_2 X_i)$$

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

Note:- We are not taking n β_2 because one variable X_1 is already percent. So no need for n, co$_2$ they are one & the same.

(LRM) = Classical linear regression Modes) Normal equation Y is dependent upon X. X is independent.)

Q. Find the value of $\hat{\beta}_1 \& \hat{\beta}_2$

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_i \qquad \rightarrow \qquad (1)$$

$$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i \qquad \rightarrow \qquad (2)$$

$$\sum X_i Y_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2 \rightarrow \qquad (3)$$

Dividing equator (2) by n

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

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}$$

$$\hat{\beta}_1 = \hat{\beta}_2 X - \bar{Y}$$

Now after further simplification we get the value of $\hat{\beta}_2$ as

$$\hat{\beta}_2 = \frac{\sum xy}{\sum \ i}$$

SUMMARY AND CONCLUSIONS:

1. The key concept underlying regression analysis is the concept of the **conditional expectation function (CEF), or population regression function (PRF).** Our objective in regression analysis is to find out how the average value of the dependent variable (or regressand) varies with the given value of the explanatory variable (or regressor).

2. This lesson largely deals with **linear PRFs,** that is, regressions that are linear in the parameters. They may or may not be linear in the regressand or the regressors.

3. For empirical purposes, it is the **stochastic PRF** that matters. The **stochastic disturbance term** *ui* plays a critical role in estimating the PRF.

4. The PRF is an idealized concept, since in practice one rarely has access to the entire population of interest. Usually, one has a sample of observations from the population. Therefore, one uses the **stochastic sample regression function (SRF)** to estimate the PRF.

II

Introduction:

To estimate the population regression function (PRF) on the basis of the sample regression function (SRF) as accurately as possible, we will discuss two generally used methods of estimation:

(1) **Ordinary least squares (OLS)** and

(2) Maximum likelihood (ML).

By and large, it is the method of OLS that is used extensively in regression analysis primarily because it is intuitively appealing and mathematically much simpler than the method of maximum likelihood. Besides, as we will show later, in the linear regression context the two methods generally give similar results.

Objectives:

1. The key objective is to find the the least-squares estimators, in the class of unbiased linear estimators, have minimum variance, that is, they are BLUE.

2. The **goodness of fit** of the fitted regression line to a set of data; that is, we shall find out how “well” the sample regression line fits the data .

Gauss-Markov Theorem/Blue:

The least-squares estimates possess some ideal or optimum properties, these properties are contained in the well-known **Gauss–Markov theorem.** To understand this theorem, we need to consider the **best linear unbiasedness property** of an estimator.

BLUE: - Best Linear-Unbiased Estimator.

MVUE: - Minimum Variance unbiased Estimator.

- If in BLUE, L is not there, because Linearity in co-effects are required not in X &Y.

The properties if Least-Square are known as the BLUE.

Properties

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

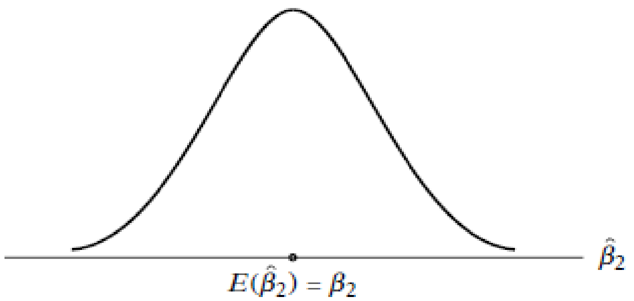
2. It is unbiased i.e its average value, $E(\hat{\beta}_2)$, is = true value of β_2.

3. Has minimum variance in class of all linear unbiased estimators.

(Note:- An unbiased estimator with the least variance is known as an efficient variable.)

Gauss Theorm:- Give the assumption of the classical linear regression Model the least squares estimators; in the class of unbiased linear estimator have minimum variance, that is they are BLUE.

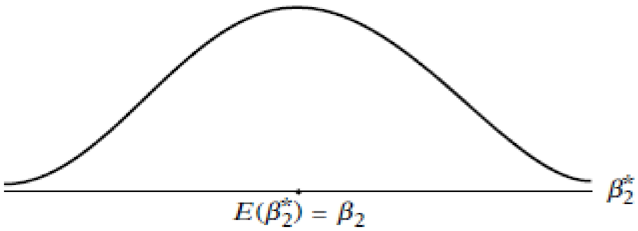
a) The mean of the $\hat{\beta}_2$values. $EC(\hat{\beta}_2)$is equal to the true value of β_2. $\hat{\beta}_2$is an unbiased estimator.



(a) Sampling distribution of β_2

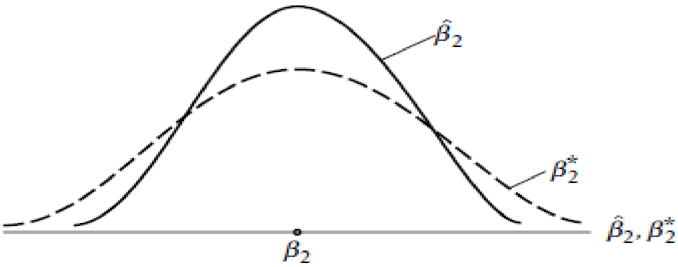
b)

- Sample distribution of β_2, an alternative estimator of β_2.
- $\hat{\beta}_2$& β_2^*. are linear estimators that is they are linear function of Y.
- β_2^* like β_2 is unbiased that is, its average or expected value is equal to β_2.



(*b*) Sampling distribution of β_2^*

c) The variance of β_2^* is larger than the variance of $\hat{\beta}_2$.One would choose the BLUE estimator



(*c*) Sampling distributions of β_2 and β_2^*

G.M. Theorem makes no assumption about the probability distribution of the random variable u_i and therefore of Y_i.

- As long as the assumption of CLRM are satisfied, the theorem holds.
- If any of the assumption doesn't hold, the theoram is invalid.

Derivation of R^2

Coefficient of determination (r^2).

A measure of "Goodness of fit"

- Goodness to fit of the fitted regression line fits the data; that is we shall find out how will the sample regression line fits the data.
- The coefficient of determination r^2 (Two variable case) or R^2 (multiple regression) is a sum many measure that tells how will the sample regression line fits the data.

$r^2 = 0$ $r^2 = 1$

% 100 of variation in Y is explanatory by X)

Y = Dependent variable

X= Explanatory variable

Greater the extent of the overlap, the greater the variance in Y is explained by X. r^2 simply a numerical measure of this overlap.

r^2 computation

$$Y_i = \hat{Y}_1 + \hat{u}_i$$

in the derivation form

$$y_i = \hat{y}_1 + \hat{u}$$

Squaring both side.

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

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum \hat{u}_i^2 + 2\sum \hat{y}_i \hat{u}_i)$$

$$\sum y_i^2 = \hat{\beta}_2^2 \sum X_i^2 + \sum \hat{u}_i^2$$

$$\sum \hat{y}_i \hat{u}_i = 0$$

$$\sum \hat{y}_i = \hat{\beta}_2 X_i$$

TSS = ESS + RSS

Where a) TSS = Total sum of squares.

i.e. $Ey^2 = \Sigma(Y_i - \bar{Y})^2$

b) ESS =Estimated sum of squares.

i.e. $E\hat{Y}_1^2 = E(\hat{Y}_i - \bar{\hat{Y}})^2 = E(\hat{y} - \bar{y}_+)^2 = \hat{\beta}_2^2 \sum X_i^2$

c) RSS= Residual sum of squares.

i.e. $E\hat{u}_1^2$

Dividing between by TSS $\frac{\text{TSS}}{\text{TSS}} = \frac{\text{ESS}}{\text{TSS}} + \frac{RSS}{TSS}$

$$1=\frac{\sum(\hat{Y}_1-\bar{Y})^2}{\sum(Y_i-\bar{Y})^2}+\frac{\sum\hat{u}_1^2}{\sum(Y_i-\bar{Y})^2}$$

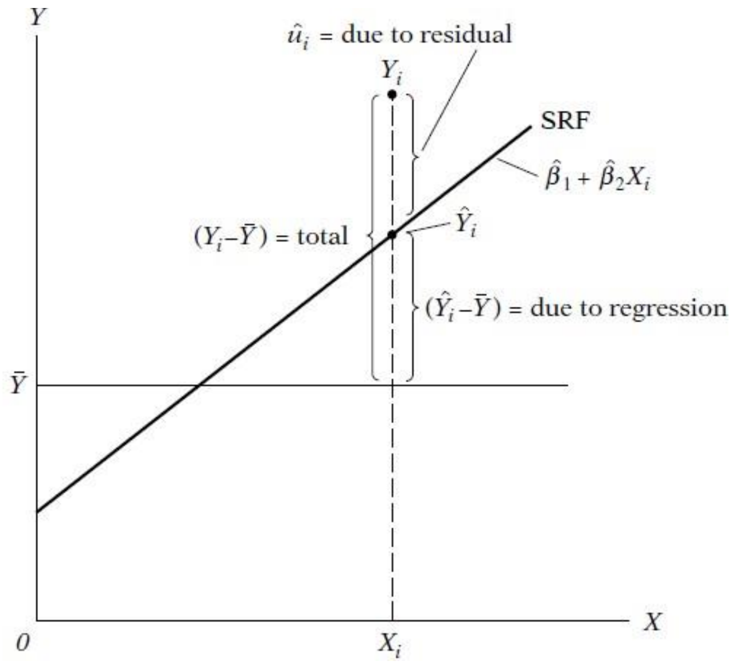
$$1=\frac{\sum(\hat{Y}_i-\bar{Y})^2}{\sum(Y_i-\bar{Y})^2}=\frac{\sum\hat{u}_i}{\sum(Y_i-\bar{Y})^2}$$

$$\left[r^2=\frac{\sum(\hat{Y}_i-\bar{Y})^2}{\sum(Y_i-\bar{Y})^2}\right]$$

$1 - r^2 = \frac{\text{RSS}}{\text{TSS}}$

$r^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$

r^2 thus defined is known as the (sample) coefficient of determination and is the most commonly used measure of goodness of fit.


Y
$\hat{u}_i$ = due to residual
Y_i
SRF
$\hat{\beta}_1 + \hat{\beta}_2 X_i$
$(Y_i - \bar{Y})$ = total
$\hat{Y}_i$
$(\hat{Y}_i - \bar{Y})$ = due to regression
$\bar{Y}$
0
X_i
X

Breakdown of the variation of Y_i into two components.

r^2 measure the proportion or % of the two variable in Y explained by regression model.

Two properties of r^2

1. It is a non negative quantity.

2. Its limits are $0 \leq r^2 \leq 1$.

An r^2 1 means a perfect fit r^2 of 0 means no relation.

A quantity closely related to but conceptually very much different from *r* 2 is the coefficient of correlation, is a measure of the degree of association between two variables. It can be computed from

$$r = \pm\sqrt{r}\ 2$$

Some of the properties of *r* are as follows:

1. It can be positive or negative, the sign depending on the sign of the term in the numerator of, which measures the sample covariation of two variables.

2. It lies between the limits of −1 and +1; that is, $-1 \leq r \leq 1$.

3. It is symmetrical in nature; that is, the coefficient of correlation between
X and *Y*(*rXY*) is the same as that between *Y* and *X*(*rYX*).

4. It is independent of the origin and scale; that is, if we define *X*i=aXi + C* and*Y*i= bYi + d*, where *a* > 0, *b* > 0, and *c* and *d* are constants, then *r* between *X** and *Y** is the same as that between the original variables*X* and *Y*.

5. If *X* and *Y* are statistically independent the correlation coefficient between them is zero; but if *r* = 0, it does not mean that two variables are independent. In other words, **zero** correlation does not necessarily imply independence.

6. It is a measure of linear association or linear dependence only; it has no meaning for describing nonlinear relations.

SUMMARY AND CONCLUSIONS:

The important topics and concepts developed in this lesson can be summarized as follows.

1. Based on these assumptions, the least-squares estimators take on certain properties summarized in the Gauss–Markov theorem, which states that in the class of linear unbiased estimators, the least-squares estimators have minimum variance. In short, they are BLUE.

2. The precision of OLS estimators is measured by their standard errors.

3. The overall goodness of fit of the regression model is measured by the coefficient of determination, r 2. It tells what proportion of the variation in the dependent variable, or regressand, is explained by the explanatory variable, or regressor. This r 2 lies between 0 and 1; the closer it is to 1, the better is the fit.

4. A concept related to the coefficient of determination is the coefficient of correlation, r. It is a measure of linear association between two variables and it lies between −1 and +1.

III

INTRODUCTION:

If our objective is to estimate β1 and β2 only, the method of OLS will be suffice. But in regression analysis our objective is not only to obtain ˆ β1 and ˆ β2 but also to draw inferences about the true β1 and β2. For example, we would like to know how close ˆ β1 and ˆ β2 are to their counterparts in the population or how close ˆYi is to the true E(Y | Xi). To that end, we must not only specify the functional form of the model, but also make certain assumptions about the manner in which Yi are generated. To see why this requirement is needed, look at the PRF: Yi = β1 + β2Xi + ui . It shows that Yi depends on both Xi and ui . Therefore, unless we are specific about how Xi and ui are created or generated, there is no way we can make any statistical inference In this lesson, we will study about the various methods through which the regression models draw inferences about the various parameters. Basically, there are three methods through which we do this:-

1. The classical linear regression model (CLRM).

2. Generalized least square (GLS).

3. Maximum Likelihood estimation (ML)

OBJECTIVES:

1. In regression analysis our objective is not only to obtain $\beta\hat{}_1$ and $\beta\hat{}_2$ but also to draw inferences about the true β_1 and β_2. For example, we would like to know how close $\beta\hat{}_1$ and $\beta\hat{}_2$ are to their counterparts in the population or how close $Y\hat{}_i$ is to the true $E(Y \mid X_i)$.

2. Look at the PRF: $Y_i = \beta_1 + \beta_2 X_i + u_i$. It shows that Y_i depends on both X_i and u_i . The

assumptions made about the X_i variable(s) and the error term are extremely critical to the valid interpretation of the regression estimates.

3. Our objective is to first discuss the assumptions in the context of the two-variable regression model,we extend them to multiple regression models, that is, models in which there is more than one regressor.

THE CLASSICAL LINEAR REGRESSION MODEL:

The assumptions underlying the method of least squares

The Gaussian, standard, or classical linear regression model (CLRM), which is the cornerstone of most econometric theory, makes 10 assumption.

Assumption 1: Linear regression model. The regression model is linear in the parameters,

$Y_i = \beta_1 + \beta_2X_i + u_i$

Assumption 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.

Assumption 3: Zero mean value of disturbance *u_i*. 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, we have

$E(ui \mid X_i)= 0$

Assumption 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 variance of u_i, are identical. Symbolically, we have

$$\begin{aligned} \text{var}\,(u_i \mid X_i) &= E(u_i \mid X_i)^2 \\ &= E(u_i^2 \mid X_i) \text{ because of Assumption 3} \\ &= \sigma^2 \end{aligned}$$

Where var stands for variance

Assumption 5: No autocorrelation between the disturbances. Given any two X values, X_i and X_j (i≠j) the correlation between any two u_i and u_j (i≠j) is zero. Symbolically

$$\mathrm{Cov}\ (u_i\ u_i\ |X_{i,}X_j) = E\{[u_i - E(u_j)]|\ X_i\}\ \{[u_i - E(u_j)]|\ X_i)$$

$$= E(u_i\ |X_i)\ (u_j\ |X_j)$$

$$= 0$$

Where i and j are two different observation and where cov means covariance.

Assumption 6: Zero covariance between u_i and X_i or $E(u_iX_i) = 0$ Formally,

$$\mathrm{Cov}\ (u_i\ X_i) = E[u_i - E(u_j)][X_i - E(x_I)]$$

$$= E[u_i\ (X_i - E(X_i))] \text{ Since } E(u_i) = 0$$

$$= E[u_iX_i) - E(X_i)\ E(u_i) \text{ Since } E(X_i) \text{ is nonstochastic}$$

$$= E[u_iX_i) \text{ Since } E(u_i) = 0$$

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

Assumption 7: The number of observation *n* must be greater than the number of parameters to be estimated. Alternatively, the number of observation n must be greater than the number of explanatory variables.

Assumption 8: Variability in X values. The X values in a given sample and not all be the same. Technically, var (X) must be a finite positive number.

Assumption 9: The regression model is correctly model in correctly specified. Alternatively, there is no specification bias or error in the model used in empirical analysis.

Assumption 10: There is no perfect multicolinearity. That is, there are no perfect linear relationships among the explanatory variable.

GENERALISED LEAST SQUARE (GLS)

OLS method doesn't follow this strategy & therefore doesn't make use of the information contained in the unequal variability of the dependent variable Y.

But GLS takes such information into accent explicitly & is therefore capable of producing estimators that are BLUE.

$$Y_i = \beta_1 + \beta_2 X_i + u \qquad \rightarrow \qquad (1)$$

Which for case of algebraic manipulation

$$Y_i = \beta_i X_{oi} + \beta_2 X_i + u_i \qquad \rightarrow \qquad (2) \qquad X_{0i} = 1$$

$$\frac{Y_i}{\sigma_i} = \beta_1 \left(\frac{X_{oi}}{\sigma_i}\right) + \beta_2 \left(\frac{X_i}{\sigma_i}\right)\left(\frac{u_i}{\sigma_i}\right) \qquad \rightarrow \qquad (3) \qquad \text{for each i}$$

$$Y_i^* = \beta_1^* X_{oi}^* + \beta_2^* X_i^* + u_i^* \qquad \rightarrow \qquad (4)$$

{Where transformed, variable are that are divided by σ_i}. We use the notation.

$\sigma^2 \rightarrow$ heteroscedastic variable

What is the purpose of transforming the original mode?

Notice the following feature of the transformed error term u_i^*

Var ($u_i^* = \sum (u_i^*)^2 = \sum \frac{(u_o)^2}{\sigma_1}$

$= \frac{1}{\sigma_1^2} \sum (u_1^2)$ {σ^2 is known)

$= \frac{1}{\sigma_1^2} (\sigma_1^2)$ $\Sigma(\mu_1^2) = \sigma_1^2$

This procedure of transforming original variable in such a way that the transformed variable satisfy the assumption of the classical model & then apply OLS to then is known as the method of GLS.

In short GLS is OLS on the transformed variables that satisfy the standard last sq. assumption.

Maximum Likelihood estimation (ML)

Assumption 2 variable modes

$Y_1 = \beta_1 + \beta_1 X_i \quad u_1$

Y_1 are normal Σ distributer

$f(Y_1 Y_2 \ldots\ldots Y_n / \beta_1 + \beta_2 X_i + \sigma^2)$

$= f(Y_1 / \beta_1 + \beta_2 X_i + \sigma^2)\, f(Y_2 / \beta_1 + \beta_2 X_i + \sigma^2) \ldots\ldots f(Y_n / \beta_1 + \beta_2)$ → (1)

When $f(Y_1) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2} \frac{(Y_i - \beta_1 - \beta_1 X_i)^2}{\sigma^2}\right\}$ → (2)

Exp mean e to the paru of expression indicator by { }

f $(Y_1) = \frac{1}{\sigma\sqrt{2\pi}}\ e^{-\frac{1}{2}\frac{(Y_i-\beta 1+\beta 2X_i)^2}{\sigma 2}}$

Subtract (2) in (1)

f$(Y_1,Y_2,Y_n / \beta_1+ \beta_2X_1 +\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}$ exp $\{-\frac{1}{2}(\frac{Y_i-\beta_1-\beta_1X_i}{\sigma^2})^2\}$ → (3)

Y_1, Y_2, Y_n are known

But β_1 β_2 & σ_2 are not.

f so (3) is known as likelihood function.

Divided by LF $(\beta_{1,}\beta_2, \sigma_2)$

$\therefore$ LF $(\beta_{1,} \beta_2 , \sigma_2) = \frac{1}{\sigma\sqrt{2\pi}}\ e\ \{-\frac{1}{2}\Sigma\ (\frac{Y_i-\beta_1-\beta_1X_i}{\sigma^2})^2\}$

ML consists in estimating the unknown parameter in such a manner that the probability of observe give by Y's is highest as possible.

In LF= -n lnσ-$\frac{n}{2}$ln$(2\pi) - \frac{1}{2}\Sigma\ \frac{(Y1-\beta_1+\beta_2X_i)^2}{\sigma^2}$ → (5)

Differencing (5) parameters with $\beta_{1,}\beta_2$ & σ_2

$\frac{\partial mLF}{\partial\beta_1} = \frac{1}{\sigma^2}\ \Sigma\ (Y_1 - \beta_1 - \beta_2 X_i)\ (-1)$ → (6)

$\frac{\partial LnLF}{\partial\beta_2} = \frac{1}{\sigma^2}\ \Sigma\ (Y_1 - \beta_1 - \beta_2 X_i)\ (-X_1)$ → (7)

$\frac{\partial Ln}{\partial\sigma_2} = -\frac{n}{\sigma} + \frac{1}{2\sigma^4}\Sigma(Y_1 - \beta_1 - \beta_2 X_i)^2$ → (8)

IV
INTRODUCTION:

The classical linear regression model is that the disturbances *ui* appearing in the population regression function are homoscedastic; that is, they all have the same variance. In this lesson we examine the validity of this assumption and find out what happens if this assumption is not fulfilled. We seek answers to the following questions:

1. What is the nature of heteroscedasticity?
2. What are its consequences?
3. How does one detect it?
4. What are the remedial measures?

