MA in Economics

IIIrd Semester

THEORY OF ECONOMETRICS

Contents

Module-1: Introduction to Econometrics

Meaning - Nature and Scope of Econometrics - Distinction between Economics and Econometrics, Mathematics and Econometrics, Statistics and Econometrics - Methodology of Econometrics - Types of Econometrics

Module-2: Simple Regression Model

Simple Regression: Meaning - Basic Ideas - Significance of Disturbance Term.

Method of Estimation: Ordinary Least Squares and Maximum Likelihood Estimation - BLUE Property - Coefficient of Determination - Assumptions - Hypothesis Testing - Confidence Interval and Test of Significance Approach - Testing Regression Coefficients - Interpretation of Results.

Module-3: Multiple Regression Model

Multiple Regression: Meaning - Three Variable Regression Model - Partial Regression Coefficients - Method of Estimation - R-Square and Adjusted R-Square –

Hypothesis Testing -

Testing Individual Regression Coefficient - Overall Significance Test - ANOVA.

Introduction to Matrix Approach to Estimation of Parameters of more than Three Variables

Module-4: Practical Problems of Regression

Multicollinearity:

Nature - Causes -Consequences - Detection - Remedial Measures.

Heteroscedasticity:

Nature - Causes -Consequences - Detection - Remedial Measures.

Auto-Correlation:

Nature - Causes -Consequences - Detection - Remedial Measures.

Module-5: Dummy Variable and Dynamic Regression Models

Dummy Variable Model: Meaning - Nature - Dummy Variable Trap –

Dummy Variable Model with -

Single Qualitative Variable - Two Qualitative Variables

Dummy Variable Model with Mixture of Qualitative and Quantitative Variables.

Autoregressive and Dynamic Models: Role of Lag in Economics - Estimation Methods: Koyck’s: Adaptive Adjustment and Partial Expectation Models - Almon Approach to Distributed Lag Models.

Module-6: Simultaneous Equation Models

Nature - Simultaneous Equation Bias - Identification: Under - Exact - Over Identification

Rules of Identification - Order and Rank Condition of Identification

Estimation of Simultaneous Equations Models: ILS, 2SLS, 3SLS, LIMLE, FIMLE.

MODULE-1: INTRODUCTION TO ECONOMETRICS

Introduction and Definition

Econometrics is a set of quantitative techniques that are useful for making economic decisions. The goal of econometrics is to help us to move from qualitative theoretical work to quantitative world in which the policy makers work. Econometrics is a rapidly developing branch of economics which, broadly speaking, aims to give empirical content to economic relations. Econometrics applies economic theory and tools of statistics 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 “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, 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. (Gerhard 1968).

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 (P.A. Samuelson, T.C.Koopman, J.R.N Stone).

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 (Goldberger 1964).

Econometrics is concerned with the empirical determination of economic laws (Theil 1971).

Maddala’s definition is similar,

Econometrics is the application of statistical and mathematical methods to the analysis of economic data, with a purpose of giving empirical content to economic theories and verifying them or refuting them.

Econometric is the art and science of using statistical methods for the measurement of economic relations (Chow, 1983).

Based on the above definitions it can be conclude that, econometrics is an integration of economic theories, mathematical economics and statistics with an objective to provide numerical values to the parameters of economic relationships.

Econometrics is no more limited to testing, analyzing and estimating economic theory. Econometrics is also used now in many subjects and disciplines like Finance, Marketing, Management, Sociology etc.

Also, the advent of modern day computers and development of modern software has helped in estimation and analysis of more complex models. So computer programing is now an essential component of modern day econometrics. Therefore, econometrics is the application of mathematics, statistical methods, and computer science to economic data and is described as the branch of economics that aims to give empirical content to economic relations.

Why to Study Econometrics as Separate branch/Discipline?

Econometrics has to be studied in its own right for the following reasons:

Economic theory and econometrics:

* Economic theory makes statements or hypotheses that are mostly qualitative in nature. For example, theory of demand, theory of consumption, theory of production etc, these theories does not provide any numerical measure of the relationship. It means, these theories do not tell about how much increase or decrease in one variable because of a certain unit change in the related variable/s. This is the job of the econometrician. Econometricians provide empirical content to economic theories.

The main purpose of mathematical economics is to represent economic theory in mathematical form without regard to measurability or empirical verification of the theory. Whereas econometrics uses mathematical models developed by mathematical economist to lend empirical support to economic theory.

* Economic statistics is mainly concerned with collecting, processing, and presenting economic data in the form of charts and tables. These are the jobs of the economic statistician. The data thus collected constitute the raw data for econometric work. But the economic statistician does not go any further, not being concerned with using the collected data to test economic theories. Of course, one who does that becomes an econometrician.
* Statistics/ mathematical statistics provides many techniques for econometric analysis, but econometricians often need special and appropriate methods to deal with economic data. Because economic data is not generated by controlled experiments. Therefore, econometrician has to make appropriate modifications to apply statistical tools to analyse economic relations.

Methodology of Econometrics

What is the methodology of econometricians? How do they proceed to answer the economic problems? 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

What is a model/theory?

A model is an abstraction from a real-world process.

What is a model made up of?

* + - * + a set of assumptions
        + a set of relationships which may be... definitional (identities)

behavioral

equilibrium conditions

* + - * + a set of conclusions (propositions, theorems, lemmas, or hypotheses)
  + What’s an example of an economic model?
    - * Keynes’ theory of consumption

The fundamental psychological law … is that men(women) are disposed, as a rule and on average, to increase their consumption as their income increases, but not as much as the increase in their income.

(MPC < 1)

2. Specification of The Mathematical Model of The Theory

Keynes postulated a positive relationship between consumption and income, but he did not specify any function form. For simplicity, we can express in the following form:

*Y = β1 + β2X 0 < β2 < 1*

*Y = consumption expenditure and (*dependent variable)

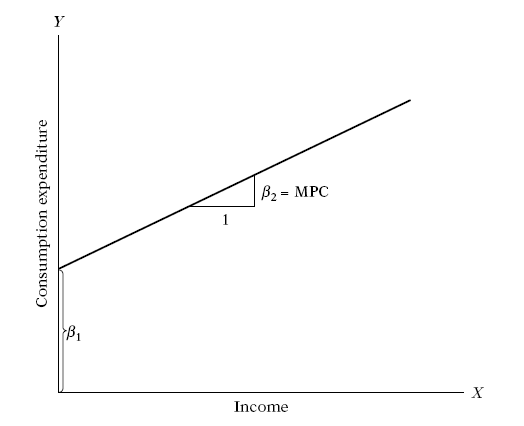
*X =* income, (independent, or explanatory variable)

*β1 =* the intercept

*β2 =* theslope coefficient

The slope coefficient *β2 measures the MPC.*

The above equation shows a linear relationship between consumption and income and it is mathematical model of consumption and income. The above equation can be represented geometrically as follows:



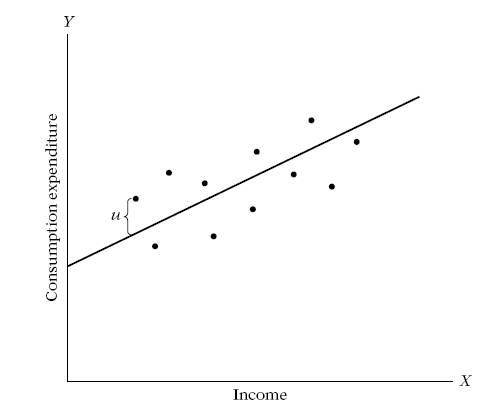
3. Specification of the Statistical or Econometric Model

Mathematical model is limited interest to econometricians, because it assumes exact or deterministic relationship between the variables. But the relationship exists between economic variables are inexact or in deterministic. It is because of individual variations in the subjects of the problem. Therefore to study the inexact relationship between economic variables, the econometrician would modify the deterministic consumption function as.

*Y = β1 + β2X + u*

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

The econometric model of consumption can be depicted as shown in the following Figure:



4. Obtaining the Data

To estimate the econometric model or to obtain the numerical values of the parameters in the model, namely, β1 and β2, we need data. Data available for an econometrician are broadly classified into four types, namely,

Cross-Sectional Data: Sample of entities at a given point in time or observation of an economic phenomenon at a given point in time.

Time Series Data: Observations over time or observation of an economic phenomenon for multiple periods of time.

Pooled Data / Pooled Cross Sections: Combined Cross Sections from different years

Panel / longitudinal Data: Time Series of each Cross Section, Same cross sectional units are followed over time

Example of Cross-Sectional Data

Monthly Income of a sample of individuals in 2019

|  |  |
| --- | --- |
| Respondent | Income (Rupees) |
| A | 72000 |
| B | 63000 |
| C | 37000 |
| D | 68000 |

Example of Time Series Data

(In Billions of Rupees)

|  |  |  |
| --- | --- | --- |
| Year | Personal Consumption Expenditure | Personal Disposal Income |
| 2011 | 7832 | 10228 |
| 2012 | 8334 | 12908 |
| 2013 | 8923 | 13627 |
| 2014 | 9289 | 14782 |
| 2015 | 9768 | 15268 |

Example of Pooled Data / Pooled Cross Sections

Monthly income of respondents from 2016 to 2019

|  |  |  |
| --- | --- | --- |
| Sample Year | Respondent | Income (in Rupees) |
| 2014 | A | 67392 |
| 2014 | B | 74839 |
| 2015 | A | 77453 |
| 2015 | C | 103672 |
| 2016 | B | 79635 |
| 2016 | D | 92346 |

Example of Panel or Longitudinal Data

GDP of Selected Countries

|  |  |  |
| --- | --- | --- |
| country | Year | Billion (in US $) |
| India | 2016 | 784 |
| India | 2017 | 839 |
| India | 2018 | 953 |
| China | 2016 | 3672 |
| China | 2017 | 4035 |
| China | 2018 | 41346 |
| Srilanka | 2016 | 158 |
| Srilanka | 2017 | 189 |
| Srilanka | 2018 | 195 |

5. Estimation of the Parameters of the Econometric Model

Once data are obtained for the variables of the model, the next task is to estimate the parameters, namely, *β1 and β2.* Numerical values of the estimated parameters give empirical content to the econometric model or theory being our testing. In our example it gives empirical content to consumption function. To estimate parameters of the model a statistical tool called regression is used and econometrics heavily relyed on this technique. For hypothetical data, let us assume we obtain estimates for *β1 and β2* of our consumption function, 128 and 0.78 respectively. Thus, the estimated consumption function is:

Yˆ = 128 + 0.78Xi

The hat on the Y indicates that it is an estimate of Y. From the above results we see that the slope coefficient (i.e., the MPC) was about 0.78, suggesting that for the study period an increase in income of one rupee led, on average, to an increase of about 78 paisa in consumption expenditure. We say on average because the relationship between consumption and income is inexact because of individual variations. The numerical value 128 is the estimated value of intercept or constant of model. It can be interpreted as if the value of X, the explanatory variable assumes zero value, then the average or mean value of the dependent variable is 128.

6. Hypothesis Testing

Hypothesis testing is to find out whether the estimates obtained in the previous step are in accord with the expectations of the theory that is being tested. Keynes expected the MPC to be positive but less than 1. In our example we found the MPC to be about 0.78. But before we accept this finding as confirmation of Keynesian consumption theory, we must enquire whether this estimate is sufficiently below unity. In other words, is 0.78 statistically less than 1? If it is, it may support Keynes’ theory.

Such confirmation or refutation of economic theories on the basis of sample evidence is based on a branch of statistical theory known as statistical inference (hypothesis testing).

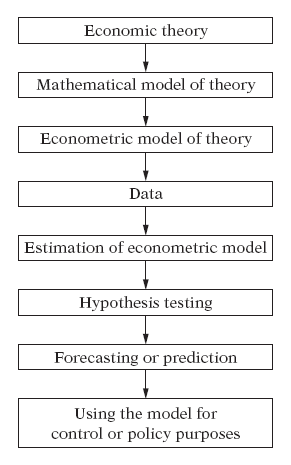
7. Forecasting of Prediction

If the chosen model is in accordance with the theory and support our hypothesis, we may use estimated model to predict the future value of the dependent variable, Y on the basis of the given values of explanatory variable, X.

8. Using the Model for Control or Policy purpose

We can also use the estimated model for produced desired value of the dependent by controlling explanatory variables of model and use the results of model to frame or modify the policy.

Anatomy of Classical Econometric Modeling



* 1. Types of Econometrics

Econometrics is divided broadly in to parts. They are theoretical econometrics and applied econometrics. Each division can be approach by using traditional or Bayesian method.

Traditional

Theoretical

Econometrics Bayesian

Applied Traditional

Bayesian

Theoretical econometrics is concerned with the development of appropriate methods for measuring economic relationships specified by econometric models. It must spell out the assumptions of this method, its properties, and what happens to these properties when one or more of the assumptions of the method are not fulfilled.

In applied econometrics we use the tools of theoretical econometrics to study some special field(s) of economics and business, such as the production function, investment function, demand and supply functions, portfolio theory, etc.

*Terminology and Notations*

The model to be estimated: Yi = ß1 + ß2Xi + ui

^ ^

The estimated model: Yi = ß1 + ß2Xi

^ ^ ^

or Yi = ß1 + ß2Xi + ui

^

where ui = Yi - Yi

Y = β1 + β2 X + u

Left hand-side Variable: Right hand-side Variable:

Dependent Independent

Explained Explanatory

Predictand Predictor

Regressand Regressor

Response Stimulus or control

Endogenous Exogenous

MODULE II

SIMPLE REGRESSION MODEL

Basic Ideas of Regression

The term regression was introduced by Francis Galton in 1886. In his research paper he found that tendency for tall parents to have tall children and for short parents to have short children, but the average height of children born from parents of a given height tended to move (or regress) toward the average height in the population as a whole (F. Galton, “*Family Likeness in Stature*”).

Galton’s Law was confirmed by Karl Pearson by collecting empirical data. He also 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. (K. Pearson and A. Lee, “*On the law of Inheritance*” 1903) By the words of Galton, this was “*Regression to mediocrity*”

Modern Meaning of Regression

The modern meaning and interpretation of regression is distinctly different. Broadly it can be defined as

Regression analysis is concerned with the study of the dependence of one variable (dependent) on one or more other variables (explanatory) with a view to estimating and/or predicting the mean or average value of the former (dependent) in terms of the known or fixed values of the latter explanatory). (Gujarati, p. 15)

Note three things about the above statement:

1) The dependency is inexact or stochastic. As we already noted in specification of econometric model, the relationship between the dependent and independent variables is inexact or stochastic.

2) In the simplest case, the functional form of the relationship is linear. Though the relation between the variables may be linear or non-linear in nature for make analysis as simple as possible the functional form is assumed to linear.

3) It is assumed that there is one-way causality flowing from the independent variable(s) to the dependent variable. Again to study the cause and effect relationship between the dependent and explanatory variables, for simplicity we assume that only explanatory variables influences on dependent variable, not vice-versa.

In spite of the fact that we assume *X* *causes Y*, regression analysis does not imply causation. The idea of causation between the chosen variable is from statistics (regression) but it should be from the economic theory.

The regression analysis is the main tool in econometric analysis. By making use of this tool we are able to estimate or measure the amount changes in one variable because of changes in the related variable/s. Purpose of an economist and the economic theories is to measure such changes in economic phenomenon, hence it is most appropriate tool for economic/econometric analysis.

Simple or Two Variable or Bivariate Regression Model

To understand the application of regression to study the dependency of one variable on other related variables in a simplest manner, here we select only two variables. One is dependent and the other is independent or explanatory variable. Hence, this model is known as two variable or bi-variate regression model.

Hypothetical example:

To understand the application of simple regression model considers the hypothetical data given in the table below.

Table: Weekly family income X ($), and consumption Y ($)

|  |  |
| --- | --- |
| X Y | 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 |
| Mean | 65 77 89 101 113 125 137 149 161 173 |

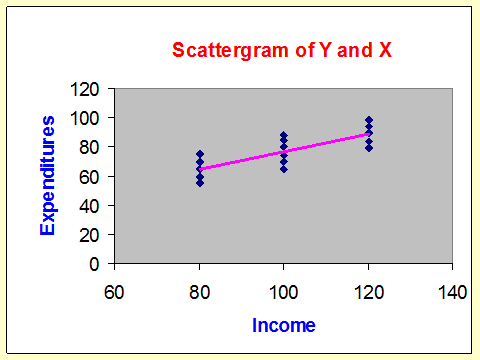
Suppose we want to study the relationship between weekly consumption expenditures (*Y*) and weekly disposable income (*X*). Suppose we have a *population* of 60 families whose weekly consumption expenditures and disposable income are given in the above table. Weekly disposable income is divided in to ten groups and weekly consumption expenditure is distributed is within this ten groups. Therefore, we have ten fixed values of X, namely, weekly disposable income and corresponding Y values, that is, weekly consumption expenditure, against each X values. So to speak, there are 10 Y sub populations.

By observation of consumption expenditure it is noted that, there is considerable variations in each income group. But the general picture one gets is on the average consumption expenditure increases with increase in income which can be observed from the last row of the table. Thus corresponding to weekly income level of $80 the average consumption expenditure is $ 65 and while corresponding to the income level of $ 240, mean consumption expenditure is $173. These mean values are called as conditional expected values, because, they depend on the given values of X. Symbolically we denote these conditional expected values as, E (Y/X), which is read, as the expected value of Y given the value of X.

By using this conditional expected value we are able to better predict the average or mean value of the dependent variable than if we do not have this concept. In other words, by having the knowledge on X variable our prediction of mean is more accurate than do not have information of X variable.

**Population Regression Function (PRF)**

If we plot the conditional distribution of *Y* given *X*, we obtain the following *scattergram* for three income groups:



The line/curve, which passes through the conditional mean values of *Y* given fixed values of *X* is known as *Population Regression Line, PRL* and the functional representation of the line is Population Regression Function (PRF)*.*

Note that the PRF can be written so as to express conditional mean of Y as a function of X,

E(Yi|Xi) = ƒ (Xi) i = 1,2,...,16

Linear, Deterministic and Stochastic PRF

In the above expression the ƒ (Xi) is in general form. That is, it is silent about the functional form. But, whether the functional form is linear or non-line we have to decide before using regression. As we already noted, for simplicity linear functional form is assumed.

A linear and deterministic regression can be written as follows:

E(Yi|Xi) = ß1 + ß2Xi i = 1,2,...,16

where ß1 and ß2 are true but unknown population parameters.

But the above deterministic PRF cannot be considered an econometric model because it is deterministic or exact, that is it does not have a random error term. To turn this into an econometric model we note that, while on the average Y increases as X increases, this is not necessarily true for every individual family in our example.

However, as the scattergram of income and expenditures shows, consumption expenditures of families that have the same income level are distributed around their group mean value.



Thus, for any given level of income, we can express the deviation of an individual family's consumption from the mean consumption of all families with that income as,

ui = Yi ‑ E(Yi |Xi)

or

Yi = E(Yi |Xi) + ui

(Systematic ) (Non- Systematic)

This is the stochastic PRF.

Here, the term, ui is a random variable that takes on positive, zero, or negative values. This is the random or stochastic error or disturbance term we first saw in Methodology part. It is a proxy for determinants of consumption other than disposable income. It also captures measurement errors, incorrect functional form, and randomness in human response.

Right hand side of the above expression has two components, namely, Systematic and Non- Systematic. The systematic part, E(Yi |Xi), systematically measure the variations in Y because of changes in X. but the systematic part is unable to measure the total variation in Y. Residual amount of change in Y is non-systematic and represented by random error term, ui.

Above stochastic PRF specifies no functional form for the E(Yi |Xi).

Assuming E(Yi | Xi) is linear, the PRF becomes

Yi = ß1 + ß2Xi + ui

which is both linear and stochastic.

Using the stochastic *PRF* and the data from our hypothetical example of weekly consumption expenditures, we can write

*Y1 = 55 = ß1 + ß2(80) + u1*

*Y2 = 60 = ß1 + ß2(80) + u2*

*.......................*

*.......................*

*Y5 = 75 = ß1 + ß2(80) + u5*

From the above discussion of deterministic and stochastic PRF we learn that, the total variations in the dependent variable cannot be explained by the explanatory variables included in the model. Hence, stochastic regression model is more appropriate.

**Significance of Disturbance Term in Econometric Model**

As we noted earlier disturbance term is surrogate or proxy for all variables that are omitted from the model but they collectively affect dependent variable, Y. Then the immediate question

is why not include all such variables into the model or use multiple regression model . The reasons are as follows:

1. Vagueness of theory: The theory, if any, determining the variations of Y may be, and often is, incomplete and also unclear about number of variables influences on Y. We might know for certain that price of certain commodity, X influences its demand Y, but we might be ignorant or unsure about the all other variables affecting Y. Therefore, ui may be used as a substitute or proxy for all the excluded or omitted variables from the model.
2. Unavailability of data: Although we know what some of the omitted variables are and therefore consider them in a regression model, we may not have data about these variables. It is a common problem in empirical analysis. For example, family wealth is a potential explanatory variable in determining consumption expenditure, but unfortunately data on wealth is not available. Though we have omitted wealth from the model, but its influence as taken care by the disturbance term.
3. Core variables versus peripheral variables: Assume that in demand model besides price, income , prices of related commodities, taste and preference, and expectation also affect demand of certain commodity. But it is possible that the influence of some of these variables may be so small and not significant. In such conditions it does not need to introduce them into the model explicitly and their combined effect can be treated as a random variable ui.
4. Intrinsic randomness in human behavior: Even if we succeed in introducing all the relevant variables into the model, there is bound to be some “intrinsic” randomness in individual Y’s that cannot be explained no matter how hard we try. The disturbances, the u’s, may very well reflect this intrinsic randomness.
5. Poor proxy variables: Regression model assumes that the variables Y and X are measured accurately, in practice the data may be plagued by errors of measurement. The error term u may in this case then also represent the errors of measurement.
6. Principle of parsimony: one of the principles of model building is to keep the regression model as simple as possible. If the model is able to estimate the behavior of dependent variable, Y “substantially” with two or three explanatory variables and theory is not strong enough to suggest what other variables might be included, why introduce more variables? Let ui represent all other variables.
7. Wrong functional form: Even if we are able to include all variables explaining a phenomenon and obtain data on these variables, very often we do not know the form of the functional relationship between the dependent and the explanatory variables. If we fit a linear model though the relationship is non-linear and vice-versa we are committed an error. If we introduce the disturbance term to the model it taken care this error also.