OBJECTIVES:

1. Understand the meaning of heteroskedasticity and homoskedasticity through examples.

2. Understand the consequences of heteroskedasticity on OLS estimates.

3. Detect heteroskedasticity through graph inspection.

4. Detect heteroskedasticity through formal econometric tests.

5. Distinguish among the wide range of available tests for detecting heteroskedasticity.

HETEROSCEDASTICITY

Where the conditional variance of the Y population varies with X. This situation in known appropriately as heteroscedasticity or unequal spread or variance.

$$E\left(u_i^2\right)=\sigma_i^2$$

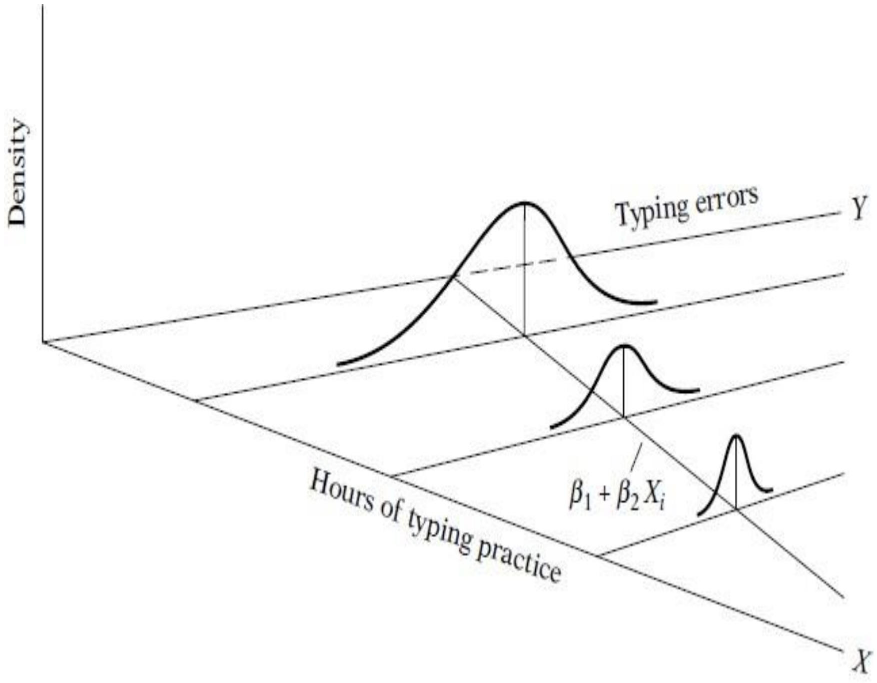


Illustration of heteroscedasticity.

Higher income families on the arrange save more than the lower income family, but there is more variability in their savings

Nature of Heteroscedasticity

2. It’s an error learning model, as people learn, their error of behavoiur become smaller over time.
3. As income grow, people have more discretionary income & hence more scope for choice about the disposition of their income.
4. As data collecting techniques increases σ_i^2 is likely to decrease.
5. If can also arise as a result of the presence of collinear.
6. It is skewness in the distribution of one or more regressions included in the model.
7. Incorrect data transformation.
8. Incorrect functional form.

1.3.1.1 OLS Estimation in the Presence of Heteroscedasticity

$$E\left(u_i^2\right)=\sigma_i^2$$

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

Applying the usual formula the OLS estimator is β_2 is

$$\beta 2 = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$$

$$= \frac{n\Sigma x_i y_i}{n\Sigma X_i} - \frac{\Sigma x_i y_i}{(\Sigma X_i)^2}$$

$$\therefore \quad Var\hat{\beta}_2 = \frac{\sigma_i^2}{\sum x_i^2}$$

$$\therefore \quad Var\beta_2^2 = \frac{\sum ne\sigma_i^2}{(\sum x_i^2)^2}$$

$$\therefore \quad Var(\hat{\beta}_2) = \frac{\sigma_i^2}{\sum x_i^2}$$

DETECTION OR TEST

1.3.2.1 Informal Methods

1. **Nature of Problem: -** Very often nature of the problem under consideration suggests whether heteroscedasticity is likely to be encountered.

2. **Graphical Problem**: - If there is no empirical information about the nature of heteroscedasticity, in practice one can do the regression analysis on the assumption that there is no heteroscedasticity & then do a postmortem

examination of the residual squared $\hat{u}_i^2$ to see if they exhibit any systematic pattern.

1.3.3.2 Formal Methods

1. **Park Test**: - Park formalized the graphical method by suggesting that σ_i^2 is same function of the explanatory variable X_i.

   His suggested function was

$$\sigma_i^2 = \sigma^2 X_i^{\beta} e^{vi}$$
$$or$$
$$\ln\sigma_i^2 = \ln\sigma^2 + \beta\ln X_i + V_i$$

   Since σ_i^2 is generally not known. Park suggested using $\hat{u}_i$, as a proxy & running following regression.

$$\ln\hat{u}_i^2 = \ln\sigma^2 + \beta\ln X_i + V_i$$
$$= \alpha + \beta\ln X_i + V_i$$

- If β turn out to be statistically significant, it would suggest that heteroscedasticity is present in the data.
- Park test is two stage procedure

  a) We run the OLS regression disregarding the heteroscedasticity question.

  b) Run the regression.

2. **Glejser Test**: - It is as Park test. He suggests regressing the absolute values of $\hat{u}_1$ on the X variable.

$$|\hat{u}_i| = \sqrt{\beta_1 + \beta_2 X_i^2} + V_i$$

3. **Spearman's Rank Correlation Test**:-

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

d_i = difference in the rank

n= no. of individual.

4. **GoldFeld Quandt Test**: - One of the popular methods, in which of one assumes that the heteroscedasticity variance σ_i^2 is positively related to one of the explanatory valuables in the regression model.

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

Suppose σ_i^2 is positively related to X_i

$\sigma_i^2 = \sigma^2 X_i^2$

5. **Breusch Pagan Godfrey Test (BPG Test)**:-

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots\ldots\ldots \beta_k X_k + u_i$$

$$\sigma_i^2 = f(\alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 \ldots\ldots\ldots + \alpha_m z_m$$

$$\sigma_i^2 = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 \ldots\ldots\ \alpha_m z_m$$

(Linear functions)

$$\alpha_0 = \alpha_1 = \alpha_2 \ldots\ldots\ldots\ldots \alpha_m = 0$$

{No heteroscedasticity, no relation between two}

Run the regression

$$\hat{u}_1^2 = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 \ldots\ldots\ldots\ldots \alpha_m z_m$$

$\theta = \frac{1}{2}$ ESS

6. **White Test**: - (Most logical for all)

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_2$$

Error may be related between X$_1$ & X$_2$.

$$\hat{u}_1^2 = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_1^2 + + \alpha_4 X_2^2 + \alpha_5 X_1 X_2 + v_i$$

of $\alpha_0 = \alpha_1 = \alpha_2 \; .... \; \alpha_n = 0$ (No heteroscedasticity)

White test can be a test of heteroscedasticity or specification error or both.

CONSEQUENCES OF USING OLS IN THE PRESENCE OF HETEROSCEDASTICITY

As we have seen, both $\hat{\beta}_2^*$ and $\hat{\beta}_2$ are (linear) unbiased estimators: In repeated sampling, on the average, $\hat{\beta}_2^*$ and $\hat{\beta}_2$ will equal the true β$_2$; that is, they are both unbiased estimators. But we know that it is $\hat{\beta}_2^*$ that is efficient that is, has the smallest variance. What happens to our confidence interval, hypotheses testing, and other procedures if we continue to use the OLS estimator $\hat{\beta}_2^*$? We distinguish two cases.

OLS Estimation allowing for heteroscedasticity

Suppose we use $\hat{\beta}_2^*$ and use the variance formula given in var ($\hat{\beta}_2$) = $\frac{\Sigma x_1^2 \sigma_1^2}{(\Sigma x_1^2)^2}$, which takes into account heteroscedasticity explicitly. Using this variance, and assuming σ_i^2 are known, can we establish confidence intervals and test hypotheses with the usual *t* and *F* test? The answer generally is no because it can be shown that var($\hat{\beta}_2^*$) < var($\hat{\beta}_2^*$

),[5] which means that confidence intervals based on the latter will be unnecessarily larger. As a result, the *t* and *F* test are likely to give us inaccurate results in that var ($\hat{\beta}_2$)is overly large and what appears to be a statistically insignificant coefficient (because the t value is smaller than what is appropriate) may in fact be significant if the correct confidence intervals were established on the basis of the GLS procedure.

OLS Estimation disregarding heteroscedasticity

The situation can become serious if we not only use $\hat{\beta}_2$ but also continue to use the usual (Homoscedasticity) variance formula given in var ($\hat{\beta}_2$)= $\frac{\sigma^2}{\Sigma x_1^2}$ even if heteroscedasticity is present or suspected: Note that this is the more likely case of the two we discuss here running in standard OLS regression package and ignoring)or being ignorant (or being ignorant of) heteroscedasticity will yield variance of $\hat{\beta}_2$. First of all car ($\hat{\beta}_2$)is a biased estimator of var ($\hat{\beta}_2$) that is, on the average it over estimates or underestimates the latter, and in general we cannot tell whether the bias is positive (overestimation) or negative (underestimation) because it depends on the nature of the relationship between $\sigma_i^2$and the values taken by the explanatory variable X,. The bias arise from the fact that $\hat{\sigma}^2$, the conventional estimator of $\hat{\sigma}^2$, namely Σ $\hat{u}_i^2$(n-2) is no longer an unbiased estimator of the latter when heteroscedasticity in present . As a result, we can no longer rely on the conventionally computed confidence intervals and the conventionally employed *t* and F tests. In short, if we persist in using the usual testing procedures despite heteroscedasticity, whatever conclusions we draw or inferences we make may be very misleading.

To throw more light on this topic, we refer to a Monte Carlo study conducted by Davidson and MacKonnon. They consider the following simple model, which in our notation is

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

They assume that $\beta_1 = 1$, $\beta_2 = 1$, and $u_i \sim N(0, X_i^{\alpha})$.

From the preceding discussion it is clear that heteroscedasticity is potentially a serious problem and the researcher needs to know whether it is present in a given situation. If its presence is detected, then one can take corrective action, such as using the weighted least-squares regression or some other technique. Before we turn to examining the various corrective procedures, however, we must first find out whether the various corrective procedures, however, we must first find out whether heteroscedasticity is present or likely to be present in a given case.

REMEDIAL MEASURES

When σ_i^2 is known: The method of weighted least squares

As we have seen, if σ_i^2 is known, the most straight forward method of correcting heteroscedasticity is by means of weighted least squares, for the estimators thus obtained are BLUE.

When σ_i^2 is not known

If true σ_i^2 are known, we can use the WLS method to obtain BLUE estimators. Since the true σ_i^2 are rarely known, is there a way of obtaining consistent (in the statistical sense) estimates of the variances and co-variances of OLS estimators even if there is heteroscedasticity? The answer is yes.

White's Heteroscedasticity-Consistent Variances and Standard Errors. White has shown that this estimate can be performed so that asymptotically valid (i.e., large-sample) statistical inferences can be made about the true parameter values. We will not present the mathematical details, for they are beyond the scope of this book. Nowadays, several computer package present White’s heteroscedasticity-corrected variances and standard errors along with the usual OLS variances and standard errors. Incidentally, White’s heteroscedasticity corrected standard errors are also known as robust standard errors.

IV

INTRODUCTION:

There are generally three types of data that are available for empirical analysis: (1) cross section, (2) time series, and (3) combination of cross section and time series, also known as pooled data. In developing the classical linear regression model (CLRM) we made several assumptions. However, we noted that *not* all these assumptions would hold in every type of data. As a matter of fact, we saw in the previous lesson that the assumption of homoscedasticity, or equal error variance, may not be always tenable in cross-sectional data. In other words, cross-sectional data are often plagued by the problem of heteroscedasticity.

However, in cross-section studies, data are often collected on the basis of a random sample of cross-sectional units, such as households (in a consumption

function analysis) or firms (in an investment study analysis) so that there is no prior reason to believe that the error term pertaining to one household or a firm is correlated with the error term of another household or firm. If by chance such a correlation is observed in cross-sectional units, it is called **spatial autocorrelation,** that is, correlation in space rather than over time. However, it is important to remember that, in cross-sectional analysis, the ordering of the data must have some logic, or economic interest, to make sense of any determination of whether (spatial) autocorrelation is present or not. The situation, however, is likely to be very different if we are dealing with time series data, for the observations in such data follow a natural ordering over time so that successive observations are likely to exhibit intercorrelations, especially if the time interval between successive observations is short, such as a day, a week, or a month rather than a year. If you observe stock price indexes, such as the Dow Jones or S&P 500 over successive days, it is not unusual to find that these indexes move up or down for several days in succession. Obviously, in situations like this, the assumption of **no auto,** or **serial, correlation** in the error terms that underlies the CLRM will be violated.

In this lesson we take a critical look at this assumption with a view to answering the following questions:

1. What is the nature of autocorrelation?

2. What are the theoretical and practical consequences of autocorrelation?

3. Since the assumption of no autocorrelation relates to the unobservable disturbances *ut*, how does one know that there is autocorrelation in any given situation? Notice that we now use the subscript *t* to emphasize that we are dealing with time series data.

4. How does one remedy the problem of autocorrelation?

OBJECTIVES:

1. Understand the meaning of autocorrelation.

2. Understand the consequences of autocorrelation on OLS estimates.

3. Detect autocorrelation through graph inspection.

4. Detect autocorrelation through formal econometric tests.

5. Distinguish among the wide range of available tests for detecting autocorrelation..

WHAT IS AUTOCORRELATION

Correlation between members of series of observation ordered in time (as in time series data) or space as in cross-sectional data)

Auto doesn't exist in the disturbance u_1)

$\Sigma(u_i\ u_j) = 0 \qquad i \neq j$

NATURE OF AUTOCORRELATION:

1. **Inertia**: - Silent feature of most of the time series is inertia or sluggishness. Well known, time series such as GNI price Index.

2. **Specification Bias: Excluded variable case**: - Residuals (which are proxies of u_i) may suggest that same variable that were originally candidates but were not included in the model for a variety of reasons should be included.

$$Y_i = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + u_i$$

Y = Quantity of beef demanded.

X_2 = Price of beef

X_3 = Consumer income

X_4 = Price of Pork

t = Time

AFTER REGRESSION:-

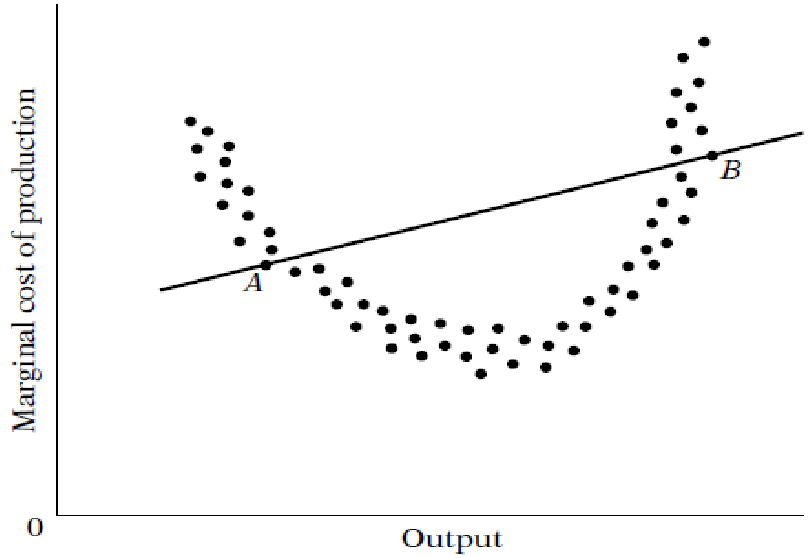
$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + V_t$$

3. **Specification Bias: Incorrect functional form**:-

Marginal $\text{Cost}_t = \beta_1 + \beta_2 \text{ output} + \beta_3 \text{output}_i^2 + u_i$

But we get the following model.

$MC_t = \alpha_1 + \alpha_2 \text{ output}_t + V_i$



MC curve corresponding to the 'true' model is along with the "incorrect" linear cost curve.

Specification bias: incorrect functional form.

4. **Cobweb Phenomenon**: - The supply of many agricultural commodities reflects the so called cobweb Phenomenon. Where supply reacts to price with a lag of one time period because supply decisions takes time implement.

$$\text{Supply}_t = \beta_1 + \beta_2\, P_{t-1} + u_t$$

5. **Lag**: -

$$\text{Consumption: - } \beta_1 + \beta_2 \text{ Income} + \beta_2 \text{ Consumption}_{t-1} + u_t$$

6. **Manipulation of data**: - In empirical analysis the raw data are often manipulated.

7. **Data Transformation**:-

$Y_t = \beta_1 + \beta_2 X_t + u_t \qquad \rightarrow 1$

Y = Consumption, X = Income

$Y_{(t-1)} = \beta_1 + \beta_2 X_{(t-1)} + u_{(t-1)}$ → 2 Previous Period

$Y_{(t-1)}$, $X_{(t-1)}$, $u_{(t-1)}$ are lagged values of X_1 Y & U

Sub. (II) from (I) we get

$\Delta Y_t = \beta_2 \, \Delta X_t + \Delta u_t$ → Δ first difference operator

FOR EMPIRICAL PURPOSE

$\Delta Y_t = \beta_2 \Delta X_t + V_t$ → $V_t = \Delta u_t = (u_t - u_{t-1})$

TEST OF AUTOCORRELATION:

Graphical Method:-

- Plot any of error
- Error term & there exists non-stationary

Stationary

$Y_t = \rho Y_{t-1} + u_t$

$Y_t = Y_{t-1} + u_t$ $(\rho=1)$

$Y_t - Y_{t-1} = u_t$

Now assume there is lag operation (L)

$(Ly_t = Y_{t-1})$

$Y_t - LY_t = U_t$

$y_t (1-L) = U_t$

if (1-L) = 0

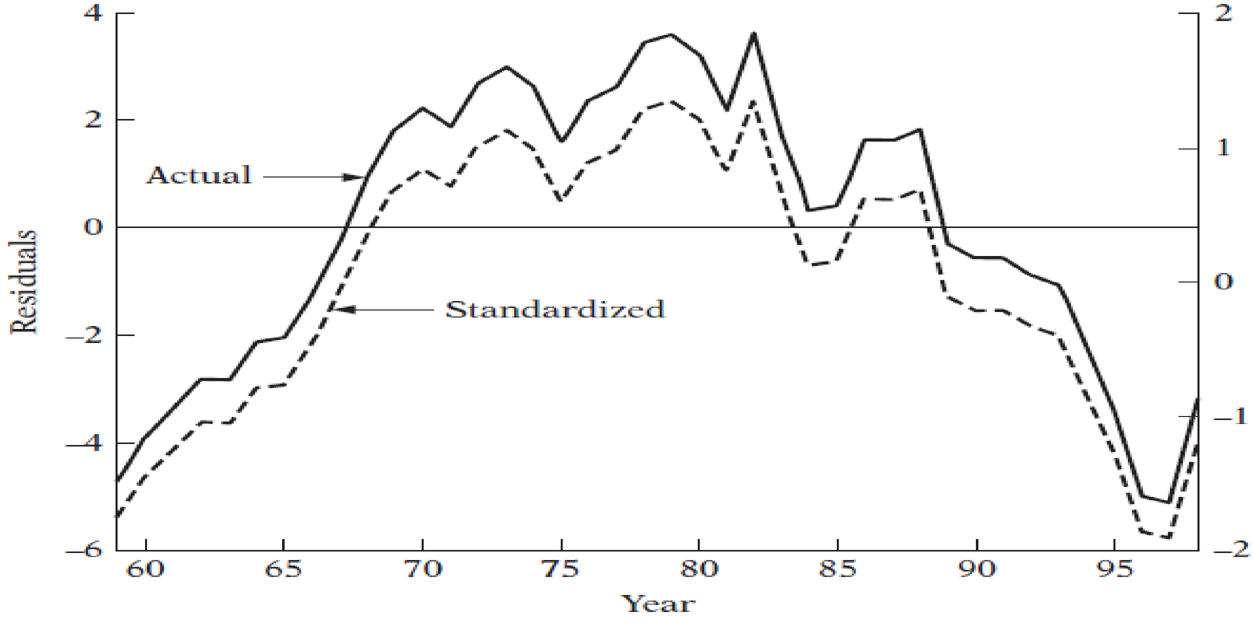
L = 1

This is known as unit root.

(When root is unit autocorrelation is there) (Non stationary & unit rest is same)

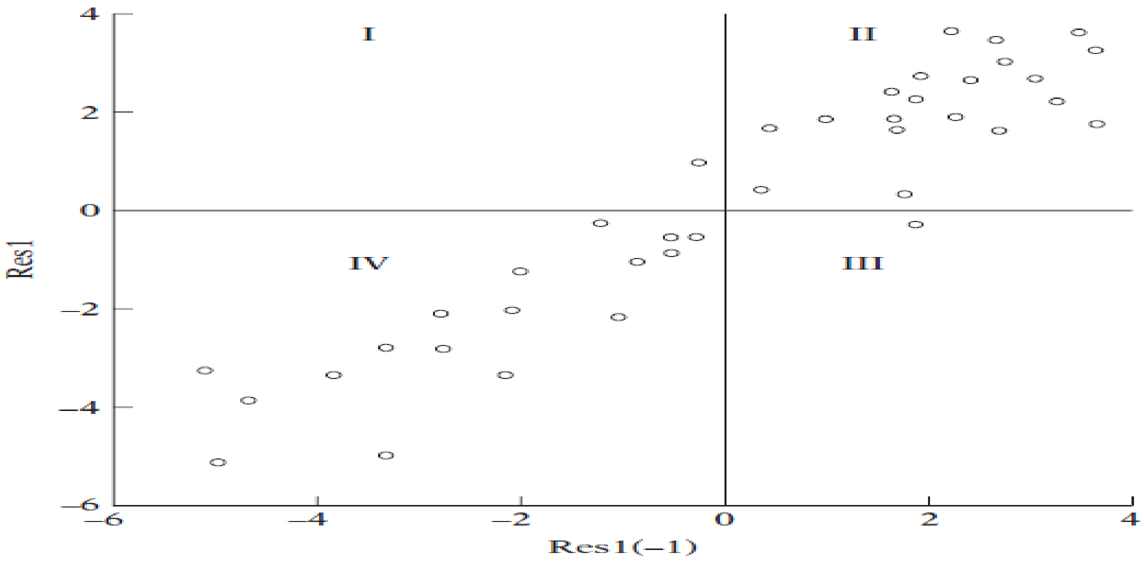
There are various ways of examine the residuals (error)

a) Time sequence plot



Residuals and standardized residuals from the wages–productivity regression (

b) Standardized residual



Current residuals versus lagged residuals.

The Runs Test:-

Initially, we have several residuals that are negative, then there is a series of positive residuals, and then there are several residuals that are negative. If these residuals were purely random, could we observe such a pattern? Intuitively, it seems unlikely. This intuition can be checked by the so-called runs test, sometimes also know as the Geary test, a nonparametric test.

(---------)(+++++++++++++++++++)(--------------)

This is also a crude method.

We now define a run as an uninterrupted sequence of one symbol or attribute, such as + or -. We further define the length of a run as the number of elements in it.

<u>Durbin Watson test</u>:-

→ Also known as Durbin Watson *d* Test.