For all these reasons, the stochastic disturbances ui assume an extremely critical role in econometric analysis.

**The Sample Regression Function (SRF)**

The discussion so far we have is related to the population of Y values for the fixed X values. But in most practical situations what we have sample values of Y corresponding to some given values of X. That is what face is sampling problems. Because of both theoretical and practical reasons the population is unobservable, hence, we should estimate PRF on the basis of sample information.

To introduce the sampling procedure, let consider our earlier example of weekly consumption expenditure and weekly income of 60 families. Now assume that this information is not known and only information we had is randomly selected sample values of Y for fixed Xs, given in the following table. We have only one Y value for given Xs. Now the question is, can we estimate PRF from this sample data? The answer is we may not be able estimate the PRF accurately due to because of sampling fluctuation, which can be observed one more random sample drawn from 60 families given below.

A random sample 1 from the population

**Y X**

**70 80**

**65 100**

**90 120**

**95 140**

**110 160**

**115 180**

**120 200**

**140 220**

**155 240**

**150 260**

Another random sample2 from the population

**Y X**

**55 80**

**88 100**

**90 120**

**80 140**

**118 160**

**120 180**

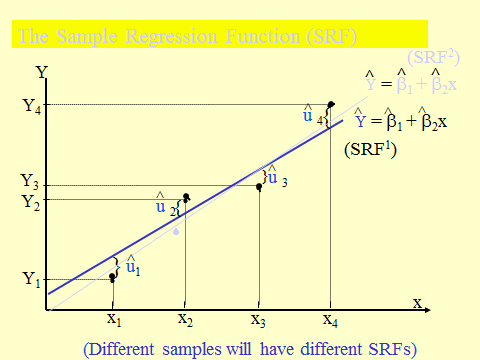
**145 200**

**135 220**

**145 240**

**175 260**

Sample information in the above tables represents the population regression line, but because of sampling fluctuations they are at best an approximation of the true PR. In general, we would get N different SRFs for N different samples, and these SRFs are not likely to be the same.



The sample counterpart of the true but unknown population regression function is called *sample regression function, SRF*.

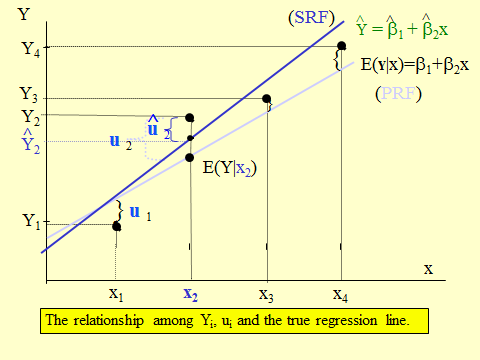
Below is the non-stochastic (deterministic) version of the SRF

*^  ^  ^*

*Yi = ß1 + ß2Xi*

*^* *^ ^*

where *Yi* is an estimator of *E(Yi|Xi)*, and *ß1* and *ß2* are estimators of *ß1* and *ß2*, respectively. Because of *sampling fluctuations*, the above *SRF* is an approximation (estimate) of the true *PRF*.



We can express the above non-stochastic *SRF* into a stochastic *SRF* as follows.

First, we define *sample residuals* as

*^   ^*

*ui = Yi ‑ Yi*

Next, we rewrite this as

*^  ^*

*Yi = Yi + ui*

*^  ^  ^*

Finally, we substitute *ß1 + ß2Xi*for *Yi* in the above,

*^ ^ ^*

*Yi = ß1 + ß2Xi + ui*

**Two Variable Regression- The Problem of Estimation**

As we were noted in the methodology of econometrics, once the econometric model is specified the next task is to estimate the PRF on the basis of SRF as accurately as possible. In the literature there are two techniques are widely used. They are Ordinary Least Squares (OLS) and Maximum Likelihood Estimation (MLE). Out of these two the simple and most popular method is OLS.

**Method of Ordinary Least Squares (OLS)**

The method of ordinary least squares is attributed to Carl Friedrich Gauss, a German mathematician. The method of least squares has some very attractive statistical properties, because of this it is considered as the most powerful and popular methods of regression analysis. To understand this method, we first explain the least-squares principle.

Recall PRF

Yi = ß1 + ß2Xi + ui

Since PRF is not directly observable, we estimate SRF

^ ^  ^

Yi = ß1 + ß2Xi + ui

^ ^

Yi = Yi + ui

But the question is how this SRF is itself determined? For this we need to proceed as follows:

In term of residuals SRF is written as

*^*  ^

ui  = Yi  - Yi

^  ^

= Yi - ß1 + ß2Xi

With the given ‘n’ pairs of observations on X and Y, we would like to determine the SRF in such a manner that it is as close as possible to actual values of Y.

To this end, if we choose sum of the residuals as small as possible, i.e,

*^*

∑ui = ∑(*Yi – Yi )*

Though it is intuitively appealing but it is not a good criterion. Because this principle assign equal weight to all uis.

We can avoid this problem if we adopt the least-squares criterion, which states that the SRF can be fixed in such a way that

^   ^

∑ ui2 = ∑ (Yi - Yi)2

^ ^

= ∑ (Yi - ß1 ‑ ß2Xi)2

^ ^

It is based on the least-squares principle, which suggests choosing *ß1* and *ß2* such that, for a given sample,∑ ui2 , the Sum of Squared Residuals (*RSS*) is the smallest, i.e., minimize.

For a given sample, the method of least squares provides us with unique estimates of β1 and β2 that give the smallest possible value of ∑uˆ2 i. This can be a straightforward exercise in differential calculus. The differential calculus yields the following normal equations:

∑Yi = nβˆ 1 + βˆ 2 Xi

∑Yi Xi = βˆ 1 Xi + βˆ 2 X2

Solve these two equations, and you will obtain the OLS estimators,

*\_ \_ \_*

*^ ∑ [(Xi ‑X)(Yi ‑Y)]*

*ß2=*

*∑(Xi ‑X)2*

and

*^ \_ ^  \_*

*ß1 = Y ‑ ß2 X*

*\_ \_*

where *Y* and *X* are the sample means of *Y* and *X*, respectively.

The estimators are known as the least-squares estimators, for they are derived from the least-squares principle.

**Numerical Features of OLS Estimators & SRF**

1. The OLS estimators depend on observable sample quantities *Xi*and *Yi*.

2. They are linear functions of *Yi*.

3. They are point estimators.

4. The two OLS estimators of *ß1* and *ß2* yield a linear sample regression line whose RSS

is smaller than any other straight line.

5. The OLS sample regression line passes through the sample means of *X* and *Y*.

*^*

6. The mean value of the estimated *Y* is equal to the mean value of actual *Y*

7. The mean value of sample residuals, u i, is zero.

8. Sample residuals, ui and *Xi* are uncorrelated.

9. Sample residuals and predicted *Yi* are uncorrelated.

**The Classical Linear Regression Model: The Assumptions Underlying the Method of Least Squares**

In regression analysis our objective is not only to obtain βˆ1 and βˆ2 but also to draw inferences about the true β1 and β2. For example, we would like to know how close βˆ1 and βˆ2 are to their counterparts in the population or how close Yˆi is to the true E(Y | Xi).

Look at the PRF: Yi = β1 + β2Xi + ui . It shows that Yi depends on both Xi and ui . The assumptions made about the Xi variable(s) and the error term are extremely critical to the valid interpretation of the regression estimates.

The Gaussian, standard, or classical linear regression model (CLRM), makes 10 assumptions.

ASSUMPTION 1

Linear Regression Model: The regression model is linear in the parameters, though it may or may not be linear in the variables.

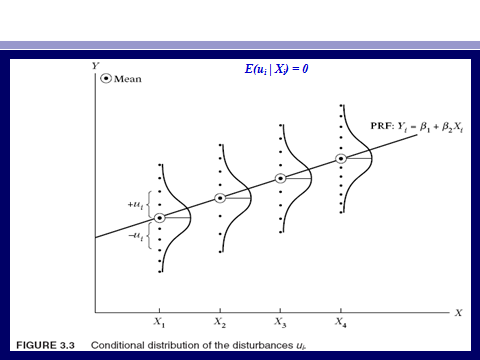
That is the regression model is Yi = β1 + β2 Xi + ui

ASSUMPTION 2 : Fixed X Values or X Values Independent of the Error Term:

Values taken by the regressor X may be considered fixed in repeated samples. This means that our regression analysis is *conditional regression analysis*, that is, conditional on the given values of the regressor (s) *X.* Keeping the value of the independent variable *X fixed,* we draw at random a value for Y and *we can repeat this* process for all the *X values.*

ASSUMPTION 3 Zero Mean Value of Disturbance ui:

Given the value of Xi, the mean, or expected, value of the random disturbance term ui is zero. Symbolically, we have E(ui|Xi) = 0. Each Y population corresponding to a given X is distributed around its mean value with some Y values above the mean and some below it. The mean value of these deviations corresponding to any given X should be zero which can be observed for the following figure.



ASSUMPTION 4 Homoscedasticity or Constant Variance of ui:

The variance of the error, or disturbance, term is the same regardless of the value of X. Symbolically, var (ui) = E[ui − E(ui|Xi)]2

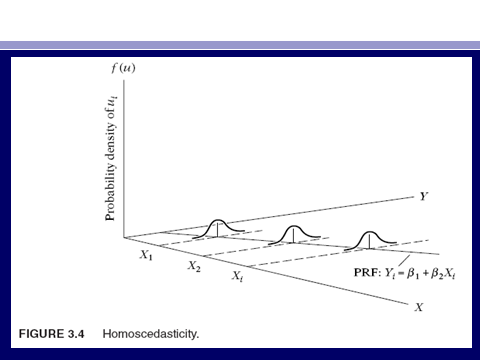
= E(u2 i |Xi)

= σ2

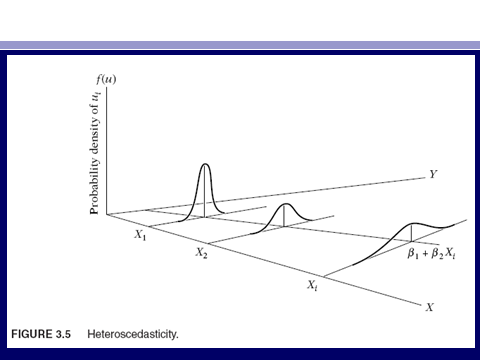
where var stands for variance

homo means equal and scedasticity means spread, that is equal spread. Stated differently, it means that the Y populations corresponding to various X values have the same variance.