→ One of the good methods as the *d* statistic is based on the estimated residuals, which are computed in regression analysis

$$d = \frac{\sum(\hat{u}_{t-} - \hat{u}_{t-1})^2}{\sum(\hat{u}_t)^2}$$

This tells where there exists autocorrelation or not

$$\frac{\sum \hat{u}_t^2}{\sum \hat{u}_t^2} + \frac{\sum \hat{u}_{t-1}^2}{\sum \hat{u}_t^2} - \frac{2\sum \hat{u}_t + \hat{u}_{t-1}}{\sum \hat{u}_t^2}$$

$\simeq$ 1 + 1 (by nearly) - $\frac{2\sum \hat{u}_t - \hat{u}_{t-1}}{\sum \hat{u}_t}$

$\simeq$ 2 (1 - $\frac{\sum \hat{u}_t - \hat{u}_{t-1}}{\sum \hat{u}_t}$ $)$

d $\simeq 2(1-\hat{\rho}$ 2 (1- (-1) = 4

2 (1 - (1) = 0

d will be 0 ≤ d ≤ 4

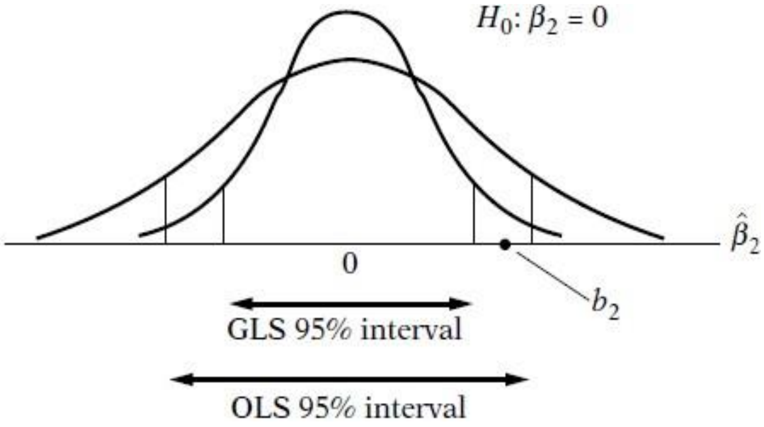
because ρ = -1 ≤ ρ ≤ 1

→ d $\simeq$ 2 → no autocorrelation

→ d $\simeq$ 0 or 4 (closer) there is autocorrelation

CONSEQUENCES OF AUTOCORRELATION:

OLS Estimation allowing for Autocorrelation.



GLS and OLS 95% confidence intervals.

To establish confidence interval to test hypotheses, one should be GLS & not OLS even though the estimators derived from the latter are unbiased & consistent.

Estimation Disregarding Autocorrelation.

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{(n-2)}$$

Unbiased estimate of σ² i.e $\sum(\hat{\sigma}_i^2) = \sigma^2$

$$\sum(\hat{\sigma}^2 - \frac{\sigma^2\{n-[2/(1-\rho)]\}-2\rho r}{n-2}$$

REMEDIAL MEASURES OF AUTOCORRELATION:

1. Try to find out if the autocorrelation is pure autocorrelation or not because of the result of the mis-specification of the model.
2. Transformation of original model, so that in the transformed model we do not have the problem of (Pure) autocorrelation.
3. In case of large sample we can Newey-West method to obtain standard error of OLS estimators that are corrected for auto correlation.
4. In some situation we can continue to use the OLS method.

SUMMARY AND CONCLUSIONS:

1. If the assumption of the classical linear regression model—that the errors or disturbances *ut* entering into the population regression function (PRF) are random or uncorrelated—is violated, the problem of serial or autocorrelation arises.

2. Autocorrelation can arise for several reasons, such as inertia or sluggishness of economic time series, specification bias resulting from excluding important variables from the model or using incorrect functional form, the cobweb phenomenon, data massaging, and data transformation.

3. Although in the presence of autocorrelation the OLS estimators remain unbiased, consistent, and asymptotically normally distributed, they are no longer efficient. As a consequence, the usual *t*, *F*, and $\chi 2$ tests cannot be legitimately applied. Hence, remedial results may be called for.

4. The remedy depends on the nature of the interdependence among the disturbances *ut*. But since the disturbances are unobservable, the common practice is to assume that they are generated by some mechanism.

LETS SUM IT UP:

In last, we can say that this lesson in many ways similar to the preceding lesson on heteroscedasticity in that under both heteroscedasticity and autocorrelation the usual OLS estimators, although linear, unbiased, and asymptotically (i.e., in large samples) normally distributed, are no longer minimum variance among all linear unbiased estimators. In short, they are not efficient relative to other linear and unbiased estimators. Put differently, they may not be BLUE. As a result, the usual, t, F, and χ2 may not be valid.

EXCERCISES:

State whether the following statements are true or false. Briefly justify your answer.

a. When autocorrelation is present, OLS estimators are biased as well as inefficient.
b. The Durbin–Watson d test assumes that the variance of the error term ut is homoscedastic.
c. The first-difference transformation to eliminate autocorrelation assumes that the coefficient of autocorrelation ρ is −1.
d. The R2 values of two models, one involving regression in the first difference form and another in the level form, are not directly comparable.
e. A significant Durbin–Watson d does not necessarily mean there is autocorrelation of the first order.
f. In the presence of autocorrelation, the conventionally computed variance and standard errors of forecast values are inefficient.
g. The exclusion of an important variable(s) from a regression model may give a significant d value.

Given a sample of 50 observations and 4 explanatory variables, what can you say about autocorrelation if (a) d = 1.05? (b) d = 1.40? (c) d = 2.50?
(d) d = 3.97?

In a sequence of 17 residuals, 11 positive and 6 negative, the number of runs was 3. Is there evidence of autocorrelation? Would the answer change if there were 14 runs?

Explain the Durbin-Watson and Runs Test for detecting autocorrelation?

Elaborate the various remedial measures of autocorrelation?

2.7 Suggested Reading / References:

1. Baltagi, B.H.(1998). Econometrics, Springer, New York.

2. Chow,G.C.(1983). Econometrics, McGraw Hill, New York.

3. Goldberger, A.S.(1998). Introductory Econometrics, Harvard University Press, Cambridge, Mass.

4. Green, W.(2000). Econometrics, Prentice Hall of India, New Delhi.

5. Gujarati, D.N.(1995). Basic Econometrics. McGraw Hill, New Delhi.

6. Koutsoyiannis,A.(1977). Theory of Econometrics(2nd Esdn.). The Macmillan Press Ltd. London.

7. Maddala, G.S.(1997). Econometrics, McGraw Hill; New York.

LESSON-3

MULTICOLLINEARITY

STRUCTURE

INTRODUCTION:

The assumption 10 of the classical linear regression model (CLRM) is that there is no multicollinearity among the regressors included in the regression model. In this lesson we take a critical look at this assumption by seeking answers to the following questions:

1. What is the nature of multicollinearity?

2. Is multicollinearity really a problem?

3. What are its practical consequences?

4. How does one detect it?

5. What remedial measures can be taken to alleviate the problem of multicollinearity?

OBJECTIVES:

1. Understand the meaning of multicollinearity.

2. Understand the consequences of multicollinearity on OLS estimates.

3. Detect multicollinearity. through rule of thumb inspection.

4. Detect multicollinearity. through formal econometric tests.

5. Distinguish among the wide range of available tests for detecting multicollinearity..

MULTICOLLINEARITY

It means the existence of a perfect or exact linear relationship among some all explanatory variables of a regression model.

$X_{2i} = \lambda X_{3i}$ → perfect multicollinearity

It is due to Ragnar Frisch

$\lambda_1 X_1 + \lambda_2 X_2 + \ldots\ldots\ldots\ldots\ldots + \lambda_k X_k = 0$

$\lambda_1\ \lambda_2 \ldots\ldots\ldots\ldots \lambda_k$ are constants

The term multicollinearity is used in a broader sense to include the case of perfect multicollionearity.

$\lambda_1 X_1 + \lambda_2 X_2 + \ldots\ldots\ldots\ldots\ldots + \lambda_k X_k + V_i = 0$

Where V_i is a stochastic error term.

Difference between perfect & less than perfect multicollinearly assumed.

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \cdots - \frac{\lambda_k}{\lambda_2} X_{ki}$$

Linear combination $\lambda_2 \neq 0$

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \cdots - \frac{\lambda_k}{\lambda_2} X_{ki} - \frac{1}{\lambda_2} v_i$$

which shows that X_2 is not an exact linear. Combination of other X's because it is also determined by the stochastic error term V_i.

3.3.1 NATURE/ SOURCE:

If multicollinearity is perfect in the sense, the regression coefficients of the *X* variables are indeterminate and their standard errors are infinite. If multicollinearity is less than perfect, the regression coefficients, although determinate, possess large standard errors (in relation to the coefficients themselves), which means the coefficients cannot be estimated with great precision or accuracy. The following can be reasons for the existence of multicollinearity:

1. Data collection method
2. Constraints on the model.
3. Model specification.
4. An over determined model.

3.3.2 REMEDIAL MEASURES

Do Nothing

The "do nothing" school of thought is expressed by Blanchard as follows:

When students run their first ordinary least squares (OLS) regression, the first problem that they usually encounter is that of multicollinearity. Many of them conclude that there is something wrong with OLS; some resort to new and often creative techniques to or around the problem. But we tell them, this is wrong, Multicollineaity is God's will, not a problem with OLS or statistical technique in general.

What Blanchard is saying is that multicollinearity is essentially a data deficiency problem (micronumerosity, again) and some times we have no choice over the data we have available for empirical analysis.

Rule of Thumb Procedures

One can try the following rules of thumb to ad

dress the problem of multicollinearity, the success depending on the severity of the multicollinearity problem.

1. **A priori information.** Suppose we consider the model

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

where Y = consumption, X_2 = income, and X_3 = wealth. As noted before, income and wealth variables tend to be highly collinear. But suppose a priori we believe that $\beta_3 = 0.10\beta_2$; that is, the rate of change of consumption with respect to wealth is one-tenth the corresponding rate with respect to income. We can then run the following regression:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \mathbf{0.10}\beta_2 X_{3i} + u$$
$$= \beta_i + \beta_2 X_i + u_i$$

Where $X_i + 0.1X_{3i}$. Once we obtain $\hat{\beta}_2$, we can estimate $\hat{\beta}_3$ from the postulated relationship between $\beta_2$and β_3.

2. **Combining cross-sectional and time series data.** A variant of the extraneous or a priori information technique is the combination of cross-sectional and time-series data, known as pooling the data. Suppose we want to study the demand for automobiles in the United States and assume we have time series data on the number of cars sold, average price of the car,

$$InY_1 = \beta_1 + \beta_2 InP_1 + \beta_3 InI_t + u_t$$

Where Y = number of cars sold, P = average price, I = income, and t = time. Out objective is to estimate the price elasticity β_2 and income elasticity β_3.

In time series data the price and income variables generally tend to be highly collinear. Therefore, if we run the proceeding regression, we shall be faced with the usual multicollinearity problem. A way out of this has been suggested by Tobin. He says that if we have cross-sectional data (for example, data generated by consumer panels, or budget studies conducted by various private and governmental agencies), we can obtain a fairly reliable estimate of the income elasticity β_3 because in such data, which are at a point in time, the prices do not vary much. Let the cross-sectionally estimated income elasticity be $\hat{\beta}_3$. Using this estimate, we may write the preceding times series regression as

$$Y_t^* = \beta_1 + \beta_2 InP_t + u_t$$

Where Y* = In Y - $\hat{\beta}_3$ In I, that is, Y* represents that value of Y after removing from it the effect of income. We can now obtain an estimate of the price elasticity β_2 from the preceding regression.

3) Dropping a variable (s) and specification bias. When faced with severe multicollinearity, one of the “simplest” things to do is to drop one of the collinear variables. Thus, in our consumption-income-wealth illustration, which shows that, whereas in the original model the income variable was statistically insignificant, it is now ‘highly’ significant.

But in dropping a variable from the model we may be committing specification bias or specification error. Specification bias arises from incorrect specification of the model used in the analysis. Thus, if economic theory says that income and wealth should both be included in the model explaining the consumption expenditure, dropping the wealth variable would constitute specification bias.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

But we mistakenly fit the model

$$Y_i = b_1 + b_{12} X_{2i} + \hat{u}_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Then it can be shown that

$$E(b_{12}) = \beta_2 + \beta_3 b_{32} \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

where b_{32} = slope coefficient in the regression of X_3 on X_2. Therefore, it is obvious that b_{12} will be a biased estimate of β_2 as long as b_{32} is different from zero (it is assumed that $\beta_3$is different from zero; otherwise there is no sense in including X_3 in the original model). Of course, if b_{32} is zero, we have no multicollinearity problem to begin with. It is also clear from that if both b_{32} and β_3 are positive (or both are negative), $E(b_{12})$ will be greater than β_2; hence, on the average b_{12} will overestimate β_2, leading to a positive bias. Similarly, if the product b_{32} β_3 is negative, on the average b_{12} will underestimate β_2, leading to a negative bias.

4) **Transformation of variables.** Suppose we have time series data on consumption expenditure, income and wealth. One reason for high multicollinearity between income and wealth in such data is that over time both the variables tend to move in the same direction. One way of minimizing this dependence is to proceed as follows.

If the relation

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Holds at time t, it must also hold at time t – 1 because the origin of time is arbitrary anyway. Therefore, we have

$$Y_{t-1} = \beta_1 + \beta_2 X_{2,t-1} + \beta_3 X_{3,t-1} + u_{t-1} \text{\_\_\_\_\_\_\_\_\_\_} (4)$$

If we subtract (3) from (1), we obtain

$$Y_t - Y_{t-1} = \beta_2 (X_{2t} - X_{2,t-1}) + \beta_3 (X_{3t} - X_{3,t-1}) + v_t \text{ ....} (5)$$

Where $v_t = u_t - u_{t-1}$. Equation (5) is known as the first difference form because we run the regression, not on the original variables, but on the differences of successive values of the variables.

The first difference regression model often reduces the severity of multicollinearity because, although the levels of X_2 and X_3 may be highly correlated, there is no a priori reason to believe that their differences will also be highly correlated.

As we shall see in the lessons on time series econometrics, an incidental advantage of the first – difference transformation is that it may make a nonstationary time series stationary. In those lessons we will see the importance of stationary time series Another commonly used transformation in practice is the ratio transformation.

Consider the model:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \text{ ....} (6)$$

Where Y is consumption expenditure in real dollars, X_2 is GDP, and X_3 is total population. Since GDP and population grow over time, they are likely to be correlated. One “Solution” to this problem is to express the model on a per capita basis, that is, by dividing (6) by X_3, to obtain:

$$\frac{Y_t}{X_{3t}} = \beta_1\left(\frac{1}{X_{3t}}\right) + \beta_2\left(\frac{X_{2t}}{X_{3t}}\right) + \beta_3 + \left(\frac{u_t}{X_{3t}}\right) \ldots\ldots\ldots\ldots\ldots(7)$$

Such a transformation may reduce collinearity in the original variables.

But the first – difference or ratio transformations are not without problems. For instance, the error term v_t in ( ) may not satisfy one of the assumptions of the classical linear regression model, namely, that the disturbances are serially uncorrelated.

5) **Additional or new data.** Since multicollinearity is a sample feature, it is possible that in another sample involving the same variables collinearity may be so serious as in the first sample. Sometimes simply increasing the size of the sample (if possible) may attenuate the collinearity problem. For example, in the three-variable model we saw that

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

Now as the sample size increases, $\sum x_{2i}^2$ will generally increase. Therefore, for any given r_{23}, the variance of $\hat{\beta}_2$ will decrease, thus decreasing the standard error, which will enable us to estimate β_2 more precisely.

6) **Other methods of remedying multicollinearity.** Multivariate statistical techniques such as factor analysis and principal components or techniques such as ridge regression are often employed to 'solve' the problem of multicollinearity. Unfortunately, these techniques are beyond the scope of this book, for they cannot be discussed competently without resorting to matrix algebra.

SUMMARY AND CONCLUSIONS:

1. One of the assumptions of the classical linear regression model is that there is no multicollinearity among the explanatory variables, the X’s. Broadly interpreted, multicollinearity refers to the situation where there is either an exact or approximately exact linear relationship among the X variables.

2. The consequences of multicollinearity are as follows: If there is perfect collinearity among the X’s, their regression coefficients are indeterminate and their standard errors are not defined. If collinearity is high but not perfect, estimation of regression coefficients is possible but their standard errors tend to be large. As a result, the population values of the coefficients cannot be estimated precisely. However, if the objective is to estimate linear combinations of these coefficients, the estimable functions, this can be done even in the presence of perfect multicollinearity

3. Although there are no sure methods of detecting collinearity, there are several indicators of it, which are as follows:

(a) The clearest sign of multicollinearity is when R2 is very high but none of the regression coefficients is statistically significant on the basis of the conventional t test. This case is, of course, extreme.

(b) In models involving just two explanatory variables, a fairly good idea of collinearity can be obtained by examining the zero-order, or simple, correlation coefficient between the two variables. If this correlation is high, multicollinearity is generally the culprit.

(c) However, the zero-order correlation coefficients can be misleading in models involving more than two X variables since it is possible to have low zero-order correlations and yet find high multicollinearity. In situations like these, one may need to examine the partial correlation coefficients.

(d) If R2 is high but the partial correlations are low, multicollinearity is a possibility. Here one or more variables may be superfluous. But if R2 is high and the partial correlations are also high, multicollinearity may not be readily detectable. Also, as pointed out by C. Robert, Krishna Kumar, John O'Hagan, and Brendan McCabe, there are some statistical problems with the partial correlation test suggested by Farrar and Glauber.

(e) Therefore, one may regress each of the Xi variables on the remaining X variables in the model and find out the corresponding coefficients of determination R2 . A high R2 would suggest that Xi is highly correlated with the rest of the X's. Thus, one may drop that Xi from the model, provided it does not lead to serious specification bias.

4. Detection of multicollinearity is half the battle. The other half is concerned with how to get rid of the problem. Again there are no sure methods, only a few rules of thumb. Some of these rules are as follows: (1) using extraneous or prior information, (2) combining cross-sectional and time series data, (3) omitting a highly collinear variable, (4) transforming data, and (5) obtaining additional or new data. Of course, which of these rules will work in practice will depend on the nature of the data and severity of the collinearity problem.

5. We noted the role of multicollinearity in prediction and pointed out that unless the collinearity structure continues in the future sample it is hazardous to use the estimated regression that has been plagued by multicollinearity for the purpose of forecasting.

LETS SUM IT UP:

In the concluding remarks, we can say that in cases of near or high multicollinearity, one is likely to encounter the following consequences:

1. Although BLUE, the OLS estimators have large variances and covariances, making precise estimation difficult.

2. Because of consequence 1, the confidence intervals tend to be much wider, leading to the acceptance of the “zero null hypothesis” (i.e., the true population coefficient is zero) more readily.

3. Also because of consequence 1, the *t* ratio of one or more coefficients tends to be statistically insignificant.

4. Although the *t* ratio of one or more coefficients is statistically insignificant, *R*2, the overall measure of goodness of fit, can be very high.

5. The OLS estimators and their standard errors can be sensitive to
small changes in the data.

<u>EXCERCISES:</u>

What do you mean by multicollinearity?

What is Rule of Thumb?

How can we detect multicollinearity?

Q.4.State with reason whether the following statements are true, false, or uncertain:

a. Despite perfect multicollinearity, OLS estimators are BLUE.

b. In cases of high multicollinearity, it is not possible to assess the individual significance of one or more partial regression coefficients.

c. If an auxiliary regression shows that a particular R2 is high, there is definite evidence of high collinearity.

d. High pair-wise correlations do not suggest that there is high multicollinearity.

e. Multicollinearity is harmless if the objective of the analysis is prediction only.

f. Ceteris paribus, the higher the VIF is, the larger the variances of OLS estimators.

g. The tolerance (TOL) is a better measure of multicollinearity than the VIF.

h. You will not obtain a high R2 value in a multiple regression if all the partial slope coefficients are individually statistically insignificant on the basis of the usual t test.

i. In the regression of Y on X2 and X3, suppose there is little variability in the values of X3. This would increase var (ˆ β3). In the extreme, if all
X3 are identical, var (ˆ β3) is infinite.

Q.5 a. Show that if r1i = 0 for i = 2, 3, . . . , k then R1.2 3. . . k = 0

b. What is the importance of this finding for the regression of variable X1(=Y) on X2, X3, . . . , Xk?

3.7 Suggested Reading / References:

1. Baltagi, B.H.(1998). Econometrics, Springer, New York.

2. Chow,G.C.(1983). Econometrics, McGraw Hill, New York.

3. Goldberger, A.S.(1998). Introductory Econometrics, Harvard University Press, Cambridge, Mass.

4. Green, W.(2000). Econometrics, Prentice Hall of India, New Delhi.

5. Gujarati, D.N.(1995). Basic Econometrics. McGraw Hill, New Delhi.

6. Koutsoyiannis,A.(1977). Theory of Econometrics(2nd Esdn.). The Macmillan Press Ltd. London.

7. Maddala, G.S.(1997). Econometrics, McGraw Hill; New York.

LESSON-4

MODEL MIS-SPECIFICATION VERSUS PURE AUTOCORRELATION

STRUCTURE

INTRODUCTION

OBJECTIVES

CORRECTING FOR (PURE) AUTOCORRELATION

THE METHOD OF GENERALIZED LEAST SQUARES (GLS)

OLS VERSUS FGLS AND HAC

COEXISTENCE OF AUTOCORRELATION AND HETEROSCEDASTICITY

SUMMARY AND CONCLUSIONS

LETS SUM IT UP

EXCERCISES

SUGGESTED READING / REFERENCES

INTRODUCTION:

Let us return to our wages productivity regression. There we saw that the d value was 0.1229 and based on the Durbin-Watson d test we concluded that there was positive correlation in the error term. Could this correlation have arisen because our model was not correctly specified? Since the data underlying regression is time series data, it is quite possible that both wages and productivity exhibit trends. If that is the case, then we need to include the time or trend, t, variable in the model to see the relationship between wages and productivity net of the trends in the two variables.

To test this, we included the trend variable and obtained the following results.

$$\hat{Y}_t = 1.4752 + 1.3057 X_t - 0.9032 t$$

$$se = (13.18) \quad (0.2765) \quad (0.4203)$$

$$t = (0.1119) (4.7230) \; (-2.1490)$$

$$R^2 = 0.9631; \qquad d = 0.204$$

The interpretation of this model is straightforward: Over time, the index of real wages has been decreasing by about 0.90 units per year. After allowing for this if the productivity index went up by a unit, on average, the real wage index went up by about 1.30 units, although this number is not statistically different from one (why?). What is interesting to note is that even allowing for the trend variable, the d value is still very low, suggesting pure autocorrelation and not necessarily specification error.