Put simply, the variation around the regression line (which is the line of average relationship between Y and X) is the same across the X values; it neither increases or decreases as X varies. Which can be observed from the following figure:



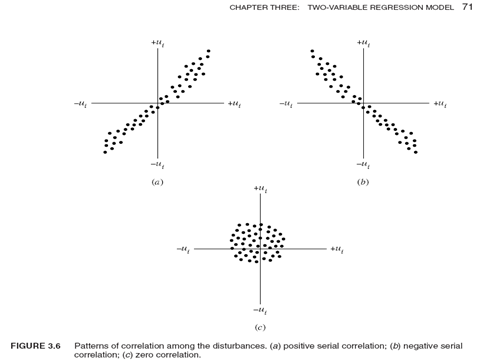
where the conditional variance of the Y population varies with X. This situation is known as heteroscedasticity, or unequal spread, or variance. Symbolically, in this situation can be written as var (ui | Xi) = σ2i and shown in the figure given below:

**

ASSUMPTION 5 No Autocorrelation between the Disturbances:

Given any two X values, Xi and Xj (i ≠ j), the correlation between any two ui and uj(i ≠ j) is zero. In short, the observations are sampled independently. Symbolically, cov(ui, uj|Xi, Xj) = 0 cov(ui, uj) = 0, where i and j are two different observations and where cov means covariance.

The disturbances ui and uj are uncorrelated, i.e., no serial correlation or no autocorrelation. This means that, given Xi , the deviations of any two Y values from their mean value do not exhibit patterns. In the following figure, (a) and (b) represent a systematic pattern with positive and negative autocorrelation respectively whereas (c) shows no autocorrelation.\



ASSUMPTION 6 : Zero correlation between Xi and Ui.

The disturbance u and explanatory variable X are uncorrelated. The PRF assumes that X and u (which may represent the influence of all the omitted variables) have separate (and additive) influence on Y. But if X and u are correlated, it is not possible to assess their individual effects on Y. Thus, if X and u are positively correlated, X increases when u increases and it decreases when u decreases. Similarly, if X and u are negatively correlated, X increases when u decreases and it decreases when u increases. In either case, it is difficult to isolate the influence of X and u on Y.

ASSUMPTION 7 The Number of observations n must be greater than the number of parameters to be estimated: Alternatively, the number of observations must be greater than the number of explanatory variables. For example I f we have only single observation of Y and X, there is no way to estimate the two unknowns, β1 and β2. We need at least two pairs of observations to estimate the two unknowns.

ASSUMPTION 8: Variability in X values. The X values in a given sample must not all be the same and technically the variance of X should be finite positive value. If all the X values are identical, then Xi = X¯ and the denominator of that equation will be zero, making it impossible to estimate β2 and therefore β1.

ASSUMPTION 9 Regression model is correctly Specified

An econometric investigation begins with the specification of the econometric model underlying the phenomenon of interest. Some important questions that arise in the specification of the model include the following:

(1) What variables should be included in the model?

(2) What is the functional form of the model? Is it linear in the parameters, the variables, or both?

(3) What are the probabilistic assumptions made about the Yi , the Xi, and the ui entering the model?

A model should include all relevant variable, correct functional form and appropriate assumption related to Y, X and u.

ASSUMPTION 10: There is no perfect multicollinearity.

Perfect multicollinearity means existence of an exact linear relationship between two or more independent variables. When there is perfect multicollinearity, regression coefficients cannot be estimated as they will be of the indeterminate form, 0/0. Further, standard errors of the estimated coefficients tends to infinity as they will equal σ2/0, implying complete lack of precision. Hence regression analysis assumes no perfect multicollinearity among explanatory variables.

ASSUMPTION 11:The population error term, ui, follows the normal distribution

As far as estimation is concerned, all we need to generate “good" estimates and we can obtain from assumptions *1-10*.

But in order to draw statistical inference (i.e., test hypotheses,) we also need to make an assumption regarding the probability distribution of the population error term, *ui*.

Any regression model that satisfies assumptions 1-10 is known as a Classical Linear Regression Model (CLRM), which is suitable for estimation alone. Any regression model that satisfies assumptions 1-11 is known as the Classical Normal Linear Regression Model (CNLR) , suitable not only for estimation but also for testing hypotheses.

**Precision of Ordinary Least Squares Estimators**

OLS estimates are point estimators which are considered as random variables as they change from sample to sample. Thus, we must assess their precision or reliability. In statistics, the precision of estimates is measured by their standard errors. All else the same, the smaller the standard error of a point estimate, the more precise that estimate is.

The least-squares estimates are a function of the sample data. But since the data change from sample to sample, the estimates will change. Therefore, what is needed is some measure of “reliability” or precision of the estimators βˆ1 and βˆ2.

Which can be obtained by using following formula:

σ2 is the constant or homoscedastic variance of ui of Assumption 4.

σ2 itself is estimated by the following formula:

ˆσ2 = / n-2

where ˆσ2 is the OLS estimator of the true but unknown σ2 and where the expression n−2 is known as the number of degrees of freedom (df)  is the residual sum of squares (RSS). Once  is known, ˆσ2 can be easily computed. The term number of degrees of freedom means the total number of observations in the sample (= n) less the number of independent (linear) constraints or restrictions put on them. In other words, it is the number of independent observations out of a total of n observations.

**Properties of OLS estimator: Gauss - Markov Theorem**

Given the assumptions of the classical linear model, OLS estimators possess the some important ideal statistical properties. These properties are contained in the famous Gauss – Markov theorem. Based on this theorem the OLS estimator are said to be Best Linear Unbiased estimators (BLUE). This property holds good if the following conditions are satisfied by OLS estimators.

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

^  ^  ^

Yi = ß1 + ß2Xi

1. It is unbiased, that is, its average or expected value, E(βˆ 2), is equal to the true value, β2.

*^*

An estimator, β , is said to be an unbiased estimator of the population parameter β if, in repeated sampling, its expected or mean value equals the true parameter, β ,

^

E(β ) = β

Or ^ ^

Bias(β ) = E(β ) - β = 0

Note that unbiasedness is a property of repeated sampling, not of a single sample.

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

*^*

An estimator, β , is said to be the best linear unbiased estimator (BLUE) of the true population parameter, β , if it is...

* + linear
  + unbiased
  + minimum variance in the class of all linear unbiased estimators of β.

Thus BLUE is equivalent to linear and efficient.

**The Coefficient of Determination r2: A Measure of “Goodness of Fit”**

We need criteria for assessing the overall fit of the sample regression line or how well the sample regression fits data well. In practice, because of sampling fluctuation, it is impossible to obtain a perfect fit. In general, there will be some positive ûi and some negative ones. How good the sample regression line fits the sample data depends on how small the individual ûi are. The simple sample coefficient of determination, r2, is a measure that summarizes this information. Hence, r2 is a summary measure, that tells how well the sample regression line fits the data.

We calculate r2 as follows. Recall that,

Yi = *ß1 + ß2Xi* + ui PRF

*^  ^*

*Yi = ß1 + ß2Xi* SRF

^ ^

Yi = Yi + ui

Write this in deviation form,

^ ^

yi = yi + ui

Square both sides and sum over all observations, ^ ^ ^ ^

∑y2 = ∑y2 + ∑u2 + 2∑yiui

^ ^

∑y2 = ∑y2 + ∑ui2

The above equality consists of three sums of squares, or variations:

1. ∑y2 which represents total variation of actual Y values about their sample means. We call this *total sum of squares* (TSS)

^

2. ∑y2, which represents variation of estimated Y values about their sample mean, known

as *explained sum of squares* (ESS)

3. ∑ui2 representing residual or unexplained variation in Y values about the regression line, the *residual sum of squares* (RSS)

In short,

^

∑yi2 = ∑yi2 + ∑ui2

can be written

TSS = ESS + RSS

Divide both sides of the last equality by TSS,

1 = (ESS/TSS) + (RSS/TSS)

^ ^

1 = ∑yi2/∑yi2 + ∑ui2/∑yi2

Therefore, we can define goodness of fit as follows

^

r2 = ESS/TSS = ∑yi2/∑yi2

r2 measures the proportion (percentage) of total variation in Y that is explained by the regression model as a whole.

It has the following properties:

* + It is bounded between zero and one.
  + In the bivariate regression model, the square root of r2 is equal to simple sample correlation coefficient between Y and X,

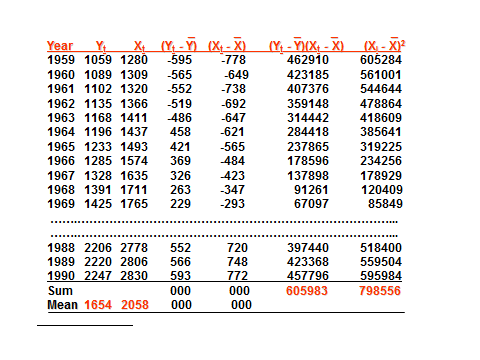
rXY = √r2 = R

*AN EXAMPLE :*Estimating Simple Consumption Function for the U.S. - 1959‑1990

The standard Keynesian consumption function expresses real consumption expenditures, Y, as a linear function of real disposable income, X,

Yt = β1 + β2Xt + ut

Here we use seasonally adjusted, personal expenditures on nondurable goods and services in billions of 1987 dollars to quantify the dependent variable. And Xt, seasonally adjusted personal disposable income in billions of 1987 dollars is used.



^ \_ \_ \_

β2 = Σ(Xi - X)(Yi - Y)/Σ(Xi - X)2

= 605983/798556 = 0.7588

^ \_ ^ \_

β1 = Y - β2X = 1654 - 0.7588(2058) = 92.3896

Thus, the estimated model can be written in either of the following two alternative forms,

^ ^

Yt = 92 + 0.76Xt or Yt = 92 + 0.76Xt + ut

We interpret the estimation results as follows:

According to the estimated value of β2 (i.e., 0.76), if the real personal disposable income increases by one dollar, on average real consumption expenditures are expected to increase by $0.76.

According to the estimated value of the intercept term, the mean systematic effect of other determinants of consumption expenditures not explicitly included in the model is $92 billion.