To test this, we regress Y on X and X^2 to test for the possibility that the real wage index may be nonlinearly related to the productivity index. The results of this regression are as follows:

$$\hat{Y}_t = -16.2181 + 1.9488 X_t - 0.0079 X_t^2$$

$$t = (-5.4891) (24.9868) (-15.9363)$$

$$R^2 = 0.9947 \qquad d = 1.0$$

These results are interesting. All the coefficients are statistically highly significant, the p values being extremely small. From the negative quadratic term, it seems that although the real wage index increases as the productivity index increases, it increases at a decreasing rate. But look at the d value. It still suggests positive autocorrelation in the residuals, for d_L = 1.391 and d_U = 1.60 and the estimated d value lies below d_L.

It may be safe to conclude from the proceeding analysis that our wages-productivity regression probably suffers from pure autocorrelation and not necessarily from specification bias. Knowing the consequences of autocorrelation, we may therefore want to take some corrective action. We will do so shortly.

Incidentally, for all the wages productivity regression that we have presented above, we applied the Jarque–Bera test of normality and found that the residuals were normally distributed, which is comforting because the d terms assumes normally of the error term.

OBJECTIVES:

1. The key objective is to find what are the criteria in choosing a model for empirical analysis.

2. Our objective is to find what types of model mis- specification errors is one likely to encounter in practice.

3. The another objective is to find how does one evaluate the performance of competing models?

CORRECTING FOR (PURE) AUTOCORRELATION:

THE METHOD OF GENERALIZED LEAST SQUARES (GLS):

Knowing the consequences of autocorrelation, especially the lack of efficiency of OLS estimators, we may need to remedy the problem. The remedy depends on the knowledge one has about the nature of interdependence among the disturbances, that is, knowledge about the structure of autocorrelation.

As a starter, consider the two-variable regression model:

$$Y_t = \beta_1 + \beta_2 X_t + \mathbf{u}_t$$

And assume that the error term follows the AR(1) scheme, namely,

$$(u_t - pu_{t-1}) = \varepsilon_t \quad -1 < p < 1$$

Now we consider two cases: (1) p is known and (2) is not known but has to be estimated.

When is known

If the coefficient of first-order autocorrelation is known, the problem of autocorrelation can be easily solved. Hence,

$$Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + u_{t-1} \qquad 1$$

Multiplying by on both sides, we obtain

$$\rho Y_{t-1} = \rho\beta_1 + \rho\beta_2 X_{t-1} + pu_{t-1} \qquad 2$$

Subtracting (2) from (1) gives

$$(Y_t - \rho Y_{t-1}) = \beta_1(1-\rho) + \beta_2(X_t - \rho X_{t-1}) + \varepsilon_t \qquad 3$$

Where $\varepsilon_t = (u_t - pu_{t-1})$

We can express (3) as

$$Y_t^* = \beta_1^* + \beta_2^* X_t^* + \varepsilon_t \qquad 4$$

Where $\beta_1^* = \beta_1(1-\rho), Y_t^* = (Y_t - \rho Y_{t-1}), X_t^* = (X_t - pX_{t-1}), and \beta_2^* = \beta_2 \qquad 5$

Since the error term in (4) satisfies the usual OLS assumptions, we can apply OLS to the transformed variables Y* and X* and obtain estimators with all the optimum properties, namely, BLUE. In effect, running is tantamount to using generalized least squares (GLS) discussed in the previous lesson – recall that GLS is nothing but OLS applied to the transformed model that satisfies the classical assumptions.

Regression (4) is known as the generalized, or quasi, difference equation. It involves regressing Y on X, not in the original form, but in the difference form, which is obtained by subtracting a proportion $(=\rho)$ of the value of a variable in the previous time period from its value in the current time period. In this differencing procedure we lose one observation because the first observation has no antecedent. To avoid this loss of one observation, the first observation on Y and X is transformed as follows. $Y_1\sqrt{1-p^2}\, and X_1\sqrt{1-p^2}.$ This transformation is known as the Prais-Winsten transformation.

OLS VERSUS FGLS AND HAC

The practical problem facing the researcher is this: In the presence of auto-correlation, OLS estimators, although unbiased, consistent, and asymptotically normally distributed, are not efficient. Therefore, the usual inference procedure based on the t, F, and χ^2 tests is no longer appropriate. On the other hand, FGLS (Feasible GLS and EGLS: Estimated GLS) HAC (Heteroscedasticity and autocorrelation estimation) produce estimators that are efficient, but the finite, or small-sample, properties of these estimators are not well documented. This means in small samples the FGLS and HAC might actually do worse than OLS. As a matter of fact, in a Monte Carlo study Griliches and Rao found that if the sample is relatively small and the coefficient of auto-correlation, , is less than 0.3, OLS is as good or better than FGLS. As a practical matter, then, one may use OLS in small samples in which the estimated rho is, say, less than 0.3. Of course, what is a large and what is a small sample are relative questions, and one has to use some practical judgement. If you have only 15 to 20 observations, the sample may be small, but if you have, say, 50 or more observations, the sample may be reasonably large.

Coexistence of Autocorrelation and Heteroscedasticity

What happens if a regression model suffers from both heteroscedasticity and autocorrelation? Can we solve the problem sequentially, that is, take care of heteroscedasticity first and then autocorrelation? As a matter of fact, one author contends that “Autoregression can only be detected after the heteroscedasticity is controlled for”. But can we develop an omnipotent test that can solve these and other problems (e.g., model specification) simultaneously? Yes, such tests exist, but their discussion will take us far afield. It is better to leave them for references.

SUMMARY AND CONCLUSIONS

1. If the assumption of the classical linear regression model that the errors or disturbances u_t entering into the population regression function (PRF) are random or uncorrelated – is violated, the problem of serial or autocorrelation arises.

2. Autocorrelation can arise for several reasons, such as inertia or sluggishness of economic time series, specification bias resulting from excluding important variables from the model or using incorrect functional form, the cobweb phenomenon, data massaging, and data transformation. As a result, it is useful to distinguish between pure autocorrelation and "induced" autocorrelation because of one or more factors just discussed.

3. Although in the presence of autocorrelation the OLS estimators remains unbiased, consistent, and asymptotically normally distributed, they are no longer efficient. As a consequence, the usual t, F, and χ^2 tests cannot be legitimately applied. Hence, remedial results may be called for.

4. The remedy depends on the nature of the interdependence among the disturbances u_t. But since the disturbances are unobservable, the common practice is to assume that they are generated by some mechanism.

5. The mechanism that is commonly assumed is the Markov first-order autoregressive scheme, which assumes that the disturbance in the current time period is linearly related to the disturbance term in the previous time period, the coefficient of autocorrelation p providing the extent of the interdependence. This mechanism is known as the AR(1) scheme.

6. If the AR(1) scheme is valid and the coefficient of autocorrelation is known, the serial correlation problem can be easily attacked by transforming the data following the generalized difference procedure. The AR(1) Scheme can be easily generalized to an AR(p). One can also assume a moving average (MA) mechanism or a mixture of AR and MA schemes, known as ARMA. This topic will be discussed in the lessons on time series econometrics.

7. Even if we use an AR(1) scheme, the coefficient of autocorrelation is not known a priori. We considered several methods of estimating p, such as the Durbin-Watson d, Theil-Nagar modified d, Cochrane-Orcutt (C-O) iterative procedure, C-O two step method, and the Durbin two-step procedure. In large samples, these methods generally yield similar estimates of p, although in small samples they perform differently. In practice, the C-O interative method has become quite popular.

8. Using any of the methods just discussed, we can use the generalized difference method to estimate the parameters of the transformed model by OLS, which essentially amounts to GLS. But since we estimate $\rho(=\hat{p})$ we call the method of estimation as feasible, or estimated, GLS, or FGLS or EGLS for short.

1. In using EGLS, one has to be careful in dropping the first observation, for in small samples the inclusion or exclusion of the first observation can make a dramatic difference in the results. Therefore, in small samples it is advisable to transform the first observation according to the Prais-Winsten procedure. In large samples, however, it makes little difference if the first observation is included or not.

10. It is very important to note that the method of EGLS has the usual optimum statistical properties only in large samples. In small samples, OLS may actually do better that EGLS, especially if $p<0.3$.

11. Instead of using EGLS, we can still use OLS but the correct the standard errors for autocorrelation by the Newey-West HAC procedure. Strictly speaking, this procedure is valid in large samples. One advantages of the HAC procedure is that it not only corrects for autocorrelation but also for heteroscedasticity, if it is present.

12. Of course, before remediation comes detection of autocorrelation. There are formal and informal methods of detection. Among the informal methods, once can simply plot the actual or standardized residuals, or plot current residuals against past residuals. Among formal methods, one can use the runs test, Durbin Watson *d* test, asymptotic normality test, Berenblutt-Webb test, and Breusch-Godfrey (BG) test. Of these, the most popular and routinely used is the Durbin-Watson *d* test, for it is much more general in that it allows for both AR and MA error structures as well as the presence of lagged regressed as an explanatory variable. But keep in mind that it is a large sample test.

LETS SUM IT UP:

In concluding remarks, we can say that if particular model is not specified correctly, we face the problem of model specification error or model specification bias.

EXCERCISES:

Q1 State Breusch Pagan Godfrey test.

Q2 What happens to OLS estimation in presence of autocorrelation?

Q3 What is EGLS or FGLS?

Q4 Does heteroscedasticity makes the estimators biased? Explain.

Q5 Describe correlation for pure autocorrelation.

Q6 Describe multicollinearity its test and remedial measures.

Suggested Reading / References:

1. Baltagi, B.H.(1998). Econometrics, Springer, New York.

2. Chow,G.C.(1983). Econometrics, McGraw Hill, New York.

3. Goldberger, A.S.(1998). Introductory Econometrics, Harvard University Press, Cambridge, Mass.

4. Green, W.(2000). Econometrics, Prentice Hall of India, New Delhi.

5. Gujarati, D.N.(1995). Basic Econometrics. McGraw Hill, New Delhi.

6. Koutsoyiannis,A.(1977). Theory of Econometrics(2nd Esdn.). The Macmillan Press Ltd. London.

7. Maddala, G.S.(1997). Econometrics, McGraw Hill; New York.

UNIT-3

ECONOMETRIC MODELING: MODEL SPECIFICATION AND DIAGNOSTIC TESTING

- .

LESSON-1

MODEL SPECIFICATION

STRUCTURE

INTRODUCTION

OBJECTIVES

MODEL SELECTION CRITERIA

TYPES OF SPECIFICATION ERRORS

CONSEQUENCES OF MODEL SPECIFICATION ERRORS

UNDERFITTING A MODEL (OMITTING A RELEVANT VARIABLE)

INCLUSION OF AN IRRELEVANT VARIABLE (OVERFITTING A MODEL)

TESTS OF SPECIFICATION ERRORS

DETECTING THE PRESENCE OF UNNECESSARY VARIABLES (OVER FITTING A MODEL)

TESTS FOR OMITTED VARIABLES AND INCORRECT FUNCTIONAL FORM

THE DURBIN-WATSON *D* STATISTICS ONCE AGAIN

SUMMARY AND CONCLUSIONS

LETS SUM IT UP

EXCERCISES

SUGGESTED READING / REFERENCES

INTRODUCTION:

In regression analysis specification is the process of developing a regression model. This process consists of selecting an appropriate functional form for the model and choosing which variables to include. As a first step of regression analysis, a person specifies the model. If an estimated model is misspecified, it will be biased and inconsistent.

Specification error occurs when an independent variable is correlated with the error term. There are several different causes of specification error:

- incorrect functional form
- a variable omitted from the model may have a relationship with both the dependent variable and one or more of the independent variables (omitted-variable bias);[2]
- an irrelevant variable may be included in the model
- the dependent variable may be part of a system of simultaneous equations (simultaneity bias)measurement errors may affect the independent variables

One of the assumptions of the classical linear regression model (CLRM) Assumption 9, is that the regression model used in the analysis is "Correctly" specified: If the model is not "Correctly" specified, we encounter the problem of model specification error or model specification bias. In this lesson we take a close and critical look at this assumption, because searching for the correct model is like searching for the Holy Grail.

OBJECTIVES:

1. Understand the model selection criteria for empirical analysis.

2. Understand the specification errors.

3 Understand the consequences of model specification errors on OLS estimates.

4. Detect specification errors through formal econometric tests.

5. Distinguish among the wide range of available tests for detecting specification errors.

MODEL SELECTION CRITERIA:

According to Hendry and Richard, model chosen for empirical analysis should satisfy the following criteria;

1. Be data admissible: that is, predictions made from the model must be logically possible.
2. Be consistent with theory; that is, it must make good economic sense. For example, if Milton Friedman’s permanent income hypothesis holds, the intercept value in the regression of permanent consumption on permanent income is expected to be zero.
3. Have weakly exogenous regressors; that is, the explanatory variables, or regressors, must be uncorrelated with the error term.
4. Exhibit parameter constancy: that is, the value s of the parameters should be stable. Otherwise, forecasting will be difficulty. As Friedman notes, “The only relevant test of the validity of a hypothesis (Model) is comparison of its predictions with experience.” In the absence of parameter constancy, such predictions will not be reliable.
5. Pure Random: Exhibit data coherency; that is, the residual estimated from the model must be purely random (technically, white noise). In other words, if the regression model is adequate, the residuals from this model must be white noise. If that is not the case, there is some specification error in the model. Shortly, we will explore the nature of specification error(s).
6. Be encompassing: that is the model should encompass or include all the rival models in the sense that it is capable of explaining their results. In short, other models cannot be an improvement over the chosen model.

TYPES OF SPECIFICATION ERRORS

Assume that on the basis of the criteria just listed we arrive at the model that we accept as a good model. To be concrete, let this model be

$$Y_i=\beta_1+\beta_2X_i+\beta_3X_i^2+\beta_4X_i^3+u_{2i} \qquad \rightarrow(1)$$

Where Y = total cost of production and X=output. Equation (1) is the familiar text book example of the cubic total cost function,.

But suppose for some reason (say, laziness in plotting the scatter gram) a researcher decides to use the following model:

$$Y_i=\alpha_1+\alpha_2X_i+\alpha_3X_i^2+u_{2i} \qquad \rightarrow(2)$$

Note that we have changed the notation to distinguish this model from the true model. Since (1) is assumed true, adopting (2) would constitute a specification error, the error consisting in omitting a relevant variable(X_i^3). Therefore, the error term u_{2i}. In (2) is in fact

$$u_{2i}=u_{li}+\beta_4X_i^3 \qquad \rightarrow(3)$$

We shall see shortly the importance of this relationship.

Now suppose that another researcher uses the following model;

$$Y_i=\lambda_1+\lambda_2X_i+\lambda_3X_i^2+\lambda_4X_i^3+\lambda_5X_i^4+u_{3i} \qquad \rightarrow(4)$$

If (1) is the "Truth," (4) also constitutes a specification error, the error here consisting in including an unnecessary or irrelevant variable in the sense that the true model assumes λ_5To be zero. The new error term is in fact

$$u_{3i}=u_{1i}-\lambda_5X_i^4 \qquad \rightarrow(5)$$

$=u_{li}$ Since $\lambda_5= 0$ in the true model.

Now assume that yet another researcher postulates the following mode:

$$\text{In } Y_i=\gamma_1+\gamma_2X_i+\gamma_3X_i^2+\gamma_4X_i^3+u_{4i} \qquad \rightarrow(6)$$

In relation to the true model, (6) would also constitute a specification bias, the bias here being the use of the wrong functional form: In (1) Y appears linearly, whereas in (6) it appears log-olinearly.

Finally, consider the researcher who uses the following model:

$$Y^*_i = \beta^*_i + \beta^*_2 X^*_i + \beta^*_3 X^{*2}_i + \beta^*_4 X^{*3}_i + u^*_i \qquad \rightarrow(7)$$

Where $Y^*_i = Y_i + \varepsilon_i$ and $X^*_i = X_i + w_i$, ε_i and w_i Being the errors of measurement. What (7) states is that instead of using the true Y_i And X_i we use their proxies, Y^*_i and X^*_i Which may contain errors of measurement. Therefore, in (7) we commit the errors of measurement bias. In applied work data are plagued by errors of approximations or errors of incomplete coverage or simply errors of omitting some observations. In the social sciences we often depend on secondary data and usually have no way of knowing the types of errors, if any, made by the primary data-collecting agency.

Another type of specification error relates to the way the stochastic error μ_i (or μ_t) enters regression model. Consider for instance, the following bivariate regression model without the intercept term;

$$Y_i = \beta X_i u_I \qquad \rightarrow(8)$$

Where the stochastic error term enters multiplicatively with the property that. satisfies the assumptions of the CLRM, against the following model

$$Y_i = \alpha X_i + u_i \qquad \rightarrow(9)$$

Where the error term enters additively. Although the variables are the same in the two models, we have denoted the slope coefficient in (8) by β and the sple coefificient in (9) by α Now if (8) is the "correct" or "true" model, would the estimated α provide an unbiased estimate of the true β^2 That is, will $E(\hat{\alpha}) =$ If that is not the case, improper stochastic specification of the error term will constitute another source of specification error.

To sum up, in developing an impirical model, one is likely to commit one or more of the following specification errors:

1. Omission of a relevant variable(s)
2. Inclusion of an unnecessary variable(s)

3. Adopting the wrong functional form
4. Errors of measurement
5. In correct specification of the stochastic error term

Before turning to an examination of these specification errors in some detail, it may be fruitful to distinguish between model specification errors and model mis-specification errors. The first four types of error discussed above are essentially in the nature of model specification errors in that we have in mind a 'true" model but somehow we donot estimate the correct model. In model mis-specification errors, we do not know what the true model is to begin with. In this context one may recall the controversy between the Keynesians and the monetarists. The monetarists give primacy to money in explaining changes in GDP, whereas the Keynesians emphasize the role of government expenditure to explain changes in GDP. So to speak there are two competing models.

In what follows, we will first consider model specification errors and then examine model mis-specification errors.

CONSEQUENCES OF MODEL SPECIFICATION ERRORS

Whatever the sources of specification errors, what are the consequences? To keep the discussion simple, we will answer this question in the context of the three-variable model and consider in this section the first two types of specification errors discussed earlier, namely (1) underfitting a model, that is, omitting relevant variables, and (2) overfitting a model, that is, including unnecessary variables. Our discussion here can be easily generalized to more than two regressors, but with tedious algebra.., matric algebra becomes almost a necessity once we go beyond the three variable case.

Underfitting a Model (Omitting a Relevant Variable)

Suppose the true model is

$Y_i = \beta_1 + \beta_2X_{2i} + \beta_3X_{3i} + u_i$

But for some reason we fit the following model:

$Y_i = \alpha_1 + \alpha_2 X_{2i} + v_i$

The consequences of omitting variable X_3 are as follows:

1. If the left-out, or omitted, variable X_3 is correlated with the included variable X_2 that is r_{23}, the correlation coefficient between the two variables is nonzero, $\hat{\alpha}_1 and \hat{\alpha}_2$ are basied as well as inconsistent. That is $E(\hat{\alpha}_1) \neq \beta_1$ and $E(\hat{\alpha}_2) \neq \beta_2$ the bias does not disappear as the sample size get larger.
2. Even if X_2 and X_3 are not correlated $\hat{\alpha}_1$ although $\hat{\alpha}_2$ is now unbiased.
3. The disturbance variance σ^2 is incorrectly estimated.
4. The conventionally measured variance $\hat{\alpha}_1(= \sigma^2 / \sum x_{2i}^2$ is a biased estimator of the variance of the true estimator $\hat{\beta}_1$
5. In consequence, the usual confidence interval and hypothesis-testing procedures are likely to give misleading conclusions about the statistical significance of the estimated parameters.
6. As another consequence, the forecasts based on the incorrect mode l and the forecast (confidence) intervals will be unreliable.

$$E(\hat{\alpha}_2) = \beta_2 + \beta_3 b_{32}$$

Where $b_{32}$is the slope in the regression of the excluded variable X_3 on the included variable X_2(b32 = $\Sigma x_{3i} x_{2i} / \Sigma x_{2i}^2$). As shows, $\hat{\alpha}_2$ is biased, unless β_3 and β_{32} or both are zero. We rule out β_3 being zero, because in that case we do not have specification error to being with. The coefficient β_{32} will be zero if X_2 and X_3 are uncorrelated, which is unlikely in most economic data.

Now let us examine the variances of $\hat{\alpha}_2$and $\hat{\beta}_2$

$\text{Var}(\hat{\alpha}_2) = \frac{\sigma^2}{\Sigma x_{2i}^2}$

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

Where VIF (a measure of collinearity) is the variance inflation factor [=1/(1-r_{23}^2)] is the correlation coefficient between variable X_2 and X_3.

INCLUSION OF AN IRRELEVANT VARIABLE (OVERFITTING A MODEL)

Now let us assume that

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i$$

Is the truth, but we fit the following model.

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + u_i$$

And thus commit the specification error of including an unnecessary variable in the model.

The consequences of this specification error are as follows:

1. The OLS estimators of the parameters of the “incorrect” model are all unbiased and consistent, that is $E(\alpha_1) = \beta_1$, $E(\hat{\alpha}_2) = \beta_2$,and $E(\hat{\alpha}_3) = \beta_3 = 0$
2. 2. The error variance σ^2 Is correctly estimated.
3. The usual confidence interval and hypothesis-testing procedures remain valid.
4. However, the estimated α's will be generally inefficient, that is, their variances will be generally larger than those of the $\hat{\beta}s$of the true model.

From the usual OLS formula we know that

$$\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\Sigma x_{2i}^2}$$

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

Therefore $$\frac{\text{Var}(\hat{\alpha}_2)}{\text{var}(\hat{\beta}_2)} = \frac{1}{1-r_{23}^2}$$

Since $0 \le r_{23}^2 \le 1$, it follows that $(\hat{\alpha}_2) \ge \text{var}\hat{\beta}_2$; that is, the variance of $\hat{\alpha}_2$ is generally greater than the variance of $\hat{\beta}_2$ even though, on average $\hat{\alpha}_2 = \hat{\beta}_2$.

The implication of this finding is that the including of the unnecessary variable X_3 makes the variance of $\hat{\alpha}_2$ larger than necessary, thereby making $\hat{\alpha}_2$ less precise. This is also true of $\hat{\alpha}_1$

TESTS OF SPECIFICATION ERRORS

Detecting the presence of unnecessary variables (Over fitting a model)

Suppose we develop a K-variable model to explain a phenomenon:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \text{................} + \beta_k X_{ki} + u_i$$

However, we are not totally sure that, say, the variable X_k really belongs in the model. One simple way to find this out is to test the significance that we are not sure whether, say β_k with the usual *t* test: $t = \hat{\beta}_k / se(\hat{\beta}_k)$ But suppose that we are not sure whether, say, X_3 and X_4 legitimately belong in the model. This can be easily ascertained by the F test. Thus, detecting the presence of an irrelevant variable(or variables) is not a difficult task.

It is, however, very important to remember that in carrying out these tests of significance we have a specific model in mind. We accept that model as the maintained hypothesis or the "truth," however tentative it may be. Given that model, then, we can find out whether one or more regressors are really relevant by the usual *t* and *f* tests. But note carefully that we should not use the *t* and *f* tests to build a model iteratively, that is, we should not say that initially Y is related to X_2 only because $\hat{\beta}_2$is statistically significant and then expand the model to include X_3 and decide to keep that variable in the model if $\hat{\beta}_3$ turns out to be statistically significant, and so on. This strategy of building model is called the bottom-up approach (starting with a smaller model and expanding it as one goes along) or by the somewhat pejorative term, data mining (other names are regression fishing, data grubbing, data snooping , and number crunching).

Tests for Omitted Variables and incorrect functional form

In practice we are never sure that the model adopted for empirical testing is "the truth, the whole truth and nothing but the truth." On the basis of theory or introspection and prior empirical work, we develop a model that we believe captures the essence of the subject under study. We then subject the model to empirical testing. After we obtain the results, we being the post mortem, keeping in mind the criteria of a good model discussed earlier. It is at this stage that we come to know if the chosen model is adequate. In determining model adequacy, we look at some broad features of the results, such as the R^2 value, the estimated coefficients in relation to their prior expectations, the Durbin-Watson statistic, and the like. If these diagnostics are reasonably good, we proclaim that the chosen model is a fair representation of reality. By the same token, if the results do not look encouraging because the R^2 value is too low or because very few coefficients are statistically significant or have the correct signs or because the Durbin-Watson *d* is too low, then we being to worry about model adequacy and look for remedies. May we have omitted an important variable, or have used the wrong functional form, or have not first differenced the time series (to remove serial correlation), and so on.

The Durbin-Watson *d* Statistics Once Again.

If we examine the routinely calculated Durbin-Watson d we see that for the linerar cost function the estimated *d* suggesting that there is positive "correlation" in the estimated residuals: for n = 10 and k' = 1 and then 5 percent *d* critical value are d_L Liewise, the computed value for the quadratic cost function is 1..38, whereas the 5 percent critical values are d_L = 0.697 and D_U= 1.641, indicating indecision. But if we use the modified *d* test we can say that there is positive "correlation" in the residuals, for the computed d is less than d_U. For the cubic cost function, the true specification, the estimated *d* value does not indicate any positive "correlation" in the residuals.

The observed positive "correlation" in the residuals when we fit the linear or quadratic model is not a measure of (first oder) serial correlation but of fact that some variable(s) that belong in the modeol are included in the error term and need to be culled out from it and introduced in their own right as explanatory variables: If we exclude the $x_1^3$from the cost function, the error term in the mis-specified model is in fact ($\mu_{li} + \beta_4 X_1^3$and it will exhibit a systematic pattern (e.g. positive autocorrelation) if X_1^3in fact affects Y significantly.

To use the Durbin-Watson test for detecting model specification error(s), we proceed as follows

1. From the assumed mode, obtain the OLS residuals.
2. If it is believed that the assumed model is mis-specified because it excludes a relevant explanatory variable, say, Z from the model, order the residuals obtained in Step 1 according to increasing values of Z. Note: The Z variable

could be one of the x variables included in the assumed model or it could be some function of that variable, such as X^2 and X^3.

3. Compute the d statistic from the residuals thus ordered by the usual d formula, namely

$$d = \left(\frac{\prod_{1=2}^{n} (\hat{\mu}_t - \hat{\mu}_{t-1})^2}{\sum_{t=1}^{n} \hat{\mu}_t^2} \right)$$

Note: The subscript t is the index of observation here and does not necessarily mean that the date are time series.

4. From the Durbin-Watson tables, if the estimated d value is significant, then one can accept the hypothesis of model mis-specification. If that turns out to be the case, the remedial measures will naturally suggest themselves. Ramsey's Reset Test. Ramsey has porposed a general test of specification error called RESET (regression specification error test0. Here we will illustrate only the simplest version of the test. To fix ideas, let us continue with out cost-output example that the cost function is linerar in output as.

   $Y_i = \lambda_1 + \lambda_1 X_i + \mu_{3i}$

Where Y= total cost and X= output. Now if we plot the residuals $\hat{\mu}_i$ obtained from this regression against $\hat{Y}_i$ the estimated Y_i from this model, we get the picture shown in figure Although $\sum \hat{\mu}_i$ and $\sum \hat{\mu}_i \hat{Y}_i$ are necessarily zero.

the residuals in this figure show a pattern in which theirmean changes systematically with $\hat{Y}_i$. This would suggest that if we introduce $\hat{Y}_i$ in some form as regressor (s), it should increase R^2. And if the increase in, R^2 is statistically significant (on the basis of the F test discussed in previous Lesson), it would suggest that the liner cost function was mis-specified. This sis essentially the idea behind RESET. The steps involved in RESET are as follow:

1 From the chosen model, obtain the estimated Y$_i$, that is $\hat{Y}_1$.

   $Y_i = \lambda_1 + \lambda_2 X_i + u_{3i}$

2 Rerun (13.4.6) introducing $\hat{Y}_1$ in some form as an additional regressor(s). From Figure ,we observe that there is a curvilinear relationship between $\hat{\mu}_i$ and $\hat{Y}_i$. Suggesting that one can introduce $\hat{Y}_i^2 and \hat{Y}_i^3$ as additional regressors Thus, we run.

   $$Y_i = \beta_1 + \beta_2 X_i + \beta_3 \hat{Y}_i^2 + \beta_4 \hat{Y}_i^3 + u_i$$

3 Let the R^2 obtained from be R$^2_{new}$ and that obtained from be R$^2_{old}$ Then we can use the F test first introduced in namely.

   $$F = \frac{(R^2_{new} - R^2_{old})/\text{number of regressors}}{(1 - R^2_{new})/(n - \text{number of parametres in the new model}}$$

   to find out if the increase in R^2 from using is statistically significant.

4 **Lagrange Multiplier (LM) Test for Adding Variables**. This is an alternative to Ramsey's RESET test. To illustrate this test, we will continue with the preceding illustrative example.

If we compare the linear cost function with the cubic cost function the former is a restricted version of the latter.The restricted regression assumes that the coefficients of the squared and cubed output terms are equal to zero. To test this, the LM test proceeds as follows;

1 Estimate the restricted regression by OLS and obtain the residuals $\hat{u}_i$.

2 If in fact the unrestricted regression is the true regression the residuals obtained in should be related to the squared and cubed output terms, that is X_i^2andX_i^3

3 This suggests that we regress the $\hat{u}_i$obtained in Step 1 on all the regressors(including those in the restricted regression) which in the present case means.

$\hat{u}_i = \alpha_1 + \alpha_2 X_i + \alpha_3 X_i^2 + \alpha_4 X_i^3 + \nu_i$

Where v is an error term with the usual properties.

4 For large-sample size, Engle has shown that n(the sample size0 times the R^2 Estimated from the (auxiliary) regression follows the chisquare distribution with df equal to the number of resrtrictions imposed by the restricted regression, two in the present example since the terms X_i^2 and $X_i^3$are dropped from the model.

Symbolically, we write.

$nR^2 \underset{asy}{\sim} X^2_{\text{(number of restrictions)}}$

Where as Y means asymptotically, that is, in large samples.

5 If the chi-square value obtained from exceeds the critical chi-square value at the chosen level of significant, we reject the restricted regression. Otherwise, we do not reject it.

SUMMARY AND CONCLUSIONS:

1. The assumption of the CLRM that the econometric model used in analysis is correctly specified has two meanings. One, there are no equation specification errors, and two, there are no model specification errors. In this lesson the major focus was on equation specification errors.

2. The equation specification errors discussed in this lesson were

(1) omission of important variable(s), (2) inclusion of superfluous variable(s), (3) adoption of the wrong function form, (4) incorrect specification of the error term ui, and (5) errors of measurement in the regressand and regressors.

3. When legitimate variables are omitted from a model, the consequences can be very serious: The OLS estimators of the variables retained in the model not only are biased but are inconsistent as well. Additionally, the variances and standard errors of these coefficients are incorrectly estimated, thereby vitiating the usual hypothesis-testing procedures.

4. The consequences of including irrelevant variables in the model are fortunately less serious: The estimators of the coefficients of the relevant as well as "irrelevant" variables remain unbiased as well as consistent, and the error variance σ2 remains correctly estimated. The only problem is that the estimated variances tend to be larger than necessary, thereby making for less precise estimation of the parameters. That is, the confidence intervals tend to be larger than necessary.

5. To detect equation specification errors, we considered several tests, such as (1) examination of residuals, (2) the Durbin–Watson *d* statistic, (3) Ramsey's RESET test, and (4) the Lagrange multiplier test**.**

6. A special kind of specification error is errors of measurement in the values of the regressand and regressors. If there are errors of measurement in the regressand only, the OLS estimators are unbiased as well as consistent but they are less efficient. If there are errors of measurement in the regressors, the OLS estimators are biased as well as inconsistent.

7. Even if errors of measurement are detected or suspected, the remedies are often not easy. The use of instrumental or proxy variables is theoretically attractive but not always practical. Thus it is very important in practice that the researcher be careful in stating the sources of his/her data, how they were collected, what definitions were used, etc. Data collected by

official agencies often come with several footnotes and the researcher should bring those to the attention of the reader

LETS SUM IT UP:

In last, we can say that specification error occurs when an independent variable is correlated with the error term. In this process we find appropriate functional form for the model and choosing which variables to include. If particular estimated model is mis-specified, it will give biased and inconsistent results.

EXCERCISES :

Consider the model

$$Yi = \beta 1 + \beta 2X^*I + ui$$

In practice we measure X*Xi such that

a. $Xi = X^*i + 5$

b. $Xi = 3X^*i$

c. $Xi = (X^*i + \varepsilon i)$, where εi is a purely random term with the usual properties

What will be the effect of these measurement errors on estimates of true

β1 and β2?

Suppose that the true model is

$$Yi = \beta 1Xi + ui \qquad (1)$$

but instead of fitting this regression through the origin you routinely fit the usual intercept-present model:

$$Yi = \alpha0 + \alpha1Xi + vi \qquad (2)$$

Assess the consequences of this specification error

Suppose that the "true" model is

$$Yi = \beta1 + \beta2X2i + ut \qquad (1)$$

but we add an "irrelevant" variable X3 to the model (irrelevant in the sense that the true β3 coefficient attached to the variable X3 is zero) and

estimate

$$Yi = \beta1 + \beta2X2i + \beta3X3i + vi \qquad (2)$$

a. Would the R2 and the adjusted R2 for model (2) be larger than that for model (1)?

b. Are the estimates of β1 and β2 obtained from (2) unbiased?

c. Does the inclusion of the "irrelevant" variable X3 affect the variances of ˆ β1 and ˆ β2?

what are the consequences of model specification errors?

What are the various tests used for detecting specification errors?

1.10 Suggested Reading / References:

1. Baltagi, B.H.(1998). Econometrics, Springer, New York.

2. Chow,G.C.(1983). Econometrics, McGraw Hill, New York.

3. Goldberger, A.S.(1998). Introductory Econometrics, Harvard University Press, Cambridge, Mass.

4. Green, W.(2000). Econometrics, Prentice Hall of India, New Delhi.

5. Gujarati, D.N.(1995). Basic Econometrics. McGraw Hill, New Delhi.

6. Koutsoyiannis,A.(1977). Theory of Econometrics(2nd Esdn.). The Macmillan Press Ltd. London.

7. Maddala, G.S.(1997). Econometrics, McGraw Hill; New York.

LESSON-2

NESTED VERSUS NON-NESTED MODELS

STRUCTURE

INTRODUCTION:

In carrying out specification testing, it is useful to distinguish between nested and non-nested mode,s. To distinguished between the two, consider the following models:

Model A: $Y_i = \beta_1 + \beta_2X_{2i} + \beta_3X_{3i} + \beta_4X_{4i} + \beta_5X_{5i} +$

u_i Model B: $Y_i = \beta_1 + \beta_2X_{2i} + \beta_3X_{3i} + u_i$

We say that Model B is nested in Model A because it is a special case of Model A: if we estimate Model A and test the hyp0othesis that $\beta_4 = \beta_5 = 0$ and do not reject it on the basis of , say, the F test Model A reduces to Model B. If we add variable X_4 to Model B, then Model A will reduce to Model B if β_5 is zero; here we will use the t test tot est the hypothesis that the coefficient of X_5 is zero.

Without calling them such, the specification error tests we have discussed previously and the restricted F are essentially tests of nested hypothesis.

Now consider the following modes:

Model C: $Y_i = \alpha_1 + \alpha_2X_{2i} + \alpha_3X_{3i} + u_i$

Model D: $Y_i = \beta_1 + \beta_2Z_{2i} + \beta_3Z_{3i} + u_i$

Where the X's And Z's are different variables. We say that Models C and D are non-nested because one cannot be derived as a special case of the other. In economics, as in other sciences, more than one competing theory may explain a phenomenon. Thus the monetarists would emphasize the role of money in explaining changes in GDP, whereas the Keynesians may explain them by changes in government expenditure.

It may be noted here that one can allow Model C and D to contain regressors that are common to both. For example, X_3 could be included in Model D and Z_2 could be included in Model C. Even then these are non nested models, because Model C does not contain Z_3 and Model D does not contain X_2.

Even if the same variables enter the model, the functional form may make two models non-nested. For example, consider the model:

Model E: $Y_i = \beta_1 + \beta_2 \mathit{InZ}_{2i} + \beta_3 \mathrm{In}Z_{3i} + w_i$

Models D and E are non-nested, as one cannot be derived as a special case of the other.

Since we already have looked at tests of nested model (t and F tests), in the following section we discuss some of the tests of non-nested model, which earlier we called model mis-specification errors.

OBJECTIVES:

1. The first objective is to distinguish between nested and non-nested models.

2. Understand the model selection criteria for empirical analysis.

3. Detect nested and non-nested models through formal econometric tests.

4. Distinguish among the wide range of available tests for detecting non-nested models.

TESTS OF NON-NESTED HYPOTHESES

According to Harvey, there are two approaches to testing non-nested hypotheses:

(1) the **discrimination approach,** where given two or more competing models, one chooses a model based on some criteria of goodness of fit, and (2) the **discerning approach** (my terminology) where, in investigating one model, we take into account information provided by other models. We consider these approaches briefly.

The Discrimination Approach:

Consider Models C and D above. Since both models involve the same dependent variable, we can choose between two (or more) models based on some goodness-of-fit criterion, such as $R2$ or adjusted $R2$, which we have already discussed. But keep in mind that in comparing two or more models, the regress and must be the same. Besides these criteria, there are other criteria that are also used. These include **Akaike's information criterion (AIC), Schwarz's information criterion (SIC),** and **Mallows's *Cp* criterion.**

The Discerning Approach:

The Non-Nested *F* Test or Encompassing *F* Test. Consider Models C and D introduced earlier. How do we choose between the two models? For this purpose suppose we estimate the following nested or *hybrid* model:

Model F: $Yi = \lambda 1 + \lambda 2X2i + \lambda 3X3i + \lambda 4Z2i + \lambda 5Z3i + ui$

Notice that Model F *nests or encompasses* models C and D. But note that C is not nested in D and D is not nested in C, so they are non-nested models.

Now if Model C is correct, $\lambda 4 = \lambda 5 = 0$, whereas Model D is correct if $\lambda 2 = \lambda 3 = 0$. This testing can be done by the usual *F* test, hence the name non-nested *F* test.

However, there are problems with this testing procedure. *First*, if the *X*’s and the *Z*’s are highly correlated, then, as noted in the lesson on multicollinearity, it is quite likely that one or more of the λ’s are individually statistically insignificant, although on the basis of the *F* test one can reject the hypothesis that all the slope coefficients are simultaneously zero. In this case, we have no way of deciding whether Model C or Model D is the correctmodel. *Second*, there is another problem. Suppose we choose Model C as the *reference hypothesis* or model, and find that all its coefficients are significant. Now we add *Z*2 or *Z*3 or both to the model and find, using the *F* test, that their incremental contribution to the explained sum of squares (ESS) is statistically insignificant. Therefore, we decide to choose Model C. But suppose we had instead chosen Model D as the reference model and found that all its coefficients were statistically significant. But when we add *X*2 or *X*3 or both to this model, we find, again using the *F* test, that their incremental contribution to ESS is insignificant. Therefore, we would have chosen model D as the correct model. Hence, “the choice of the reference hypothesis could determine the outcome of the choice model,”[33] especially if severe multicollinearity is present in the competing regressors. *Finally*, the artificially nested model *F* may not have any economic meaning.

Davidson–MacKinnon *J* Test.

Because of the problems just listed in the non-nested *F* testing procedure, alternatives have been suggested. One is the *Davidson–MacKinnon J* test. To illustrate this test, suppose we want to compare hypothesis or Model C with hypothesis or Model D. The ***J* test** proceeds as follows:

1. We estimate Model D and from it we obtain the estimated *Y* values, $\hat{Y}_{Di}$.

2. We add the predicted *Y* value in Step 1 as an additional regressor to

Model C and estimate the following model:

$Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 \hat{Y}D_i + u_i$ **(5)** where the $\hat{Y}D_i$ values are obtained from Step 1. This model is an example of the **encompassing principle,** as in the Hendry methodology.

3. Using the *t* test, test the hypothesis that $\alpha_4 = 0$.

4. If the hypothesis that $\alpha_4 = 0$ is not rejected, we can accept (i.e., not

reject) Model C as the true model because $\hat{Y}D_i$ included in (5), which represent the influence of variables not included in Model C, have no additional explanatory power beyond that contributed by Model C. In other words, Model C *encompasses* Model D in the sense that the latter model does not contain any additional information that will improve the performance of Model C. By the same token, if the null hypothesis is rejected, Model C cannot be the true model (why?).

5. Now we reverse the roles of hypotheses, or Models C and D. We now estimate Model C first, use the estimated *Y* values from this model as regressor in (5), repeat Step 4, and decide whether to accept Model D over Model C. More specifically, we estimate the following model:

$$Y_i = \beta_1 + \beta_2 Z_{2i} + \beta_3 Z_{3i} + \beta_4 \hat{Y}C_i + u_i \qquad \textbf{(6)}$$

where $\hat{Y}C_i$ are the estimated *Y* values from Model C. We now test the hypothesis that $\beta_4 = 0$. If this hypothesis is not rejected, we choose Model D over C. If the hypothesis that $\beta_4 = 0$ is rejected, choose C over D, as the latter does not improve over the performance of C.

Although it is intuitively appealing, the *J* test has some problems. Since the tests given in (5) and (6) are performed independently, we have the following likely outcomes

Hypothesis: $\alpha_4 = 0$

Hypothesis: $\beta 4 = 0$	**Do not reject**	**Reject**
Do not reject	Accept both C and D	Accept D, rejectC
Reject	Accept C, reject D	Reject both C and D

As this table shows, we will not be able to get a clear answer if the *J* testing procedure leads to the acceptance or rejection of both models. In case both models are rejected, neither model helps us to explain the behavior of *Y*. Similarly, if both models are accepted, as Kmenta notes, “the data are apparently not rich enough to discriminate between the two hypotheses [models].” Another problem with the *J* test is that when we use the *t* statistic to test the significance of the estimated *Y* variable in models (5) and (6), the *t* statistic has the standard normal distribution only asymptotically, that is, in large samples. Therefore, the *J* test may not be very powerful (in the statistical sense) in small samples because it tends to reject the true hypothesis or model more frequently than it ought to.

SUMMARY AND CONCLUSIONS:

If errors of measurement are detected or suspected, the remedies are often not easy. The use of instrumental or proxy variables is theoretically attractive but not always practical. Thus it is very important in practice that the researcher be careful in stating the sources of his/her data, how they were collected, what definitions were used, etc. Data collected by official agencies often come with several footnotes and the researcher should bring those to the attention of the reader. Model mis-specification errors can be as serious as equation specification errors. In particular, we distinguished between nested and nonnested models. To decide on the appropriate model we discussed the nonnested, or encompassing, *F* test and the Davidson–MacKinnon *J* test and pointed out the limitation of each test.

LETS SUM IT UP:

In concluding remarks, we can say that Model mis- specification errors can lead to various equation specification errors. In this lesson, we distinguished between nested and non-nested models. Hendry argues several econometric work starts with very simplified models and that not enough diagnostic tests are applied to check whether something is wrong with the maintained model. His suggested strategy is to start with a very general model and then progressively simplify it by some data based simplification tests.

EXCERCISES:

Distinguish between nested and non-nested models?

What is the discrimination approach of non nested hypotheses?

Elaborate the discerning approach of non nested hypotheses?

What is Davidson–MacKinnon J Test

2.7 Suggested Reading / References:

1. Baltagi, B.H.(1998). Econometrics, Springer, New York.

2. Chow,G.C.(1983). Econometrics, McGraw Hill, New York.

3. Goldberger, A.S.(1998). Introductory Econometrics, Harvard University Press, Cambridge, Mass.

4. Green, W.(2000). Econometrics, Prentice Hall of India, New Delhi.

5. Gujarati, D.N.(1995). Basic Econometrics. McGraw Hill, New Delhi.

6. Koutsoyiannis,A.(1977). Theory of Econometrics(2nd Esdn.). The Macmillan Press Ltd. London.

7. Maddala, G.S.(1997). Econometrics, McGraw Hill; New York.

LESSON-3

RECURSIVE LEAST SQUARES AND CHOW’S PREDICTION FAILURE TEST

STRUCTURE

INTRODUCTION

OBJECTIVES

RECURSIVE LEAST SQUARES

CHOW’S PREDICTION FAILURE TEST

SUMMARY AND CONCLUSIONS

LETS SUM IT UP

EXCERCISES

SUGGESTED READING / REFERENCES

INTRODUCTION:

We examined the question of the structural stability of a regression model involving time series data and showed how the Chow test can be used for this purpose. Specifically, you may recall that in that lesson we discussed a simple savings function savings as a function of income) for the United States for the period 1970-1995. There we saw that the savings income relationship probably changed around 1982. There we saw that the savings income relationship probably changed around 1982. Knowing the point of the structural break we were able to confirm it with the Chow test.

But what happens if we do not know the point of the structural break (or breaks)? This is where one can use recursive least squares (RELS).