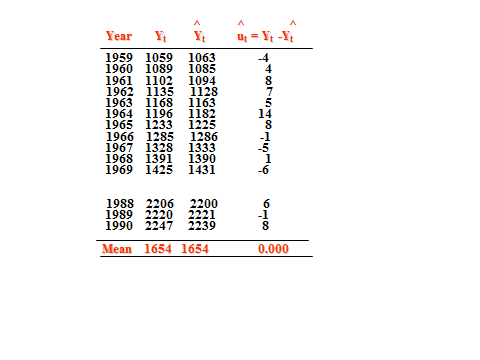
*^*

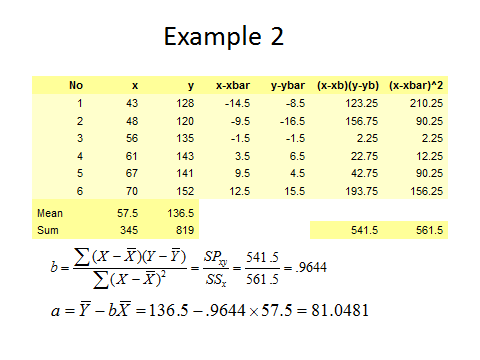
We calculate the predicted values of the dependent variable, *Yt*, by plugging values of *Xt* into the first of the above two equations.

Then by subtracting the predicted values of *Yt* from its observed values, we can calculate

^

the sample residuals, *ut*,





**Two-Variable Regression: Hypothesis Testing**

Estimation and hypothesis testing are twin branches of statistical inference. So far in the last section we discussed about estimation where we obtained numerical values for unknown parameters of the regression model by using least squares method. Now let us understand the procedure of hypothesis testing in regression context.

The very simplest sort of hypothesis test concerns the population mean from which a random sample has been drawn. In regression analysis the hypothesis testing relates to reliability of sample estimator,ß^. In other words, How “close” is ß1^ to β1? And How “close” is ß2^ to β2 ? This is because as we noted in the earlier OLS estimator are random variables as they change from sample to sample. Hence we should compare and relate the estimated sample values (ß^s) to the true population parameters (βs).

There are three approaches to hypothesis testing in the literature. They are confidence interval approach, test of significance approach and P value approach. Let us discuss these methods one by one.

1. **Confidence Approach to Hypothesis Testing**

Based on the normality distribution for random disturbance term, ui, the OLS estimators, β^1 and β^2  also assumed be to follow the normal distribution with following properties.

β^1 is normally distributed ~ N(β1, σ^β12)

And Z = (β^1- β1)/ σ^β1  is ~ N(0,1)

β^2 is normally distributed ~N(β2 ,σ^β22)

And Z = (β^2- β2)/ σ^β2 is ~ N(0,1)

(n-2) σ^2/ σ2 is distributed as the χ2(n-2)

By using the Z value when σ2  known or t value when σ2  is not known, wecan construct the confidence interval as follows for β2 ;

Pr( β2-δ < β2 < β2+δ ) = (1-α)

β2 - δ is called lower confidence bound

β2 + δ is called upper confidence bound

the interval between (β2 - δ) and (β2 + δ) is called random interval (confidence interval)

(1-α) is confidence coefficient: (0< α <1)

α is called the level of significance

The above confidence interval provides a 100(1-α) percent confidence interval for β2.

1. **Test of Significant Approach**

An alternative to the confidence-interval method is the test-of-significance approach. It is a procedure by which sample results are used to verify the truth or falsity of a null hypothesis. The decision to accept or reject null hypothesis, H0 is made on the basis of the value of the test statistic obtained from the data at hand.

Test statistic is

ˆ







t =

ˆ



Se

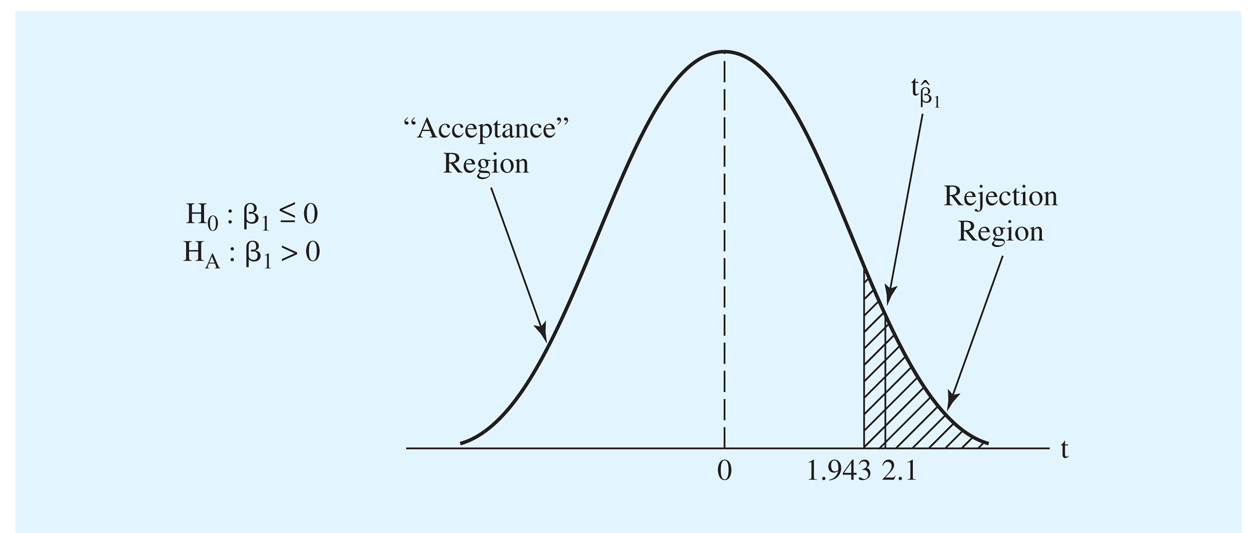
To use the above test statistic the following steps has to be follow;

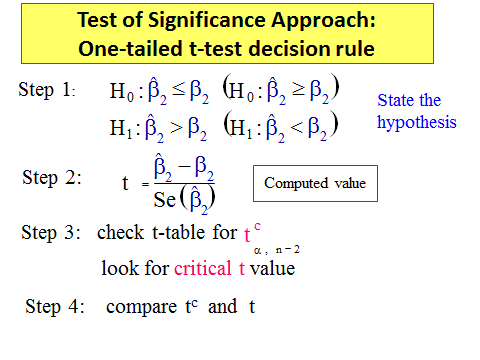
1. Set up the null and alternative hypothesis

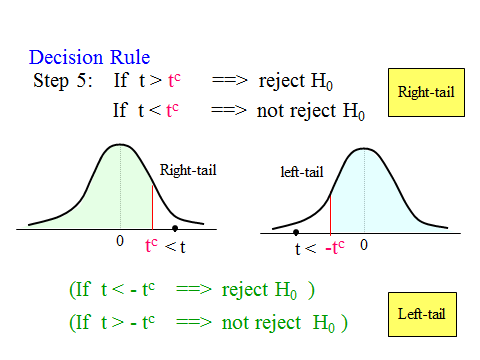
2. Choose a level of significance and therefore a critical t-value

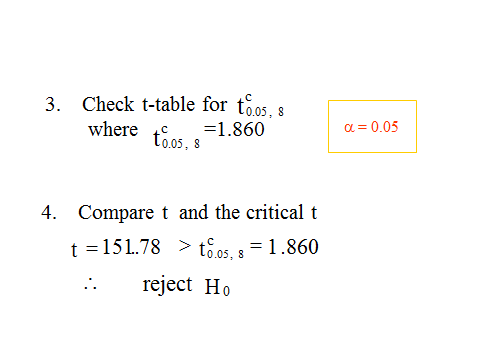
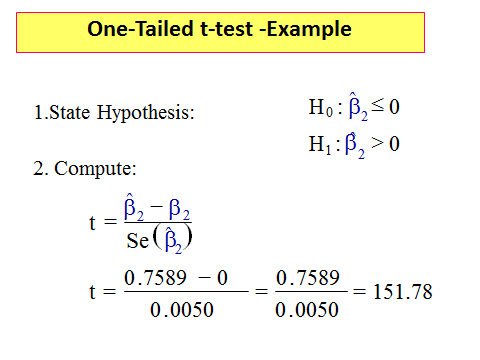
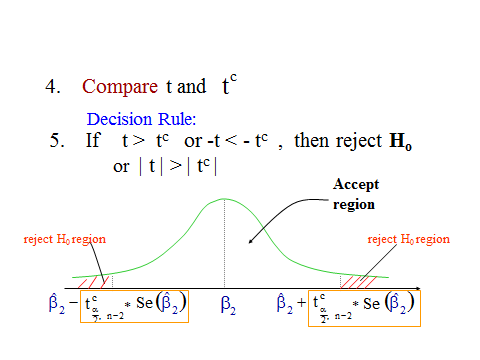
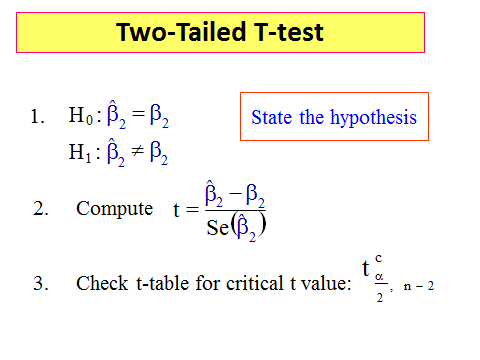
3. Run the regression and obtain an estimated t-value (or t-score)

4. Apply the decision rule by comparing calculated t-value with the  
critical t-value or table value in order to reject or not reject the null hypothesis. It is Reject H0 if |tk| > tc









1. **The Exact Level of Significance: The p Value Approach**

Once a test statistic is obtained in a given example, why not simply go to the appropriate statistical table and find out the actual probability of obtaining a value of the test statistic as much as or greater than that obtained in the example? This probability is called the p value (i.e., probability value), also known as the observed or exact level of significance or the exact probability of committing a Type I error.

More technically, the p value is defined as the lowest significance level at which a null hypothesis can be rejected.

The decision rule for p value approach is as follows:

If p. value ≤ 0.05 reject Ho otherwise do not reject

**MODULE III MULTIPLE REGRESSION MODEL**

The two variable or bivariate regression model discussed in the previous module is not often adequate in the practical applications. Because a dependent variable not just influenced by a single explanatory variable. It is influenced by several variables in practice. In other words, in practice, what we usually need is a regression model with more than just one explanatory variable. For example, the consumption expenditure not only depends on income, besides income factors like, wealth, number of persons in the family, age of the persons, religion, etc. influence on consumption expenditure. As another example, the production of a commodity depends on units of labour, amount of capital, technology, amount of raw material used, etc. Therefore the bivariate model is not adequate to explain the real world situations, hence we need to extend or modify the two variable model to cover all the variables influences on a chosen dependent variable. This is known as a multiple regression model. That is, the dependent variable, Y depends on two or more explanatory variables. Therefore a multiple regression model can be define as a model with one dependent variable and two or more independent or explanatory variable. Fortunately, with only a few exceptions, the concepts we studied in the context of the simple regression model carry over to multiple regression analysis. The simplest form of multiple regression model is tri-variate or three variable regression model. In which, one is dependent and the two are being explanatory variables.