OBJECTIVES:

1. The key objective is to find the structural stability of a regression model.

2. Use the Recursive least Squares (RELS) to find the point of structural breaks.

3. Understand the Chow test and how this test can be used for showing structural stability of a regression model.

RECURSIVE LEAST SQUARES

The basic IDEA behind RELS is very simple and can be explained with the saving –income regression.

$Y_t = \beta_1 + \beta_2X_1 + u_1$

Where Y=savings and Z = income and where the sample is for the period 1970-1995.

Suppose we first use the date for 1970-1974 and estimate the savings function, obtaining the estimates of β_1 and β_2. Then we use the data for 1970-1975 and again estimate the savings function and obtain the estimates of the two parameters. Then we use the data for 1970-1976 and re-estimate the savings model. In this fashion we go on adding an additional data point on Y and X until we exhaust the entire sample. As you can imagine, each regression run will given you a new set of estimates of β_1 and β_2 If you plot the estimated values of these parameters change. If the model under consideration is structurally stable, the changes in the estimated values of the two parameters will be small and essentially random. However, if the estimated values of the parameters change significantly, it would indicate a structural break. RELS is thus a use routine with time series data since time is ordered chronologically. It is also a useful diagnostic toll in cross-sectional data where the data are ordered by some “size” or “scale” variable, such as the employment or asset size of the firm.

Software packages such as Shazam, Eviews, and microfit now do recursive least-squares estimates routinely. RELS also generates recursive residuals on which several diagnostic tests have been based.

CHOW’S PREDICTION FAILURE TEST

Chow has shown that his test can be modified to test the predictive power of a regression model. Again, we will revert to the U.S savings-income regression for the period 1970-1995.

Suppose we estimate the savings-income regression for the period 1970-1981, obtaining $\hat{\beta}_{1,70-81}$which $\hat{\beta}_{2,70-81}$are the estimated intercept and slope coefficients based on the data for 1970-1981. Now using the actual value of income for period 1982-1995 and the intercept and slope values for the period 1970-1981, we predict the values of savings for each for 1982-1995 year. The logic here is that if there is no serious structural change in the parameter estimates for the earlier period, should not be very different from the actual values of savings prevailing in the latter period. Of course, if there is a vast difference between the actual and predicted values of savings for the latter period, it will cast doubts on the stability of the savings-income relation for the entire data period.

Whether the difference between the actual and estimated savings value is large or small can be tested by the F test as follows:

$$F = \frac{(\sum \hat{u}_t^2 - \sum \hat{u}_t^2)/n_2}{(\sum \hat{u}_t^2)/(n_1 - k)}$$

Where n_1=Number of observations in the first period (1970-1981) on which the initial regression is based n_2 number of observations in the second or forecast period $\sum \hat{u}_t^2 = RS$ when the equation estimated for all the observation (n_1-n_2) and $\sum \hat{u}_t^2 = RS$ when the equation is estimated for the first n_1 observations and k is the number of parameters estimated (two in the present instance).

SUMMARY AND CONCLUSIONS:

This lesson has discussed the functional form of the regression model.. The final sections of the lesson described hypothesis tests designed to reveal whether the assumed model had changed during the sample period, or was different for different groups of observations. These tests rely on information about when (or how) the sample is to be partitioned for the test. In many time series cases, this is unknown.Tests designed for this more complex case were considered in this lesson like Recursive least squares (RELS) and Chow’s Prediction Failure Test.

LETS SUM IT UP:

In the concluding remarks, we can say that for showing the structural stability of a regression model involving time series data we can use Chow test for this purpose. We also mentioned about the Recurssive least squares test for showing the points of structural breaks.

EXCERCISES:

State with reason whether the following statements are true or false.†

a. An observation can be influential but not an outlier.

b. An observation can be an outlier but not influential.

c. An observation can be both influential and an outlier.

d. If in the model Yi = β1 + β2Xi + β3Xi+ uiˆ β3 turns out to be statistically significant, we should retain the linear term Xi even if ˆ β2 is statistically insignificant.

e. If you estimate the model Yi = β1 + β2X2i + β3X3i + ui or Yi = α1 + β2x2i + β3x3i + ui by OLS, the estimated regression line is the same, where x2i = (X2i − ¯X 2) and x3i = (X3i − ¯X3).

Elaborate the Chow’s prediction failure test?

A regression model with K = 16 independent variables is fit using a panel of seven years of data. The sums of squares for the seven separate regressions and the pooled regression are shown below. The model with the pooled data allows a separate constant for each year. Test the hypothesis that the same coefficients apply in everyyear.

	1954	**1955**	**1956**	**1957**	**1958**	**1959**	**1960**	**All**
Observations	65	55	87	95	103	87	78	570
e_e	104	88	206	144	199	308	211	1425

Comment on the stability of estimated coefficients

through examples?

Explain the Recursive Least Squares(RELS)?

Evaluate the following statement made by Henry Theil*:

Given the present state of the art, the most sensible procedure is to interpret confidence coefficients and significance limits liberally when confidence intervals and test statistics are computed from the final regression of a regression strategy in the conventional way. That is, a 95 percent confidence coefficient may actually be an 80 percent confidence coefficient and a 1 percent significance level may actually be a10 percent level.

3.9 Suggested Reading / References:

1. Baltagi, B.H.(1998). Econometrics, Springer, New York.

2. Chow,G.C.(1983). Econometrics, McGraw Hill, New York.

3. Goldberger, A.S.(1998). Introductory Econometrics, Harvard University Press, Cambridge, Mass.

4. Green, W.(2000). Econometrics, Prentice Hall of India, New Delhi.

5. Gujarati, D.N.(1995). Basic Econometrics. McGraw Hill, New Delhi.

6. Koutsoyiannis,A.(1977). Theory of Econometrics(2nd Esdn.). The Macmillan Press Ltd. London.

7. Maddala, G.S.(1997). Econometrics, McGraw Hill; New York.

LESSON-4

NON LINEAR REGRESSION MODELS

STRUCTURE

EXCERCISES

SUGGESTED READING / REFERENCES

INTRODUCTION:

Intrinsically linear and intrinsically nonlinear regression models:

Some models may look nonlinear in the parameters but are inherently or intrinsically linear because with suitable transformation they can be made linear in the parameter regression models. But if such models cannot be linearized in the parameters, they are called intrinsically nonlinear regression models. From now on when we talk about a nonlinear regression model, we mean that it is intrinsically non linear. For brevity, we will call them NLRM.

But a simple mathematical trick will render it a linear regression model, namely,

$$\text{in}\frac{1-Y_i}{Y_i} = \beta_1 + \beta_2 X_i + u_i \quad \text{----------------1}$$

Consider now the famous Cobb-Douglas (C-D) production function. Letting Y=output, X_2 = labor input and X_3 = capital input, we will write this function in three different ways:

$$Y_i = \beta_1 X_{2i}^{\beta 2} X_{3i}^{\beta 3} e^{ui} \text{------------------------ 2}$$

or

$$\text{In}Y_i = \alpha + \beta_2 InX_{2i} + \beta_3 InX_{3i} + u_i \text{ ---------------------- 2a}$$

Where α = in β_1. Thus in this format the C-D function is intrinsically linear. Now consider this version of the C-D function.

$$Y_i = \beta_1 X_{2i}^{\beta 2} X_{3i}^{\beta 3} u_i \text{ ----------------------- 3}$$

or

$\text{InY}_i = \alpha + \beta_2 InX_{2i} + \beta_3 InX_{3i} + Inu_i$ -----------------------3a

Where α = in β_1. This model too is linear in the parameters. But now consider the following version of the C-D function.

$Y_i = \beta_1 X_{2i}^{\beta 2} X_{3i}^{\beta 3} u_i$ ----------------------------4

As we just noted, C-D versions (2a) and 3a) are intrinsically linear (in the parameter) regression models, but there is no way to transform (4) so that the transformed model can be made linear in the parameters.[2] Therefore, (4) is intrinsically a nonlinear regression model.

Another well-known but intrinsically nonlinear function is the constant elasticity of substitution (CES) production function of which the Cobb Douglas production is a special case. The CES production takes the following form:

$Yi = A[\delta K_i^{-\beta} + (1-\delta)L_i^{-\beta}]^{-1/\beta}$

Where Y= output, K= capital input, L=labour input, A= scale parameter, δ = distribution parameter ($0 < \delta < 1$) and β = substitution parameter ($\beta \geq -1$). No matter in what form you enter the stochastic error term u_i in this production function, there is no way to make it a linear (in parameter) regression model. It is intrinsically a nonlinear regression model.

OBJECTIVES:

1. Estimating nonlinear regression models: the trial and error method
2. Approaches to estimating nonlinear regression models.

ESTIMATING NONLINEAR REGRESSION MODELS: THE TRIAL AND ERROR METHOD

To set the stage, let us consider a concrete example. The data in Table relates to the management fees that a leading mutual fund in the United states pays to its investment advisors to manage its assets. The fees paid depend on the net asset value of the fund.

To see how the exponential regression model in fits the data, we can proceed by trial and error. Suppose we assume that

Initially β_1 = 0.45 and β_2 = 0.45. These are pure guesses, sometimes based on prior experience or prior empirical work or obtained by just fitting a linear regression model even though it may not be appropriate. At this stage do not worry about how these values are obtained.

Since we know the values of β_1 and β_2 we can write (2) as

$u_i = Y_i - \beta\,{}_{ie}^{\beta 2Xi} = Y_i - 45\ {}_{e}^{0.01Xi}$

Therefore

$\Sigma\square_{\square}^{2} = \Sigma\,(\square_{\square}\ \ - 0.45\square^{0.01\square\square})^{2}$

Since Y, X, $\beta_1$and β_2 are known, we can easily find the error sum of squares in (2). Remember that in OLS our objective is to find those values of the unknown parameters that will make the error sum of squares as small as possible. This will happen if the estimated & values from the model are as close as possible to the actual & values. With the given values, we obtain$\Sigma\square_{\square}^{2}$ =0.3044 But how do we know that this is the least possible error sum of squares that we can obtain.? What happens if you choose another value for β_1 and β_2 respectively? Repeating the procedure just laid down, we find that we now obtain $\Sigma\square_{\square}^{2}$ =0.0073. Obviously, this error sum of squares is much smaller than the one obtained before, namely 0.3044. But how do we know that we hve reached the lowest possible error sum of squares, for by choosing yet another set of values for the β's, we will obtain yet another error sum of squares?

As you can see, such a trial and error, or iterative, process can be easily implemented. And if one has infinite time and infinite patience, the trial and error process may ultimately produce value β_1 and β_2 that guarantee the lowest possible error sum of squares. But you might ask, how did we go from (β_1 = 0.45; β_2 =0.01) to (β_1 = 0. 50; β_2 = -0.1)? Clearly, we need some kind of algorithm that will tell us how we go from one set of values of the unknowns to another set before we stop.

APPROACHES TO ESTIMATING NONLINEAR REGRESSION MODELS:

There are several approaches, or algorithms, to NLRMs: (1) direct search or trial or error, (2)direct optimization, and (3) iterative linearization.

Direct Search or Trial-and-Error or Derivative-Free Method

In the previous section we showed how this method works. Although intutitively appealing because it doesnot require the use of calculus methods as the other methods do , this method is general not used. First, if an NLRM involves several parameters, the method become very cumbersome and computationally expensive. For example, if an NLRM involves 5 parameters and 25 alternative values for each parameter are considered, you will have to compute the error sum of squares $(25)^5$= 9,765,625 times! Second, there is no guarantee that the final set of parameter values you have selected will necessarily give you the absolute minimum error sum of squares. In the language of calculus, you may obtain a local and not an absolute minimum. In fact, no method guarantees a global minimum.

Direct Optimization

In direct optimization we differentiate the error sum of squares with respect to each unknown coefficient, or parameter, set the resulting equation to zero, and solve the resulting normal equations simultaneously. Some iterative routine is therefore called for. One routine is called the method of steepest descent. We will not discuss the technical details of this method as they are somewhat involved, but the reader can find the details in the references. Like the method of rtrial and error, the method of steepest descent also involves selecting initial tral values of the unknown parameters but then it proceeds more systematically than the hit-or-miss or trial – and-error method. One disadvantage of this method is that it may converge to the final values of the parameters extremely slowly.

Iterative Linearizartion Method

In this method we linearize a nonlinear equation around some initial values of the parameters. The linearized equation is then estimated by OLS and the initially chosen values are adjusted. These adjusted values are used to relinearize the model, and again we estimate it by OLS and readjust the estimated values. This process is continues until there is no substantial change in the estimated values from the last couple of iterations. The main technique sued in linearizing a nonlinear equation is the Taylor series expansion from calculus.

SUMMARY AND CONCLUSIONS:

The main points discussed in this lesson can be summarized as follows:

1. Although linear regression models predominate theory and practice there are occasions where non linear-in-the-parameter regression models (NLRM) are useful.
2. The mathematics underlying linear regression models is comparatively simple in that one can obtain explicit, or analytical, solutions of the coefficients of such models. The small sample and large sample theory of inference of such models is well established.
3. In contrast, for intrinsically nonlinear regression models, parameter values cannot be obtained explicitly. They have to be estimated numerically that is, by iterative procedures.

4. There are several methods of obtaining estimates of NLRMs, such as (1) trial and error, (2) non linear least squares(NLLS) and (3) Linearizartion through Taylor series expansion.
5. Computer packages now have built-in routines, such as Gauss-Newton, Newton-Raphson, and Marquard. These are all iterative routines.
6. NLLS estimators do not possess optimal properties in finite samples, but in large samples they do have such properties. Therefore, the results of NLLS in small samples must be interpreted carefully.
7. Autocorrelation, heteroscedasticity, and model specification problems can plague NLRM, as they do liner regression models.
8. We illustrated the NLLS with several examples. With the ready availability of user friendly software packages, estimation of NLRM should no longer be a mystery.

LETS SUM IT UP:

In this lesson, we extended the regression model to a form which allows nonlinearity in the parameters in the regression function. The results for interpretation, estimation, and hypothesis testing are quite similar to those for the linear model. The two crucial differences between the two models are, first, the more involved estimation procedures needed for the nonlinear model and, second, the ambiguity of the interpretation of the coefficients in the nonlinear model (since the derivatives of the regression are often nonconstant, in contrast to those in the linear model.) Finally, we added two additional levels of generality to the model. A nonlinear instrumental variables estimator is suggested to accommodate the possibility that the disturbances in the model are correlated with the included variables. In the second application, two-step nonlinear least squares is suggested as a method of allowing a model to be fit while including functions of previously estimated parameters.

EXCERCISES:

Q1 What is under fitting the model ?

Q2 describe discerning approach for testing non-nested models.

Q3 describe davidson mackinnon J test.

Q4 what is meant by intrinsically linear models?

Q5 describe recursive least square regression.

Q6 discuss miss specification and its remedies.

Suggested Reading / References:

1. Baltagi, B.H.(1998). Econometrics, Springer, New York.

2. Chow,G.C.(1983). Econometrics, McGraw Hill, New York.

3. Goldberger, A.S.(1998). Introductory Econometrics, Harvard University Press, Cambridge, Mass.

4. Green, W.(2000). Econometrics, Prentice Hall of India, New Delhi.

5. Gujarati, D.N.(1995). Basic Econometrics. McGraw Hill, New Delhi.

6. Koutsoyiannis,A.(1977). Theory of Econometrics(2nd Esdn.). The Macmillan Press Ltd. London.

7. Maddala, G.S.(1997). Econometrics, McGraw Hill; New York.

UNIT-4

Regression with qualitative variable and other Techniques

Lesson-1

Dummy variable regression models

STRUCTURE

INTRODUCTION:

In statistics and econometrics, particularly in regression analysis, a dummy variable (also known as an indicator variable, design variable, Boolean indicator, categorical variable, binary variable, or qualitative variable) is one that takes the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. Dummy variables are used as devices to sort data into mutually exclusive categories (such as smoker/non-smoker, etc.).

OBJECTIVES:

1. Understand the concept of dummy variables.

2. To understand the ANOVA(Analysis of Variance) models.

3. To understand the ANCOVA(Analysis of Covariance) models.

ANOVA Model :

A regression model in which the dependent variable is quantitative in nature but all the explanatory variables are dummies (qualitative in nature) is called an Analysis of Variance (ANOVA) model.

ANCOVA Model :

A regression model that contains a mixture of both quantitative and qualitative variables is called an Analysis of Covariance (ANCOVA) model. ANCOVA models are extensions of ANOVA models. They are statistically control for the effects of quantitative explanatory variables (also called covariates or control variables).

1.3 . 1 ANOVA MODELS:

To illustrate the ANOVA models, consider the following example.

Example 1.1

PUBLIC SCHOOL TEACHERS’ SALARIES BY GEOGRAPHICAL
REGION

Table 1.1 gives data on average salary (in dollars) of public school teachers in 50 states and the District of Columbia for the year 1985. These 51 areas are classified into three geo- graphical regions: (1) Northeast and North Central (21 states in all), (2) South (17 states in all), and (3) West (13 states in all). For the time being, do not worry about the format of the table and the other data given in the table.

Suppose we want to find out if the average annual salary (AAS) of public school teachers differs among the three geographical regions of the country. If you take the simple arithmetic average of the average salaries of the teachers in the three regions, you will find that these averages for the three regions are as follows: $24,424.14 (Northeast and North Central), $22,894 (South), and $26,158.62 (West). These numbers look different, but are they statistically different from one another? There are various statistical techniques to compare two or more mean values, which generally go by the name of analysis of variance**.** But the same objective can be accomplished within the framework of regression analysis.

EXAMPLE 1

TABLE 1 AVERAGE SALARY OF PUBLIC SCHOOL TEACHERS, BY STATE, 1986

Salary	Spending	*D*	*D*3	Salary	Spending	*D*	*D*
		2				2	3

19,583	334	1	0	22,795	336	0	1
20,263	311	1	0	21,570	292	0	1
20,325	355	1	0	22,080	298	0	1
26,800	464	1	0	22,250	373	0	1
29,470	466	1	0	20,940	285	0	1
26,610	488	1	0	21,800	253	0	1
30,678	571	1	0	22,934	272	0	1
27,170	553	1	0	18,443	230	0	1
25,853	416	1	0	19,538	264	0	1
24,500	354	1	0	20,460	312	0	1
24,274	315	1	0	21,419	275	0	1
27,170	362	1	0	25,160	342	0	1
30,168	378	1	0	22,482	394	0	0
26,525	424	1	0	20,969	250	0	0
27,360	398	1	0	27,224	544	0	0
21,690	356	1	0	25,892	404	0	0
21,974	315	1	0	22,644	340	0	0
20,816	305	1	0	24,640	282	0	0
18,095	296	1	0	22,341	229	0	0
20,939	328	1	0	25,610	293	0	0
22,644	391	1	0	26,015	370	0	0
24,624	451	0	1	25,788	412	0	0
27,186	434	0	1	29,132	360	0	0
33,990	502	0	1	41,480	834	0	0
23,382	359	0	1	25,845	376	0	0
20,627	282	0	1				

Note: *D*2 = 1 for states in the Northeast and North Central; 0 otherwise.

*D*3 = 1 for states in the South; 0 otherwise.

To see this, consider the following model:

$Yi = \beta 1 + \beta 2D2i + \beta 3i\, D3i + ui$ **(1)**

where Yi = (average) salary of public school teacher in state i

$D2i$ = 1 if the state is in the Northeast or North Central

= 0 otherwise (i.e., in other regions of the country)

$D3i$ = 1 if the state is in the South

= 0 otherwise (i.e., in other regions of the country)

Note that (1) is like any multiple regression model considered previously, except that, instead of quantitative regressors, we have only qualitative, or dummy, regressors, taking the value of 1 if the observation belongs to a particular category and 0 if it does not belong to that category or group. Hereafter, we shall designate all dummy variables by the letter D. Table 1.1 shows the dummy variables thus constructed.

What does the model tell us? Assuming that the error term satisfies the usual OLS assumptions, on taking expectation of (1) on both sides, we obtain:

Mean salary of public school teachers in the Northeast and North Central:

$E(Yi \mid D2i = 1, D3i = 0) = \beta 1 + \beta 2$ **(2)**

Mean salary of public school teachers in the South:

$$E(Y_i \mid D_{2i} = 0, D_{3i} = 1) = \beta_1 + \beta_3 \quad \textbf{(3)}$$

You might wonder how we find out the mean salary of teachers in the West. If you guessed that this is equal to β_1 , you would be absolutely right, for

Mean salary of public school teachers in the West:

$$E(Y_i \mid D_{2i} = 0, D_{3i} = 0) = \beta_1 \quad \textbf{(4)}$$

In other words, the mean salary of public school teachers in the West is given by the inter- cept, β_1, in the multiple regression (1), and the “slope” coefficients β_2 and β_3 tell by how much the mean salaries of teachers in the Northeast and North Central and in the South differ from the mean salary of teachers in the West. But how do we know if these differences are statistically significant? Before we answer this question, let us present the results based on the regression (1). Using the data given in Table 1, we obtain the following results:

$$\hat{Y}_i = 26{,}158.62 - 1734.473D_{2i} - 3264.615$$

$$\text{se} = (1128.523) \quad (1435.953) \quad (1491.615) \quad \textbf{(5)}$$

$$t = (23.\ \ 1759\ (-1.2078\ \ \ (-2.1776$$
$$(0.0000\ \ \ \ (0.2330)\ \ \ \ (0.0349\ \ R2\ \ \ \ =$$

where * indicates the *p*
values.

As these regression results show, the mean salary of teachers in the West is about

$26,158, that of teachers in the Northeast and North Central is lower by about $1734, and that of teachers in the South is lower by about $3265. The actual mean salaries in the last two regions can be easily obtained by adding these differential salaries to the mean salary of teachers in the West, as shown in Eqs. (3) and (4). Doing this, we will find that the mean salaries in the latter two regions are about $24,424 and $22,894.

But how do we know that these mean salaries are statistically different from the mean salary of teachers in the West, the comparison category? That is easy enough. All we have to do is to find out if each of the "slope" coefficients' in (5) is statistically significant. As can be seen from this regression, the estimated slope coefficient for Northeast and North Central is not statistically significant, as its *p* value is 23 percent, whereas that of the South is statistically significant, as the *p* value is only about 3.5 percent. Therefore, the overall conclusion is that statistically the mean salaries of public school teachers in the West and the Northeast and North Central are about the same but the mean salary of teachers in the South is statistically significantly lower by about $3265.

A caution is in order in interpreting these differences. The dummy variables will simply point out the differences, if they exist, but they do not suggest the reasons for the differences.