**Three Variable Regression Model**

The general form a multiple regression model can be expressed as

Yi = ß1 + ß2X2i + ... + ßkXki + ui

This model has k unknown parameters (k-1 slope coefficients and one intercept term). Thus we may refer to this as the k-variate regression model.

This simplest multiple regression model incorporates all of the concepts underlying the more general k-variate model.

The tri-variate PRF may be written as,

Yi = ß1 + ß2X2i + ß3X3i + ui

The corresponding SRF is written as either,

^ ^ ^   ^ ^ ^ ^ ^

Yi=ß1+ß2X2i+ß3X3i+ui or Yi=ß1+ß2X2i+ß3X3i

**Interpretation of Multiple Regression Equation**

Given the assumptions of the classical regression model and taking the conditional expectation of Y on both sides of tri-variate PRF, we obtain

E(Yi | X2i, X3i) = β1 + β2X2i + β3i X3i

In words, the above equation gives the conditional mean or expected value of Y conditional upon the given or fixed values of X2 and X3, same as in two-variable regression. Recall that we referred to the parameters of the simple regression model as the regression coefficients. In that model, the (only) slope coefficient, β2, represented the effect of a one-unit change in (the only) independent variable on the dependent variable.

In multiple regression we distinguish between joint and partial (individual) effects of the explanatory variables. Each partial or individual regression coefficients in a multiple regression model represents the effect of a one unit change in the corresponding X variable on the dependent variable, holding all other independent variables constant.

These are called partial regression coefficients because they are similar to partial derivatives. For example, in the tri-variate model,

ß2 = ∂Y/∂X2 and ß3 = ∂Y/∂X3

Thus ß2 measures change in Y per unit change in X2, holding X3 constant. In other words, ß2 measures the change in the mean values of Y, per unit change in X2, holding X3 constant or the ‘direct’ or ‘net’ effect of a unit change in X2 on the mean value of Y

ß3 measures change in Y per unit change in X3, holding X2 constant. It means ß3 measures the change in the mean values of Y, per unit change in X3, holding X2 constant or the ‘direct’ or ‘net’ effect of a unit change in X2 on the mean value of Y.

Holding constant indicates the assessment of the true contribution of X2 to the change in Y, we control the influence of X3 and vice versa.

Sometimes we are interested in the joint effect of simultaneous variation in X2 and X3 on the dependent variable. That is we are interested in β2 and β3 together. This effect is referred to as the joint effect of X2 and X3, as opposed to their individual or separate (i.e., partial) effects.

**Derivation of OLS Estimation of Trivariate Regression Model**

To find the OLS estimators, we need to write the SRF of trivariate PRF, as follows

^ ^ ^   ^

Yi=ß1+ß2X2i+ß3X3i+ui

As we are noted in two-variable regression model, OLS procedure consists of deriving the values unknown population parameters so that the residual sum of squares (RSS) as small as possible

.

OLS is to minimize the RSS (Σ u2)

^

min. RSS = min. Σu**^**2 = min. Σ(Y - β1 - β2X2 - β3X3)2

**^**

**^**

**^**

^

δRSS

δβ1

=2 Σ ( Y - β1- β2X2 - β3X3)(-1) = 0

**^**

**^**

**^**

**^**

δRSS

δβ2

=2 Σ ( Y - β1- β2X2 - β3X3)(-X2) = 0

**^**

**^**

**^**

**^**

δRSS

δβ3

=2 Σ ( Y - β1- β2X2 - β3X3)(-X3) = 0

**^**

**^**

**^**

^

From the above differentiated equations the following normal equation are derived:

nβ1 + β2 ∑X2  +∑X3 β3 = ∑Y

β 1∑X2  + β2 ∑X2 2 + β3 ∑X2  X3  = ∑X2 Y

β 1∑X3  + β2 ∑X2  X3  + β3 ∑X2 3 + = ∑X3 Y

Rewrite in matrix form:

n ΣX2 ΣX3

ΣX2 ΣX22ΣX2X3

ΣX3 ΣX2X3 ΣX32

β1

β2

β3

**^**

**^**

**]**

**=**

ΣY

Σ X2Y

ΣX3Y

**(X1X)**

β

**^**

**= X1Y**

**Matrix notation**

To estimate β2  and β3 the following expression has been derived from using Cramers’s rule.

**^** ( ∑yx2) (∑x32) - (∑yx3)  (∑x2x3)

β2 =

(Σ*x*22)(Σ*x*32) - (Σ*x*2*x*3)2

β3 = ( ∑yx3) (∑x22) - (∑yx2)  (∑x2x3)

(Σ*x*22)(Σ*x*32) - (Σ*x*2*x*3)2

β1 = Y - β2X2 - β3X3

**^**

**^**

**^**

**\_**

**\_**

**\_**

**The Multiple Coefficient of Determination R2**

The notation of r 2 used in two variable regression model can be easily extended to regression models containing more than two variables. Thus, in the three-variable model we would like to know the proportion of the variation in Y explained by the variables X2 and X3 jointly in three variable model. The quantity that gives this information is known as the multiple coefficient of determination and is denoted by R2; conceptually it is akin to r 2.

R2 and the Adjusted R2

The multiple coefficient of determination, *R2*, does not decrease, in fact it often increases as the number of independent variables increases. Thus *R2* can be misleading in multiple regression. To overcome this problem, we adjust R2 for the number of X variables in the model. This is done by taking into account the degrees of freedom of the sums of squares that enter the formula for *R2*.

In the k-variate model, adjusted-R2 is defined as,

\_

R2 = 1 ‑ [(RSS/n-k)/(TSS/n-1)]

= 1 ‑ [(n - 1)/(n - k)](1-R2)

Adjusted-R2 has the following properties:

\_

R2 > R2

\_

R2 ≤ 1

When working with multiple regression model, the appropriate coefficient of determination to use is the adjusted-R2

**Estimation in Multiple Regression Model: An Example**

Let us consider a trivariate regression model where the dependent variable is consumption expenditure and the explanatory variables are personal disposable income and rate of interest. In this case we simply extended a simple regression model of consisting consumption expenditure and disposable income by adding rate of interest.

Thus we have,

Yt = β1 + β2X2t + β3X3t + ut

where, Yt is consumption expenditures, X2t is personal disposable income, and X3t is a short-term interest rate, The results are shown in the following table taken from a statistical package Eviews.

Variable Coefficient Std. Error t-Statistic Prob.

==============================================================

C 101.887803 8.144611 12.50991 0.0000

INC 0.876310 0.007902 110.92411 0.0000

INT ‑3.298261 0.196013 ‑16.82814 0.0000

==============================================================

R‑squared 0.999997 Mean of dependent var 1653.813

Adjusted R‑squared 0.999988 S.D. of dependent var 385.403

S.E. of regression 12.167418 Sum of squared resid 4441.368

Log likelihood ‑143.512300 F‑statistic 299999.316

Durbin‑Watson stat 1.083619 Prob(F‑statistic) 0.000

According to the estimated coefficient on X2, a one dollar increase (decrease) in real personal disposable income is expected to increase (decrease) real consumption spending by $0.88 on average, holding the rate of interest constant.

According to the estimated coefficient on X3, a one percentage point increase (decrease) in the short‑term interest rate is expected to decrease (increase) real consumption spending by 3.3 billion dollars, holding real disposable income constant.

According to the estimated value of the intercept term, the mean systematic effect of

other determinants of consumption not explicitly included in the model is about $102

billion.

**Hypothesis Testing**

In multiple regression model, we can test several interesting hypotheses along with testing the significance of partial regression coefficients. Such as the following:

1. Testing individual partial regression coefficients
2. Testing overall significance of all coefficients
3. Testing the restriction on the variables
4. Testing coefficients equality
5. Testing the functional form of the model
6. Testing stability of the estimated regression model over time and across cross sections.

Here we discuss the procedure of the first two types of hypothesis testing.

1. **Testing Individual Partial Regression Coefficients**

To test the individual regression coefficients of a regression model, we can use the same methodology we followed in two variable model. That is the ‘t’ test can be applied for this. Let us use the results our trivariate model of consumption. Required elements to use the ‘t’ test are the estimated values of parameters, namely, ß1^ and ß2^ and their standard errors. The values of ß1^ and ß2^ 0.88 and – 3.30 respectively. The values of standard errors are 0.008 and 0.196 for ß1^ and ß2^. Then the testing procedure is as follows:

1. Holding X3 constant: Whether X2 has the effect on Y ?

Null hypothesis : H0 : β2 = 0

Alternative hypothesis : H1 : β2 ≠ 0

Which means the H0  indicates X2 has noeffect on the dependent variable , Y, where the alternative states, X2 has exert some effect on the Y.

Substituting the above information to ‘t’ test

β /

se(

0.88 - 0

0.008

t= 110

Compare with the critical value tc0.025, 12 = 2.179

Since the calculated value 110 is greater to the critical or table value, the decision is the rejection of null hypothesis and accept the alternative hypothesis. Which implies that X2, personal disposal income has some influence on the dependent variable consumption expenditure and it is statistically significant at 5 per cent level and different from zero.

1. Holding X2 constant: Whether X3 has the effect on Y ?

Null hypothesis : H0 : β3 = 0

Alternative hypothesis : H1 : β3 ≠ 0

Which means the H0  indicates X3 has noeffect on the dependent variable , Y, where the alternative states, X3 has exert some effect on the Y.

Substituting the above information to ‘t’ test

β /

se(

-3.30 - 0

0.196

t= 16.84

Compare with the critical value tc0.025, 12 = 2.179

Reject the null hypothesis because the calculated value is greater to the critical or table value and accept the alternative hypothesis. Which means that X2, personal disposal income has some influence on the dependent variable consumption expenditure and it is statistically significant at 5 per cent level and different from zero.

1. **Testing Overall Significance of All Coefficients**

For the three variable regression model

Y = β1 + β2X2 + β3X3 + u

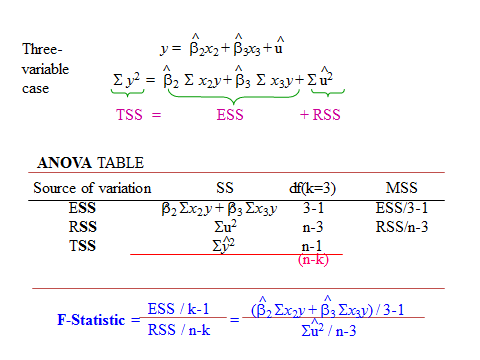
Null hypothesis : H0 : β2 = 0, β3 = 0, or H0 : β2 = β3 = 0, (all variable are zero effect)

Alternative Hypothesis H1 : β2 ≠ 0 or β3 ≠ 0 (At least one variable is not zero)

To test the above null hypothesis, we cannot continue with ‘t’ test due to its limitations in the context of overall significance. Hence the suitable test statistic is ‘F’. The steps involved are as follows:

1. Compute and obtain F statistics
2. Check for the critical Fc value (F c α, k-1, n-k)
3. Compare F and Fc, and if F > Fc ==> reject H0

Actual procedure used to test the above null hypothesis is known in the literature as Analysis of Variance (ANOVA) which make uses the Concepts we already notes in derivation of coefficient of determination, r2, namely, TSS, ESS and RSS. The following table exhibits the same



F =

MSS of **ESS**

MSS of **RSS**

=

ESS / k-1

RSS / n-k

=

Σ *y*2/(k-1)

Σ u 2 /(n-k)

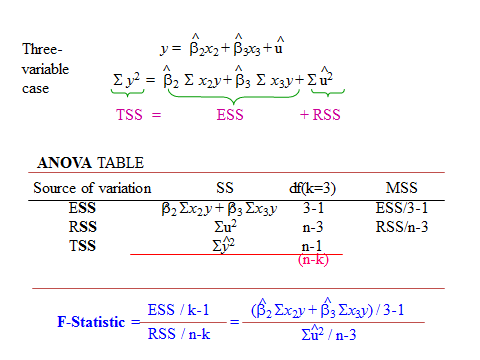
**^**

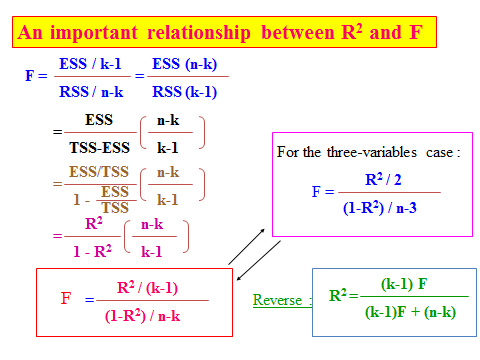
**^**

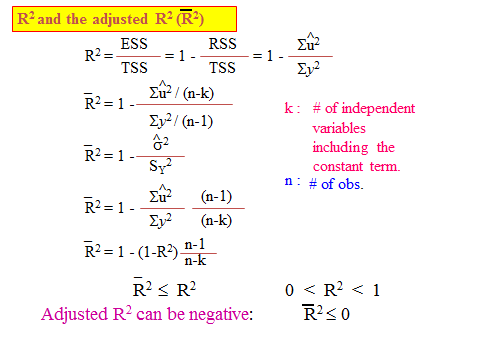
if F > Fck-1,n-k ==> reject Ho

H1 : β2 ≠ … ≠ βk ≠ 0

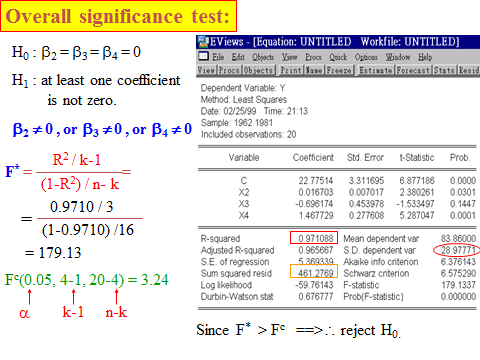
H0 : β2 = … = βk = 0

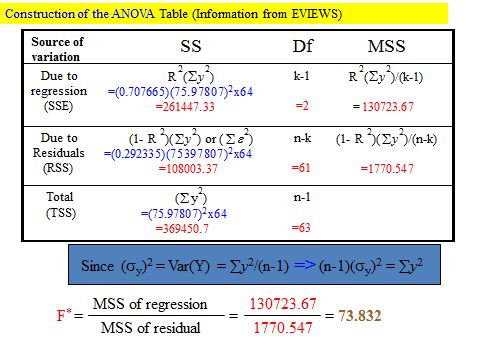






Following procedure uses the estimated value of a multiple regression to test overall significance test





Y = β1 + β2 X2 + β3 X3  + u

H0 : β2 = 0, β3= 0,

H1 : β2 ≠ 0 ; β3 ≠ 0

Compare F\* and Fc, checks the F-table:

Fc0.01, 2, 61 = 4.98

Fc0.05, 2, 61 = 3.15

Decision Rule:

Since F\*= .73.832> Fc = 4.98 (3.15) ==> reject Ho

The overall estimators are statistically significant different from zero

**Module IV Practical Problems of Regression**

In second and third modules we considered two problems, namely estimation and hypothesis testing of simple and multiple regression models. But these two problems depend on the several assumptions which we were noted in the second module. These assumptions are essential for both estimation of unknown parameters and testing hypothesis of regression model. But due to violation of these assumptions we will face several problems. Violation few assumptions have severe implications on the regression model. The three most important problems encountered by econometrician due violation of assumptions in practice are Multicollinearity, heteroscedasticity, and autocorrelation or serial correlation. In this module we will discuss what do we mean by these problems, what are their practical consequences, how does one detects them and what remedial measures can be used to remedied the problems.

**I Multicollinearity Problem**

The term multicollinearity is due to Ragnar Frisch. Multicollinearity means existence of perfect or exact linear relationship between few or all explanatory variables of a regression model. For example, the following tri-variate regression model ,

Y = β1 + β2X2 + β3X3 + u

would suffer from perfect multicollinearity if

X3 = α1+ α2X2

Note that there is no random error term in the above expression, which means |r23| = 1. It means that the correlation X2 and X3 perfect and equal to one represent perfect multicollinearity.

Multicollinearity is a sample problem. This is because independent variables in a regression equation are assumed to be non-stochastic so that population covariance between them is zero by definition. But in a sample of a given population, the independent variables may be correlated.

In the case of imperfect or near or high multicollinearity, r23 is large but not equal to one in absolute value. This would be the case if in the following regression equation,

Y = β1 + β2X2 + β3X3 + u

we had that

X3 = α1 + α2X2 + v

where v is a random error term. Because of the presence of random error term the linear relationship between X2 and X3 is not perfect or imperfect.

Further, the problem of multicollinearity as based on its definition it refers to only linear relationship of explanatory variables.

There are several reasons for arise of multicollinearity in regression model. They are as follows:

1. The data collection model employed
2. Constraints on the model or in the population being sampled.
3. Model specification
4. Overdetermined model
5. Common trend among chosen explanatory variables.

In the case of perfect multicollinearity regression coefficients remain indeterminate and their standard errors are infinite. This is because of the definition of partial regression coefficients. They give the rate of change in the average value of *Y* as one of the explanatory variable for example, *X*2 changes by a unit, holding *X*3 constant. But if *X*3 and *X*2 are perfectly collinear, there is no way *X*3 can be kept constant: As *X*2 changes, so does *X*3 by the factor *λ*. What it means, then, is that there is no way of disentangling the separate influences of *X*2 and *X*3 from the given sample.

But the perfect multicollinearity situation is very exceptional in practice, what we have is imperfect multicollinearity. We are able to determine or estimate the regression coefficients and their standard errors under the presence of imperfect multicollinearity problem. And they are best linear unbiased estimator. Though they are BLUE we will face the following problems.

**Consequences of Near or High Multicollinearity**

1. Standard errors of estimates are too large
2. This makes *t* ratios too small
3. And this increases probability of Type II Error
4. While *R2* and *adjusted R2* are not affected, we may encounter a situation where these are high and significant but none or only a few of the estimated regression coefficients are individually significant.
5. The estimated coefficients may have unexpected or "wrong" signs.
6. Estimates and their standard errors will be sensitive to changes in the sample or model specification.

**Detection of Multicollinearity**

Having studied the nature and consequences of multicollinearity, the question arises in our mind is how does one know that collinearity is present in any given situation, especially in models involving more than two explanatory variables. Before knowing the methods to detect it, it is useful to bear in mind warning given by Kmenta.

1. Multicollinearity is a question of degree and not of kind. The meaningful distinction is not between the presence and the absence of multicollinearity, but between its various degrees.

2. Multicollinearity is a feature of the sample and not of the population. Therefore, we do not “test for multicollinearity” but we measure its degree in any particular sample.

Since multicollinearity is a sample problem, we do not have any unique method to detect it and what we some are rules of thumbs. At best, these guidelines may lead us to “suspect” high multicollinearity. Nor do they tell us much regarding the severity of the consequences of high multicollinearity.

They are as follows:

1. High R2with few significant t ratios

R2, adjusted R2, F statistic, and simple correlation coefficients between the dependent variable and each individual independent variable are high but none or only a few of the estimated coefficients are individually significant.

1. High pair-wise correlations among regressors or explanatory variables.

Second rule of thumb is if pair-wise or zero order correlation among explanatory variable is high, say 0.80, then we can suspect multicollinearity is severe. But is it not true in all the cases. In some situations even if the pair-wise correlation is say 0.50 we expect multicollinearity problem. Hence this method is only a sufficient not necessary.

1. Auxiliary regressions.

Since multicollinearity arises because the linear combinations one or more of the explanatory variables with other explanatory variables, one way to find multicollinearity is run a regression for each explanatory variable, X on the remaining regressors and compute the corresponding R2 , which we designate as R2i ; each one of these regressions is called an auxiliary regression, auxiliary to the main regression of Y on the X’s. Then we can use F test and if this value exceeds the critical value, the conclusion is the presence of severe multicollinearity.

1. Eigenvalues and condition index

Next rule of thumb is the use of eigenvalues and condition index to detect multicollinearity. We can eigenvalues to compute conditional index (CI) and if the value of CI is lies between 100 and 1000 there is a moderate to strong multicollinearity. If it exceeds 1000 there is severe multicollinearity.

1. Variance-inflation factor

Variance-inflation factor (VIF), defined as

VIF(βk) = 1/(1 - R2k)

VIF may be viewed as the ratio of variance of β^k in the presence of multicollinearity to variance of β^k in the absence of multicollinearity.

When there is no multicollinearity, R2k is zero and VIF = 1/(1 - R2k) = 1/(1 - 0) = 1.

When there is perfect multicollinearity, R2k is close to 1 and VIF = 1/(1 - 1) = 1/0 → ∞.

1. Adding or dropping a few observations to or from the sample, or adding or dropping an independent variable results in significant changes in the estimated values, their signs and their statistical significance.

**Remedial Measures**

Having studied the different techniques for diagnose the multicollinearity, next is to discuss some corrections or remedial measures. The following options are available for alleviating, handling, or coping with multicollinearity:

1. Do Nothing

As mentioned above, even if every measure of intercorrelation among the independent variables points to the existence of strong multicollinearity, one need do nothing if the estimated t ratios are significant at reasonable levels and the estimates coefficients have the expected signs and reasonable magnitudes.

1. A priori information

If prior information is available for parameter of a collinear variable, we can use that

to estimate other parameter of the regression model. by doing this we can avoid the problem of multicollinearity.