1.5 CAUTION IN THE USE OF DUMMY VARIABLES:

(The dummy variable trap)

Although they are easy to incorporate in the regression models, one must use the dummy variables carefully. In particular, consider the following aspects:

1. In Example 1, to distinguish the three regions, we used only two dummy variables, $D2$ and $D3$. Why did we not use three dummies to distinguish the three regions? Suppose we do that and write the model (1) as:

$$Y_i = \alpha + \beta 1\ D1i + \beta 2\ D2i + \beta 3\ D3i + ui \qquad \textbf{(6)}$$

where $D1i$ takes a value of 1 for states in the West and 0 otherwise. Thus, we now have a dummy variable for each of the three geographical regions. Using the data in Table 1, if you were to run the regression (6), the com- puter will "refuse" to run the regression (try it). Why? The reason is that in the setup of (6) where you have a dummy variable for each category or group and also an intercept, you have a case of perfect collinearity, that is, exact linear relationships among the variables. Why? Refer to Table 1. Imagine that now we add the D1 column, taking the value of 1 whenever a state is in the West and 0 otherwise. Now if you add the three D columns hor- izontally, you will obtain a column that has 51 ones in it. But since the value of the intercept α is (implicitly) 1 for each observation, you will have a column that also contains 51 ones. In other words, the sum of the three D columns will simply reproduce the intercept column, thus leading to perfect collinearity. In this case, estimation of the model (6) is impossible.

The message here is: If a qualitative variable has m categories, intro- duce only (m − 1) dummy variables. In our example, since the qualitative variable "region" has three categories, we introduced only two dummies. If you do not follow this rule, you will fall into what is called the dummy vari- able trap, that is, the situation of perfect collinearity or perfect multi- collinearity, if there is more than one exact relationship among the vari- ables. This rule also applies if we have more than one qualitative variable in the model, an example of which is presented later. Thus we should restate the preceding rule as: For each qualitative regressor the number of dummy variables introduced must be one less than the categories of that variable. Thus, if in Example 1 we had information about the gender of the teacher, we would use an additional dummy variable (but not two) taking a value of 1 for female and 0 for male or vice versa.

2. The category for which no dummy variable is assigned is known as the base, benchmark, control, comparison, reference, or omitted cate- gory. And all comparisons are made in relation to the benchmark category.

3. The intercept value (β1) represents the mean value of the benchmark category. In Example 1, the benchmark category is the Western region. Hence, in the regression (5) the intercept value of about 26,159 repre- sents the mean salary of teachers in the Western states.

4. The coefficients attached to the dummy variables in (1) are known as the differential intercept coefficients because they tell by how much the value of the intercept that receives the value of 1 differs from the inter- cept coefficient of the benchmark category. For example, in (1.2.5), the value of about −1734 tells us that the mean salary of teachers in the Northeast or North Central is smaller by about $1734 than the mean salary of about

$26,159 for the benchmark category, the West.

5. If a qualitative variable has more than one category, as in our illus- trative example, the choice of the benchmark category is strictly up to the researcher. Sometimes the choice of the benchmark is dictated by the par- ticular problem at hand. In our illustrative example, we could have chosen the South as the benchmark category. In that case the regression results given in (5) will change, because now all comparisons are made in rela- tion to the South. Of course, this will not change the overall conclusion of our example (why?). In this case, the

intercept value will be about $22,894, which is the mean salary of teachers in the South

6. We warned above about the dummy variable trap. There is a way to circumvent this trap by introducing as many dummy variables as the num- ber of categories of that variable, provided we do not introduce the intercept in such a model. Thus, if we drop the intercept term from (.6), and con- sider the following model,

$$Y_i = \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i \qquad \textbf{(7)}$$

we do not fall into the dummy variable trap, as there is no longer perfect collinearity. *But make sure that when you run this regression, you use the no-intercept option in your regression package.*

How do we interpret regression (7)? If you take the expectation of(7), you will find that:

β_1 = mean salary of teachers in the West

β_2 = mean salary of teachers in the Northeast and North Central.

β_3 = mean salary of teachers in the South.

In other words, *with the intercept suppressed, and allowing a dummy variable for each category, we obtain directly the mean values of the various categories.* The results of (7) for our illustrative example are as follows:

$$\hat{Y}_i = 26{,}158.62D_{1i} + 24{,}424.14D_{2i} + 22{,}894D_{3i}$$

$$\begin{array}{lcccc} \text{se} & = & (887.9170) & (986.8645) & \textbf{(8)} \\ t & = & (27.5072) & (23.1987)^{*} & \end{array}$$

$R^2 = 0.0901$

where * indicates that the p values of these t ratios are very small.

As you can see, the dummy coefficients give directly the mean (salary) val- ues in the three regions, West, Northeast and North Central, and South.

6. Which is a better method of introducing a dummy variable: (1) intro- duce a dummy for each category and omit the intercept term or (2) include the intercept term and introduce only $(m - 1)$ dummies, where m is the number of categories of the dummy variable:

1.7 SUMMARY AND CONCLUSIONS:

In summary we can state the Qualitative response regression models refer to models in which the response, or regressand, variable is not quantitative or an interval scale.. The simplest possible qualitative response regression model is the binary model in which the regressand is of the yes/no or presence/absence type. We distinguished between the ANOVA and ANCOVA .

LETS SUM IT UP:

In concluding remarks, we can say that ANOVA and ANCOVA are used for analyzing the regression models which are qualitative in nature.

EXCERCISES:

What do mean by dummy variables?

Explain the Analysis of Variance (ANOVA) ?

Elaborate the Analysis of Covariance(ANCOVA)?

What are the various cautions which should be adopted in the use of dummy variables?

Give an example of ANOVA and ANCOVA models ?

1.10 Suggested Reading / References:

1. Baltagi, B.H.(1998). Econometrics, Springer, New York.

2. Chow,G.C.(1983). Econometrics, McGraw Hill, New York.

3. Goldberger, A.S.(1998). Introductory Econometrics, Harvard University Press, Cambridge, Mass.

4. Green, W.(2000). Econometrics, Prentice Hall of India, New Delhi.

5. Gujarati, D.N.(1995). Basic Econometrics. McGraw Hill, New Delhi.

6. Koutsoyiannis,A.(1977). Theory of Econometrics(2nd Esdn.). The Macmillan Press Ltd. London.

7. Maddala, G.S.(1997). Econometrics, McGraw Hill; New York.

Lesson-2

THE LINEAR PROBABILITY MODEL (LPM)

STRUCTURE

INTRODUCTION

OBJECTIVES

THE LINEAR PROBABILITY MODEL (LPM)

NON-NORMALITY OF THE DISTURBANCES *UI*

HETEROSCEDASTIC VARIANCES OF THE DISTURBANCES

NONFULFILLMENT OF $0 \le E(YI \mid X) \le 1$

SUMMARY AND CONCLUSIONS

LETS SUM IT UP

EXCERCISES

SUGGESTED READING / REFERENCES

INTRODUCTION:

In all the regression models that we have considered so far, we have implicitly assumed that the regressand, the dependent variable, or the *response* variable *Y* is quantitative, whereas the explanatory variables are either quantitative, qualitative (or dummy), or a mixture thereof. In fact, in previous Lesson, on dummy variables, we saw how the dummy regressors are introduced in a regression model and what role they play in specific situations. In this lesson we consider several models in which the regressand itself is qualitative in nature. Although increasingly used in various areas of social sciences and medical research, qualitative response regression models pose interesting estimation and interpretation challenges. In this lesson we only touch on some of the major themes in this area.

OBJECTIVES:

1. Understand the concept of Linear Probability model(LPM).

2. The another objective is to understand the various problems posed by the Linear Probability model(LPM).

THE LINEAR PROBABILITY MODEL (LPM):

To fix ideas, consider the following regression model:

$Y_i = \beta_1 + \beta_2X_i + u_i$ ------------------------- 1

Where X = family income and Y =1 if the family owns a house and 0 if it does not own a house.

Model looks like a typical linear regression model but because the regress and is binary, or dichotomous, it is called a linear probability model (LPM). This is because the conditional expectation of Y_i given X_i, $E(Y_i/X_i)$, can be interpreted as the conditional probability that the event will occur given X_i that is, Pr ($Y_i = 1/X_i$) Thus, in our example, $E(Y_i/X_i)$ Gives the probability of a family owning a house and whose income is the given amount X_i.

The justification of the name LPM for model like (1) can be seen as follows: Assuming $E(u_i) = 0$ usual (to obtain unbiased estimators) we obtain.

$E(Y_i/X_i) = \beta_1 + \beta_2X_i$ -------------------- 2

Now, if P_i = probability that Y_i =1 (that is, the event occurs), and ($1-P_i$) = probability that $Y_i = 0$ (that is, that the event does not occur), the variable Y_i has the following (probability) distribution.

___________________i

Y_i	Probability
0	$1-P_i$
1	P_i
Total	1

That is Yi follows the Beronoulli probability distribution.

Now, by the definition of mathematical expectation, we obtain.

$E(Y_i) = 0(1-P_i) + 1 (P_i) = P_i$----------------- 3

Comparing (2) with (3.), we can equate

$E(Y_i/X_i) = \beta_1 + \beta_2X_i = P_i$ --------------------- 4

That is, the conditional expectation of the model can, in fact, be interpreted as the conditional probability of Y_i in general, the expectation of a Bernoulli random variable is the probability that the random variable equals 1. In passing note that if there are n independent trials, each with a probability p of success and probability (1-p) of failure, and X of these trials represent the number of successes then X is said to follow the binomial distribution. The mean of the binomial distribution is np and its variance is m (1-p). The term success is defined in the context of the problem.

Since the probability P_i must lie between 0 and 1, we have the restriction.

$0 \leq E(Y_i/X_i) \leq 1$ ------------------ 5

That is, the conditional expectation (or conditional probability) must lie between 0 and 1.

From the preceding discussion it would seem that OLS can be easily extended to binary dependent variable regression models. So, perhaps there is nothing new here. Unfortunately, this is not the case, for the LPM poses several problems, which are as follows:

Non-Normality of the Disturbances *ui:*

Although OLS does not require the disturbances (*ui*) to be normally distributed, we assumed them to be so distributed for the purpose of statistical inference. But the assumption of normality for *ui* is not tenable for the LPMs because, like *Yi*, the disturbances *ui* also take only two values; that is, they also follow the Bernoulli distribution. This can be seen clearly if we write equation (1) as

$ui = Yi - \beta 1 - \beta 2Xi$ ------------ **(6)**

The probability distribution of *ui* is

ui	Probability

When Yi =1 $1-\beta 1-\beta 2$ Xi Pi

When $Yi = 0$ $-\beta 1-\beta 2Xi$ $(1-Pi)$ ----------- **(7)**

Obviously, *ui* cannot be assumed to be normally distributed; they follow the Bernoulli distribution. But the nonfulfillment of the normality assumption may not be so critical as it appears because we know that the OLS point estimates still remain unbiased (recall that, if the objective is point estimation, the normality assumption is not necessary). Besides, as the sample size increases indefinitely, statistical theory shows that the OLS estimators tend to be normally distributed generally.As a result, in large samples the statistical inference of the LPM will follow the usual OLS procedure under the normality assumption.

2.5 Heteroscedastic Variances of the Disturbances:

Even if $E(ui) = 0$ and cov $(ui , uj) = 0$ for *i* _= *j* (i.e., no serial correlation), it can no longer be maintained that in the LPM the disturbances are homoscedastic. This is, however, not surprising. As statistical theory shows, for a Bernoulli distribution the theoretical mean and variance are, respectively, p and $p(1-p)$, where p is the probability of success (i.e., something happening), showing that the variance is a function of the mean. Hence the error variance is heteroscedastic.

For the distribution of the error term given in (7), applying the definition of variance, the reader should verify that

$$\text{var}\,(ui) = Pi(1-Pi)$$ -------- **(8)**

That is, the variance of the error term in the LPM is heteroscedastic. Since $Pi = E(Yi \mid Xi) = \beta 1 + \beta 2Xi$, the variance of ui ultimately depends on the values of X and hence is not homoscedastic.

We already know that, in the presence of heteroscedasticity, the OLS estimators, although unbiased, are not efficient; that is, they do not have minimum variance. But the problem of heteroscedasticity, like the problem of non-normality, is not insurmountable. In Lesson 11 we discussed several methods of handling the heteroscedasticity problem. Since the variance of ui depends on $E(Yi \mid Xi)$, one way to resolve the heteroscedasticity problem is to transform the model (1) by dividing it through by

$$\sqrt{}E(Yi/Xi)\ [1 - E(Yi/Xi)] = \sqrt{}Pi\ (1 - Pi) = \text{say } \sqrt{}wi$$

that is,

$$Yi / \sqrt{}\ \ wi = \sqrt{}\beta 1 / wi + \beta 2Xi\ / \sqrt{}wi + ui / \sqrt{}wi \text{ ------------ } \mathbf{(9)}$$

As you can readily verify, the transformed error term in (9) is homoscedastic.

Therefore, after estimating (1), we can now estimate (9) by OLS, which is nothing but the *weighted least squares* (WLS) with *wi* serving as the weights.

In theory, what we have just described is fine. But in practice the true $E(Yi \mid Xi)$ is unknown; hence the weights wi are unknown. To estimate wi , we can use the following two-step procedure:

Step 1. Run the OLS regression (1) despite the heteroscedasticity problem and obtain ˆYi = estimate of the true $E(Yi \mid Xi)$. Then obtain

wiˆ =ˆYi(1−ˆYi), the estimate of wi.

Step 2. Use the estimated wi to transform the data as shown in (9)

and estimate the transformed equation by OLS (i.e., weighted least squares).

2.6 Nonfulfillment of $0 < E(Yi \mid X) < 1$

Since E(Yi | X) in the linear probability models measures the conditional probability of the event Y occurring given X, it must necessarily lie between 0 and 1. Although this is true a priori, there is no guarantee that ˆYi , the estimators of E(Yi | Xi), will necessarily fulfill this restriction, and this is the real problem with the OLS estimation of the LPM. There are two ways of finding out whether the estimated ˆYi lie between 0 and 1. One is to estimate the LPM by the usual OLS method and find out whether the estimated ˆYi lie between 0 and 1. If some are less than 0 (that is, negative), ˆYi is assumed to be zero for those cases; if they are greater than 1, they are assumed to be 1.

The second procedure is to devise an estimating technique that will guarantee that the estimated conditional probabilities ˆYi will lie between 0 and 1. The logit and probit models discussed later will guarantee that the estimated probabilities will indeed lie between the logical limits 0 and 1.

SUMMARY AND CONCLUSIONS:

In summary, we can say that the simplest possible binary regression model is the linear probability model (LPM) in which the binary response variable is regressed on the relevant explanatory variables by using the standard OLS methodology.

Simplicity may not be a virtue here, for the LPM suffers from several estimation problems. Even if some of the estimation problems can be overcome, the fundamental weakness of the LPM is that it assumes that the probability of something happening increases linearly with the level of the regressor. This very restrictive assumption can be avoided if we use the logit and probit models.

LETS SUM IT UP:

In the concluding remarks we can say that in the regression model if the regressand is binary, or dichotomous ,we call that particular model , Linear probability model.

EXCERCISES:

Q.1 What do you mean by linear probability mode(LPM)?

What are the various problems which LPM faces in the estimation?

What are the various weaknesses of the LPM?

In studying the purchase of durable goods Y (Y = 1 if purchased, Y = 0 if no purchase) as a function of several variables for a total of 762 households, Janet A. Fisher*obtained the following LPM results:

TABLE 3 :

Explanatory variable	Coefficient	Standard error
Constant	0.1411	—
1957 disposable income, $X1$	0.0251	0.0118
(Disposable income = $X1$)2, $X2$	− 0.0004	0.0004
Checking accounts, $X3$	−0.0051	0.0108
Savings accounts, $X4$	0.0013	0.0047
U.S. Savings Bonds, $X5$	−0.0079	0.0067
Housing status: rent, $X6$	−0.0469	0.0937
Housing status: own, $X7$	0.0136	0.0712
Monthly rent, $X8$	−0.7540	1.0983
Monthly mortgage payments, $X9$	−0.9809	0.5162
Personal noninstallment debt		

, X10	−0.0367	0.0326
Age, X11	0.0046	0.0084
Age squared, X12	−0.0001	0.0001
Marital status,		
X13 (1 = married)	0.1760	0.0501
Number of children, X14	0.0398	0.0358
(Number of children = X14)2,		
X15	−0.0036	0.0072
Purchase plans, X16		
(1 = planned; 0 otherwise)	0.1760	0.0384

R2 = 0.1336

Notes: All financial variables are in thousands of dollars.

Housing status: Rent (1 if rents; 0 otherwise)

Housing status:Own(1 if owns; 0 otherwise)

a. Comment generally on the fit of the equation.

b. How would you interpret the coefficient of −0.0051 attached to checking account variable? How would you rationalize the negative sign for this variable?

c. What is the rationale behind introducing the age-squared and number of children-squared variables? Why is the sign negative in both cases?

d. Assuming values of zero for all but the income variable, find out the conditional probability of a household whose income is $20,000 purchasing a durable good.

e. Estimate the conditional probability of owning durable good(s), given:

X1 = $15,000, X3 = $3000, X4 = $5000, X6 = 0, X7 = 1, X8 = $500, X9=$300, X10 = 0, X11 = 35, X13 = 1, X14 = 2, X16 = 0.

The *R*2 value in the labor-force participation regression given in Table 3 is 0.175, which is rather low. Can you test this value for statistical significance? Which test do you use and why? Comment in general on the value of *R*2 in such models.

Describe LPM and its limitations, in detail.

2.10 Suggested Reading / References:

1. Baltagi, B.H.(1998). Econometrics, Springer, New York.

2. Chow,G.C.(1983). Econometrics, McGraw Hill, New York.

3. Goldberger, A.S.(1998). Introductory Econometrics, Harvard University Press, Cambridge, Mass.

4. Green, W.(2000). Econometrics, Prentice Hall of India, New Delhi.

5. Gujarati, D.N.(1995). Basic Econometrics. McGraw Hill, New Delhi.

6. Koutsoyiannis,A.(1977). Theory of Econometrics(2nd Esdn.). The Macmillan Press Ltd. London.

7. Maddala, G.S.(1997). Econometrics, McGraw Hill; New York.

Lesson-3

The LOGIT Model And The PROBIT Model

STRUCTURE

INTRODUCTION

OBJECTIVES:

THE LOGIT MODEL

THE PROBIT MODEL

SUMMARY AND CONCLUSIONS

LETS SUM IT UP

EXCERCISES

SUGGESTED READING / REFERENCES

3.1 INTRODUCTION:

As we have seen, the LPM is plagued by several problems, such as (1) nonnormality of ui , (2) heteroscedasticity of ui , (3) possibility of ˆYi lying outside the 0–1 range, and (4) the generally lower $R2$ values. But these problems are surmountable. For example, we can use WLS to resolve the heteroscedasticity problem or increase the sample size to minimize the non-normality problem. By resorting to restricted least-squares or mathematical programming techniques we can even make the estimated probabilities lie in the 0–1 interval.

But even then the fundamental problem with the LPM is that it is not logically a very attractive model because it assumes that $Pi = E(Y = 1 \mid X)$ increases linearly with X, that is, the marginal or incremental effect of X remains constant throughout. Thus, in our home ownership example we found that as X increases by a unit ($1000), the probability of owning a house increases by the same constant amount of 0.10. This is so whether the income level is $8000, $10,000, $18,000, or $22,000. This seems patently unrealistic. In reality one would expect that Pi is nonlinearly related to Xi :

At very low income a family will not own a house but at a sufficiently high level of income, say, X*, it most likely will own a house. Any increase in income beyond X* will have little effect on the probability of owning a house. Thus, at both ends of the income distribution, the probability of owning a house will be virtually unaffected by a small increase in X. Therefore, what we need is a (probability) model that has these two features: (1) As Xi increases, $Pi = E(Y = 1 \mid X)$ increases but never steps outside the 0–1 interval, and (2) the relationship between Pi and Xi is nonlinear, that is, "one which approaches zero at slower and slower rates as Xi gets small and approaches one at slower and slower rates as Xi gets very large.''

The reader will realize that the sigmoid, or S-shaped, curve very much resembles the **cumulative distribution function** (CDF) of a random variable. Therefore, one can easily use the CDF to model regressions where the response variable is dichotomous, taking 0–1 values. The practical question now is, which CDF? For although all CDFs are S shaped, for each random variable there is a unique CDF. For historical as well as practical reasons, the CDFs commonly chosen to represent the 0–1 response models are (1) the logistic and (2) the normal, the former giving rise to the **logit** model and the latter to the **probit** (or **normit**) model.

OBJECTIVES:

1. The key objective is to understand different alternative models to LPM.

2. Understand the LOGIT model.

3. Understand the PROBIT model.

THE LOGIT MODEL:

$P_i = E(Y = 1/X_i) = \beta_1 + \beta_2X_i$--------------------1

Where X is income and Y -1 means the family owns a house. But now consider the following representation of home ownership:

$$P_i = E(Y = 1/X_i) = \frac{1}{1+e^{-(\beta 1+\beta 2X_{i})}} \text{-------------------2}$$

For ease of exposition, we write (2) as

$$P_i = \frac{1}{1+e^{-Z_1}} = \frac{e^z}{1+ e^z} \text{---------------------------------3}$$

Where $Z_i = \beta_1 + \beta_2X_i$

Equation 3 represent what is known as the cumulative logistic distribution function.

It is easy to verify that as Z_i ranges from $-\infty$ to $+\infty$, P_i Ranges between 0 and 1 and that P_i is nonlinearly related to Z_i (i.e. X_i)Thus satisfying the two requirements considered earlier. But it seems that in satisfying these requirements, we have created an estimation problem because P_i is nonlinear not only in X but also in the β'_s as can be seen clearly from (2). This means that we cannot use the familier OLS procedure to estimate the parameters. But this problems is more apparent than real because (2) can be linearized, which can be shown as follows.

If P_i the probability fo owning a house, is given by (3) then ($1-P_i$), the probability of not owning a house, is

$1 - P_i = \frac{1}{1+e^{-Z_1}}$ --------------------------4

Therefore, we can write

$\frac{-}{1-P_i} = \frac{1+e^{Z_i}}{1+e^{-Z_i}} = e^{Z_i}$------------------------------5

Now Pi/($1 - P_i$) is simply the odds ratio in favor of owning a house- the ratio of the probability that a family will own a house to the probability that it will not own a house. Thus if $P_i = 0.8$, it means that odds are 4 to 1 in favour of the family owning a house.

Now if we take the natural log of (5) we obtain a very interesting result, namely

$L_i = In\frac{P_i}{1-P_i} = Z_i$ 6

$= \beta_1 + \beta_2 X_i$

That is, L, the log of the odds ratio, is not only linear in X, but also (from the estimation viewpoint0 linear in the parameters. L is called the logit, and hence the name logit model for models like (6) Notice these features of the logit model.

1. As P goes from 0 to 1 (i.e. as Z varies from - ∞ to + ∞), the logit L goes from - ∞ to + ∞. That is, although the probabilities (of necessity) lie between 0 and 1, the logits are not so bounded.
2. Although l is linear in X, the probabilities themselves are not. This property is in contrast with the LPM model (1) where the probabilities increase linearly with X.
3. Although we have included only a single X variable, or regressor, in the preceding model, one can add as many regressors as may be dictated by the underlying theory.

4. If L, the logit, is positive, it means that when the value of the regressor(s) increases, the odds that the regress and equals 1 (meaning some event of interest happens) increases. If L is negative, the odds that the regress and equal 1 decreases as the value of X increases. To put it differently, the logit becomes negative and increasingly large in magnitude as the odds ratio decreases from 1 to 0 and becomes increasingly large and positive as the odds ratio increases from 1 to infinity.
5. More formally, the interpretation of the logit model given in (6) is as follows: β_2, the slope, measures the change in L for unit change in X, that is, it tells how the log – odds in favor of owning a house change an income changes by a unit, say $ 1000. The intercept β_1 is the value of the log odds in favour of owning a house if income is zero. Like most interpretations of intercepts, this interpretation may not have any physical meaning.
6. Given a certain level of income, say, X, if we actually want to estimate not the odds in favor of owning a house but the probability of owning a house itself, this can be done directly from (3) once the estimate of $\beta_1 + \beta_2$ are available. This, however, raises the most important question. How do we estimate β_1 and β_2 in the first place? The answer is given in the next section.
7. Whereas the LPM assumes that P_i is linearly related to X_i the logit model assumes that the log of the odds ratio is linearly related to X_i.

3.4 THE PROBIT MODEL

The estimating model that emerges from the normal CDF is popularly normit model.

To motivate the probit model, assume that in our home ownership example the decision of the *i*th Family to own a house or not depends on an unobservable utility index I_i(also known as a latent variable), that is determined by one or more explanatory variables, say income X_i in such a way that the larger the value of the index I_i The greater the probability of a family owning a house. We express the index I_i as.

$$I_i = \beta_1 + \beta_2 X_i \text{--------------------------} 1$$

Where X_i is the income of the *i*th Family.

How is the (unobservable0 index relaterd to the actual decision to own a house? As before let Y = 1if the family owns a house and Y = 0 if it does not. Now it is reasonable to assume that there is a critical or threshold level of the index, call it I_i^*such that if I_i exceeds I^*_i the family will own a house, otherwise it will not. The threshold I_i^*Like I_i is not observable, but if we assume that it is normally distributed with the same mean and variance, it is possible not only to estimate the parameters of the index given in (1). But also to get some information about the unobservable index itself. This calculation is also follows.

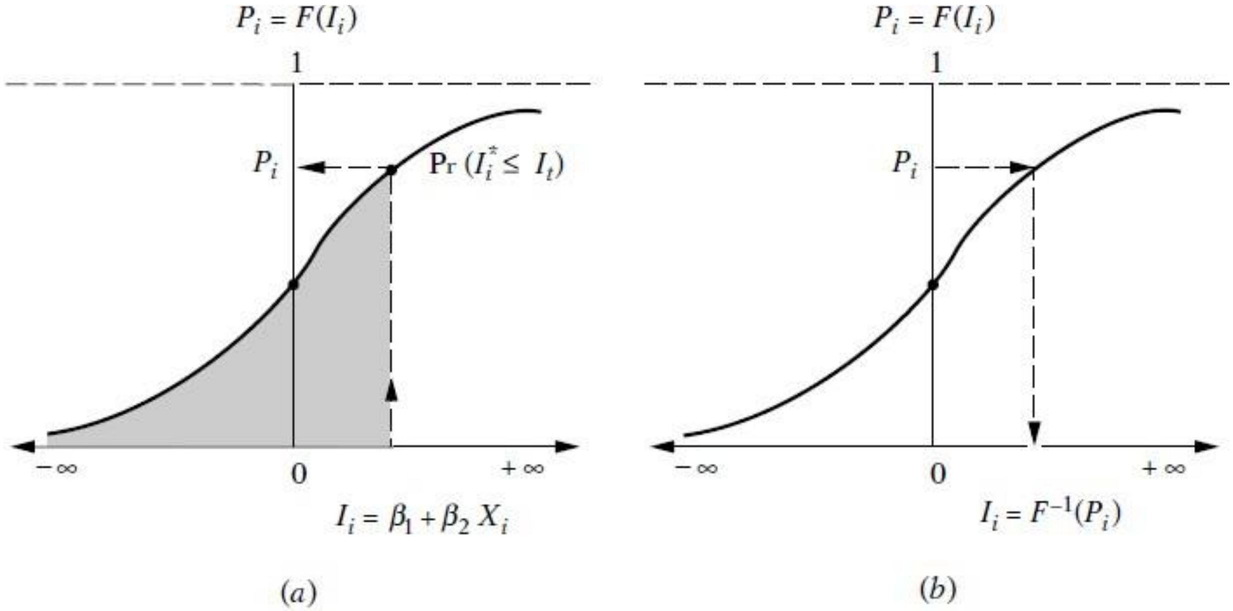
Given the assumption of normality, the probability that I_i^* is less than or equal to I_i can be computed from the standardized normal CDF as.

$$P_i = P(Y=1/X) = P(I_i^* \le I_i) = P(Z_i \le \beta_1 + \beta_2 X_i) = F(\beta_1 + \beta_2 X_i) \text{--------------------} 2$$

Where P(Y = 1/X) means the probability that an event occurs given the value(s) of the X, or explanatory, variable(s) and where Z_i is the standard normal variable, i.e. $Z \sim N(0\text{-}\sigma^2)$. F is the standard normal CDF, which written explicitly in the present context is:

$$F(I_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{I_i} e^{-z^{2/2}} dz$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta_i + \beta_2 X_i} e^{-z^{2/2}} dz \qquad \longrightarrow \qquad 3$$

Since *P* represents the probability that an event will occur, here the probability of owning a house, it is measured by the area of the standard normal curve from – ∞ as shown in Figure.



Probit model: (*a*) given I_i, read P_i from the ordinate; (*b*) given P_i, read I_i from the abscissa.

Now to obtain information, in I_i the utility index, as well asβ_i *and* β_2, we take the inverse of (2) to obtain.

$$I_i = F^{-1}(I_i) = F^{-1}(P_i) \qquad 4$$

$$= \beta_i + \beta_2 X_i$$

Where F^{-1} is the inverse of the normal CDF. What all this means can be made clear from Figure in panel a of this figure we obtain from the ordinate the (cumulative) probability of owning a house given $I_i^* \leq I_i$ whereas in panel b we obtain from the abscissa the value of I_i Given the value of P_i which is simply the reverse of the former.

3.5 SUMMARY AND CONCLUSIONS:

In the logit model the dependent variable is the log of the odds ratio, which is a linear function of the regressors. The probability function that underlies the logit model is the logistic distribution. If the data are available in grouped form, we can use OLS to estimate the parameters of the logit model, provided we take into account explicitly the heteroscedastic nature of the error term. If the data are available at the individual, or micro, level, nonlinear-in-the-parameter estimating procedures are called for. If we choose the normal distribution as the appropriate probability distribution, then we can use the probit model. This model is mathematically a bit difficult as it involves integrals. But for all practical purposes, both logit and probit models give similar results. In practice, the choice therefore depends on the ease of computation, which is not a serious problem with sophisticated statistical packages that are now readily available.

LETS SUM IT UP:

In last ,we conclude that LOGIT and PROBIT models are the alternative models to Linear Probabilty model(LPM). Although LPM, logit, and probit give qualitatively similar results, we will confine our attention to logit and probit models because of the problems with the LPM noted earlier. Between logit and probit, which model is preferable? In most applications the models are quite similar, the main difference being that the logistic distribution has slightly fatter tails. That is to say, the conditional probability Pi approaches zero or one at a slower rate in logit than in probit. Therefore, there is no compelling reason to choose one over the other. In practice many researchers choose the logit model because of its comparative mathematical simplicity.

EXCERCISES:

What are the various alternative models to Linear probability model(LPM) ?

Elaborate the LOGIT model?

Explain the PROBIT model?

Between logit and probit, which model is preferable ?

Distinguish between LOGIT and PROBIT models?

3.8 Suggested Reading / References:

1. Baltagi, B.H.(1998). Econometrics, Springer, New York.

2. Chow,G.C.(1983). Econometrics, McGraw Hill, New York.

3. Goldberger, A.S.(1998). Introductory Econometrics, Harvard University Press, Cambridge, Mass.

4. Green, W.(2000). Econometrics, Prentice Hall of India, New Delhi.

5. Gujarati, D.N.(1995). Basic Econometrics. McGraw Hill, New Delhi.

6. Koutsoyiannis,A.(1977). Theory of Econometrics(2nd Esdn.). The Macmillan Press Ltd. London.

7. Maddala, G.S.(1997). Econometrics, McGraw Hill; New York.

LESSON-4

THE TOBIT MODEL AND MODELING COUNT DATA: THE POISSON REGRESSION MODEL

STRUCTURE

INTRODUCTION:

The TOBIT model is a censored regression model. Observations on the tanent variable y* are missing (or censored) if y* is below (or above) a threshold level. This model has been used in a large number of applications where the dependent variable is observed to be zero for some individuals in the sample(automobile expenditures, medical expectations, hours worked, wages, etc.). However, on careful scrutiny we find that the censored regression model(tobit model) is inappropriate for the analysis of these problems. The Tobit model is, strictly speaking, applicable in only those situations where the latent variable can,in principle, take negative values, but these negative values are not observed because of censoring. Where the zero observations are a consequence of individual decisions, these decisions should be modeled appropriately and the tobit model should not be used mechanically.

OBJECTIVES:

1. To understand the TOBIT model.
2. To understand the POISSON Regression Model.
3. To understand the piecewise linear regression.
4. To understand the various qualitative regression models (Ordinal Logit and Probit Models, Multinomial Logit and Probit Models, and Duration Models)

THE TOBIT MODEL:

An extension of the probit model is the tobit model originally developed by James Tobin, the Nobel laureate economist. To explain this mode, we continue with our home ownership example. In the probit model our concern was with estimating the probabiolity of owning a house as a function of some socioeconomic variables. In the tobit model orur interest is in finding of some the amount of money a person or family spends on a house in relation to socioeconomic variables. Now we face a dilemma here: If a consumer does not purchase a house, obviously we have no data on housing expenditure for such consumers; we have such data only on consumers who actually purchase a house.

Thus consumers are divided into two groups, one consisting of, say n_1 consumers about whom we have information on the regressors (say, income, mortgage interest rate, number of people in the family, etc) as well as the regressand (amount of expenditure on housing) and another consisting of n_2 consumers about whom we have information only on regressors but not on the regressand. A sample in which information on the regressand is available only for some observations is known as a censored sample. Therefore, the tobit model is also known as censored regression model.

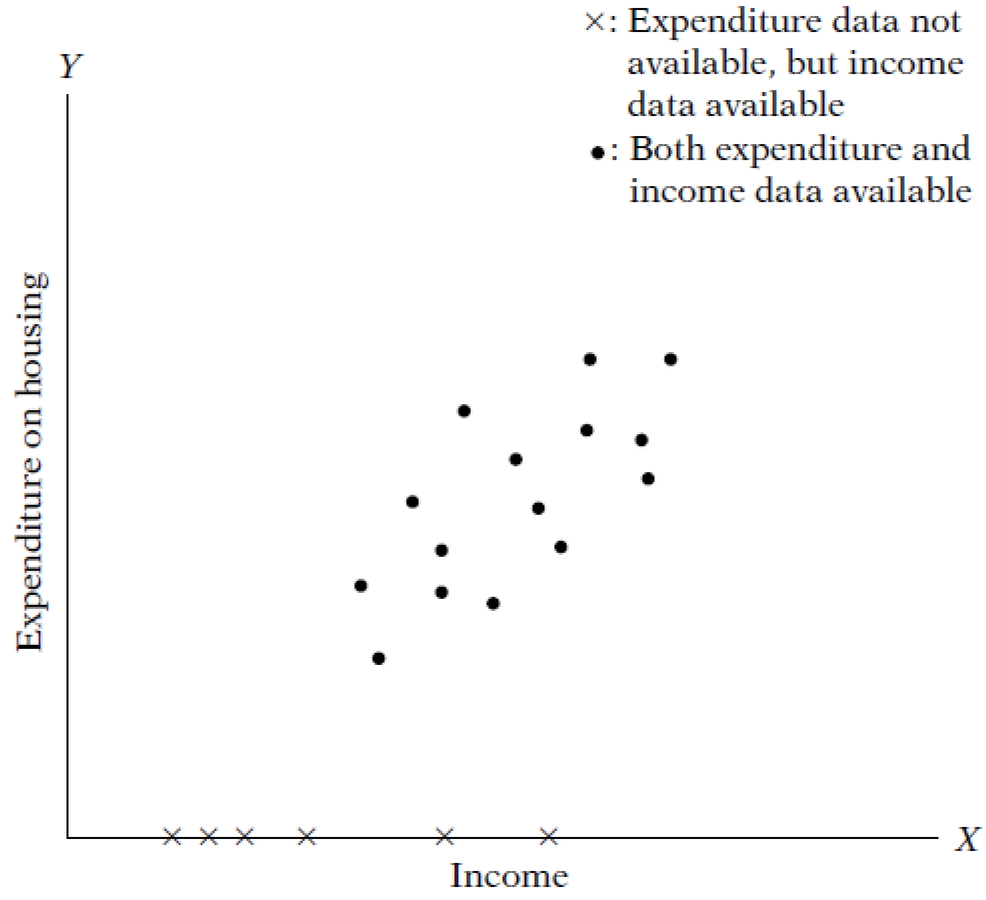
Some authors call such models limited dependent variable regression models because of the restriction put on the values taken by the regressand.

Statistically, we can express the tobit model as

$Y_i = \beta_1 + \beta_2 X_i + u_i$ if RHS > 0 --------------------- 1

$= 0$ otherwise

Where RHS= right hand side. Note. Additional X variables can be easily added to the model.



Plot of amount of monev consumer spends in buvina a house versus income.

Can we estimate regression (1) using only n_1 observations and not worry about the remaining n_2 observations? The answer is no. for the OLS estimates of the parameters obtained from the subset of n_i Observation will be biased as well as inconsistent, that is, they are biased even asymptotically.

To see this, consider Figure As the figure shows, if Y is not observed (because of censoring), all such observations(=n_2) denoted by crosses, will lie on the horizontal axis. If Y is observed, the observations(=n_1) denoted by dots, will lie in the X-Y plane. It if intuitively clear that if we estimate a regression line based on the (n_1+n_2) observations only, the resulting intercept and slope coefficients are bound to be different than if all the … observations were into account.

4.4 MODELING COUNT DATA: THE POISSON REGRESSION MODEL:

There are many phenomena where the regressand is of the count type, such as the number of vacations taken by the family per year, the number of patents received by a firm per year, the number of visits to a dentist of doctor per year, the number of visits to a grocery store per week, the number of parking or speeding tickets received per year, the number of days stayed in a hospital in a given period, the number of cars passing through a toll booth in a span of, say 5 minutes, and so on. The underlying variable in each case is discrete, taking only a finite number of values. Sometimes count data can also refer to rare, or infrequent, occurrences such as getting hit by lightning in a span of a week, winning more than one lottery within 2 weeks, or having two or more heart attacks in a span of 4 weeks. How do we model such phenomena?

Just as the Bernoulli distribution was chosen to model the yes/ no decision in the linear probability mode, the probability distribution that is specifically suited for count data is the Poisson probability distribution. The pdf of the Poisson distribution is given by.

$(Y)_i \frac{\mu^Y e^{-\mu}}{Y!} Y = 0, 1, 2$ ------------------1

Where *f*(Y) denotes the probability that the variable Y takes non-negative integer values, and where Y! (read Y factorial) stands for Y! = Y x (Y -1) x (Y-2) x 2 x 1. It can be proved that.

$E(Y) = \mu$□2

$var\ (Y) = \mu$□3

Notice an interesting features of the Poisson distribution: Its variance is the same as its mean value.

The Poisson regression model may be written as:

$Y_i = E(Y_i) + \mu_i = \mu_i + \mu_I$ -----------------4

Where the Y's are independently distributed as Poisson random variables with mean μ_i For each individual expressed as.

$\mu_i = E(Y_i) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots\ldots\ldots\ldots\ldots\ldots.. \beta_k X_{ki}$ --------------5

Where the X's are some of the variables that might effect the mean value. For example, if our count variable is the number of visits to the Metropolitan Museum of Art in New York in a given year, this number will depend on variables such as income of the consumer, admission price, distance from the museum, and parking fees.

For estimation purposes, we write the model as:

$$Y_i = \frac{\mu^Y e^{-\mu}}{Y!} + \mu_i \text{ -------------------6}$$

With μ replaced by (5) As you can readily see, the resulting regression model will be nonlinear in the parameters.

FURTHER TOPICS IN QUALITATIVE RESPONSE :REGRESSION MODELS

As noted at the outset, the topic of qualitative response regression models is vast. What we have presented in this lesson are some of the basic models in this area. For those who want to pursue this topic further, we discuss below very briefly some other models in this area. We will not pursue them here, for that would take us far away from the scope of this book.

Ordinal Logit and Probit Models:

In the bivariate logit and probit models we were interested in modeling a yes or no response variable. But often the response variable, or regressand, can have more than two outcomes and very often these outcomes are **ordinal** in nature; that is, they cannot be expressed on an interval scale. Frequently, in survey-type research the responses are on a Likert-type scale, such as “strongly agree,” “somewhat agree,” or “strongly disagree.” Or the responses in an educational survey may be “less than high school,” “high school,” “college,” or “professional degrees.” Very often these responses are coded as 0 (less than high school), 1 (high school), 2 (college), 3 (postgraduate). These are ordinal scales in that there is clear ranking among the categories but we cannot say that 2 (college education) is twice 1 (high school education) or 3 (postgraduate education) is three times 1 (high school education).

To study phenomena such as the preceding, one can extend the bivariate logit and probit models to take into account multiple ranked categories. The arithmetic gets quite involved as we have to use multistage normal and logistic probability distributions to allow for the various ranked categories.

For the underlying mathematics and some of the applications, the reader may consult the Greene and Maddala texts. At a comparatively intuitive level, the reader may consult the Liao monograph. Software packages such as Limdep, Eviews, and Shazam have routines to estimate ordered logit and probit models.

Multinomial Logit and Probit Models

In the ordered probit and logit models the response variable has more than two ordered, or ranked, categories. But there are situations where the regressand is unordered. Take, for example, the choice of transportation modeto work. The choices may be bicycle, motorbike, car, bus, or train. Although these are categorical responses, there is no ranking or order here; they are essentially nominal in character. For another example, consider occupational classifications, such as unskilled, semiskilled, and highly skilled. Again, there is no order here. Similarly, occupational choices such as self-employed, working for a private firm, working for a local government, and working for the federal government are essentially nominal in character. The techniques of multinomial logit or probit models can be employed to study such nominal categories. Again, the mathematics gets a little involved. The references cited previously will give the essentials of these techniques. And the statistical packages cited earlier can be used to implement such models, if their use is required in specific cases.

Duration Models

Consider questions such as these: (1) What determines the duration of unemployment spells? (2) What determines the life of a light bulb? (3) What factors determine the duration of a strike? (4) What determines the survival time of a HIV-positive patient? Subjects such as these are the topic of duration models, popularly known as **survival analysis** or **time-to-event data analysis.** In each of the examples cited above, the key variable is the length of time or spell length, which is modeled as a random variable. Again the mathematics involves the CDFs and PDFs of appropriate probability distributions. Although the technical details can be tedious, there are accessible books on this subject.44 Statistical packages such as Stata and Limdep can easily estimate such duration models. These packages have worked examples to aid the researcher in the use of such models.

SUMMARY AND CONCLUSION:

If we choose the normal distribution as the appropriate probability distribution, then we can use the probit model. This model is mathematically a bit difficult as it involves integrals. But for all practical purposes, both logit and probit models give similar results. In practice, the choice therefore depends on the ease of computation, which is not a serious problem with sophisticated statistical packages that are now readily available. If the response variable is of the count type, the model that is most frequently used in applied work is the Poisson regression model, which is based on the Poisson probability distribution. A model that is closely related to the probit model is the tobit model, also known as a censored regression model. In this model, the response variable is observed only if certain condition(s) are met. Thus, the question of how much one spends on a car is meaningful only if one decides to buy a car to begin with. However, Maddala notes that the tobit model is "applicable only in those cases where the latent variable [i.e., the basic variable underlying a phenomenon] can, in principle, take negative values and the observed zero values are a consequence of censoring and nonobservability."There are various extensions of the binary response regression models. These include ordered probit and logit and nominal probit and logit models. The philosophy underlying these models is the same as the simpler logit and probit models, although the mathematics gets rather complicated. Finally, we considered briefly the so-called duration models in which the duration of a phenomenon, such as unemployment or sickness, depends on several factors. In such models, the length, or the spell of duration, becomes the variable of research interest.

LETS SUM IT UP:

In the end we can say that tobit model is a censored regression model which has been used in a large number of applications where the dependent variable is observed to be zero for some individuals in the sample . Further if the response variable is of the count type, the model that is most frequently used in applied work is the Poisson regression model, which is based on the Poisson probability distribution .

EXCERCISES:

Q 1. How modeling of count data is done?

Q 2. Describe piecewise linear regression.

Q 3. Describe TOBIT model in short.

Q 4. Elaborate the Ordinal Logit and Probit Models?

Q 5. Describe the Multinomial Logit and Probit Models?

Q 6. What do you mean by Duration Models?

Suggested Reading / References:

1. Baltagi, B.H.(1998). Econometrics, Springer, New York.

2. Chow,G.C.(1983). Econometrics, McGraw Hill, New York.

3. Goldberger, A.S.(1998). Introductory Econometrics, Harvard University Press, Cambridge, Mass.

4. Green, W.(2000). Econometrics, Prentice Hall of India, New Delhi.

5. Gujarati, D.N.(1995). Basic Econometrics. McGraw Hill, New Delhi.

6. Koutsoyiannis,A.(1977). Theory of Econometrics(2nd Esdn.). The Macmillan Press Ltd. London.

7. Maddala, G.S.(1997). Econometrics, McGraw Hill; New York.