1. Drop One of the Collinear Variables

Some suggest dropping one of the collinear variables from the model. While sometimes this is wise, other times dropping a variable can cause omitted-variable bias. Of course, in some cases this bias may be more than offset by the gain in efficiency. In that case mean-squared error (MSE) of the estimated coefficients on the included variable declines indicating an improvement. But how can we tell whether MSE increases or decreases when we omit a variable from the model? If the t ratio for the variable that is a candidate for dropping is less than one in absolute value, then dropping that variable would reduce MSE of the estimated parameter on the included variable.

A corollary to the above rule is to never drop an independent variable whose estimated coefficient has a t ratio that is greater than one in absolute value, even if it has a unexpected (wrong) sign.

1. Transform the Data
2. First Difference the Variables

Differencing reduces spurious correlation that normally arises in time-series data in level form.

1. Express the Variables as Ratios

Sometimes one can greatly reduce even eliminate multicollinearity by combining two multicollinear variables by expressing them as a ratio.

1. Combining cross-sectional and time series data

In some circumstances combining cross section and time series data reduces the degree of multicollinearity. In this technique one of the parameter has to estimate based on cross section information and use it to estimate the remaining parameter. By doing this we can avoid the problem of multicollinearity.

1. Reducing collinearity in polynomial regression

Expressing the explanatory variables in polynomial form may reduce the severity of multicollinearity problem. But even if the problem persist, we have to use a special models called orthogonal polynomials.

1. Other remedial measures
2. Factor analysis and principal components

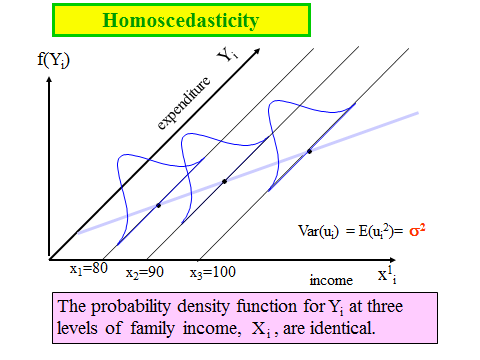
In these methods the collinear variables are reduced into a single variable (a combination of all collinear variables).

1. Ridge regression

It is a solution in case of perfect multicollinearity. We accept a (small) bias in the estimates.

**II Heteroscedasticity**

One of the assumptions of the CLR and CNLR models requires variance of the true error term to be constant, i.e., Var(ui) = σ2 for all i. That is the assumption of heteroscedasticity. When this is violated, we have the problem of heteroscedasticity-- variance of the true error term changes form observation to observation, i.e., Var(ui) = σ2i for all i. The assumption of homoscedasticity and heteroscedasticity are represented in the figures below:



The probability density function for Yi at three levels of family income, Xi , are identical.

Homoscedastic pattern of errors

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

xi

yi

0

In above figure the scattered points spread out quite equally showing homoscedastic nature of errors.

Heteroscedasticity

**.**

x 1

X =80

X =100

Yi

f(Yi)

expenditure

X=120

**.**

**.**

income

Var(ui) = E(ui2)= **σi2**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

xt

yt

0

In the above figures the variance of Yi increases as family income, Xi, increases.

**Consequences of Heteroscedasticity**

1. OLS estimators are still linear and unbiased
2. Var ( βi )s are not minimum. The true variance of βk becomes larger so that

*OLS* estimators will be inefficient

=> not the best => not efficient => not BLUE

1. *^*

*σ2 = RSS/(n-k)* will be a biased estimator of the homoscedastic error variance, *σ2*.

1. In particular, if the error variance increases as one of the independent variables increases, then *σ^2* will be biased *downward*.
2. As a result, the usual formula for estimating *Var(βk)* generate biased estimators for the true, heteroscedastic variance.
3. Since σ2 is usually underestimated in the presence of heteroscedasticity, Var(β^k) is underestimated.
4. Because of the last two results, the usual t and F tests are no longer valid when there is heteroscedasticity.
5. In particular, t ratios will be overestimated thus giving us more confidence than warranted.
6. All F statistics and R2 will be overestimated
7. Thus the probability of committing Type I Error will be higher.

**Detection of Heteroscedasticity**

In most cases involving econometric investigations, heteroscedasticity may be a matter of intuition, educated guesswork, prior empirical experience, or sheer speculation. Some of the informal and formal methods are used for detecting heteroscedasticity. Most of these methods are based on the examination of the OLS residuals uˆi since they are the ones we observe, and not the disturbances ui. One hopes that they are good estimates of ui, a hope that may be fulfilled if the sample size is fairly large.

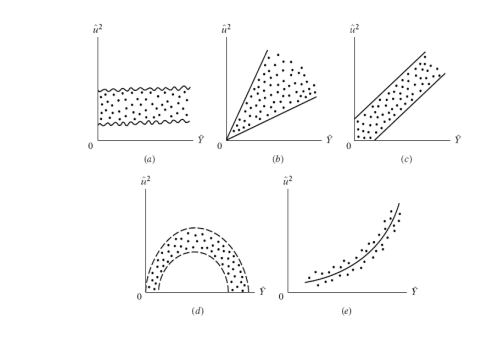
**Informal Methods**

Nature of the Problem: Very often the nature of the problem under consideration suggests whether heteroscedasticity is likely to be encountered. In cross-sectional data involving *heterogeneous* units, heteroscedasticity may be the rule rather than the exception. Thus, in a cross-sectional analysis involving the investment expenditure in relation to sales, rate of interest, etc., heteroscedasticity is generally expected if small-, medium-, and large-size firms are sampled together.

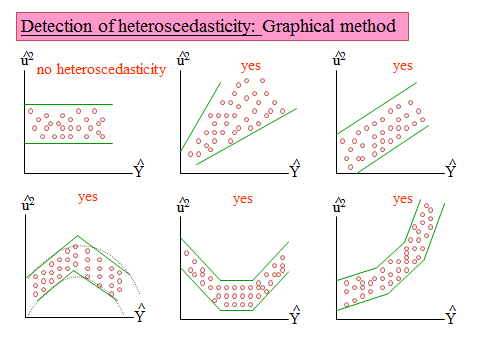
Graphical Method

If there is no a priori or empirical information about the nature of heteroscedasticity, in practice one can do the regression analysis on the assumption that there is no heteroscedasticity and then do an examination of the residual squared *u*ˆ2*i* to see if they exhibit any systematic pattern.

Although *u*ˆ2*i* are not the same thing as *u*2*i* , *they can be used as proxies* especially if the sample size is sufficiently large. An examination of the *u*ˆ2*i* may reveal patterns such as those shown in following figure. In Figure *a* we see that there is no systematic pattern between the two variables, suggesting that perhaps no heteroscedasticity is present in the data. Figure *b* to *e*, however, exhibits definite patterns. For instance, Figure *c* suggests a linear relationship, whereas Figure *d* and *e* indicates a quadratic relationship between *u*ˆ2*i* and *Y*ˆ*i*. Using such knowledge, albeit informal, one may transform the data in such a manner that the transformed data do not exhibit heteroscedasticity.



The following figure also depicts the presence or absence of heteroscedasticity.



**Formal Methods**

1, Park Test.

Park suggests that *σ*2*i* is some function of the explanatory variable *Xi*. The functional form he suggested was

*σ*2*i* = *σ*2*Xβi evi*

or

ln *σ*2*i* = ln *σ*2 + *β* ln *Xi* + *vi*

where *vi* is the stochastic disturbance term. Since *σ*2*i* is generally not known, Park suggests using *u*ˆ2*i* as a proxy and running the following regression:

ln *u*ˆ2*i* = ln *σ*2 + *β* ln *Xi* + *vi* = *α* + *β* ln *Xi* + *vi*

If *β* turns out to be *statistically significant*, it would suggest that *heteroscedasticity* is present in the data. If it turns out to be insignificant, we may accept the assumption of homoscedasticity.

2.Glejser Test.

Test is similar in spirit to the Park test. After obtaining the residuals *u*ˆ*i* from the OLS regression, Glejser suggests regressing the absolute values of *u*ˆ*i* on the *X* variable that is thought to be closely associated with *σ*2*i* . In his experiments, Glejser used the following functional forms: where *vi* is the error term.



But Goldfeld and Quandt point out some limitation of this test. They are the error term *vi* has some problems in that its expected value is nonzero, it is serially correlated and ironically it is heteroscedastic.

1. Spearman’s Rank Correlation Test.

rs = 1 − 6 (d2i/(n(n2 − 1))

where *di* = difference in the ranks assigned to two different characteristics of the *i*th individual or phenomenon and *n* = number of individuals or phenomena ranked.

The preceding rank correlation coefficient can be used to detect heteroscedasticity as follows:

Assume *Yi* = *β*1 + *β*2*Xi* + *ui .*

Step 1. Fit the regression to the data on *Y* and *X* and obtain the residuals *u*ˆ*i*

Step 2. Ignoring the sign of *u*ˆ*i*, that is, taking their absolute value |*u*ˆ*i*|, rank both |*u*ˆ*i*| and *Xi* (or *Y*ˆ*i*) according to an ascending or descending order and compute the Spearman’s rank correlation coefficient using the formula given previously.

Step 3. Assuming that the population rank correlation coefficient *ρs* is zero and *n >* 8, the significance of the sample *rs* can be tested by the *t* test as follows: with df = *n* − 2.



1. The Goldfeld-Quandt Test

This test is applicable in both small and large samples

It assumes heteroscedastic error variance σ2i is proportional to the square of an independent variable, σ2i = σ2Xi2

The mechanics of this test is as follows

STEP 1: Order all observations on the basis of the independent variable suspected of causing heteroscedasticity from smallest to largest

STEP 2: Exclude middle 1/3 of reordered observations

STEP 3: Estimate the model using the two sub-samples and obtain the RSS from each. Denote the RSS from the sub-sample of small X's, RSS1 and that from the sub-sample of large X's, RSS2.

STEP 4: Construct F = RSS2/RSS1. In the CNLR model this has an F distribution with (n2-k) and (n1-k) degrees of freedom in the numerator and denominator, where k is the number of parameters in the model

STEP 5: Use the above F statistic to test the null hypothesis of no heteroscedasticity,

H0: σ21 = σ 22. The decision rule is as usual. That is if the calculated value of F is greater than F critical value reject the null hypothesis of no heteroscedasticity or homoscedasticity.

1. Breusch–Pagan–Godfrey Test:

The success of the Goldfeld–Quandt test depends not only on the value of *c* (the number of central observations to be omitted) but also on identifying the correct *X* variable with which to order the observations. This limitation of this test can be avoided if we consider the Breusch–Pagan–Godfrey (BPG) test.

The following steps are involved in this test:

Step 1. Estimate *Yi* = *β*1 + *β*2*X*2*i* + ·· ·+*βkXki* + *ui*  by OLS and obtain the residuals ˆ*u*1, ˆ*u*2, *. . .* , ˆ*un.*

Step 2. Obtain ˜*σ*2 = ˆ*u*2*i /n*. [*Note:* The OLS estimator is ˆ*u*2*i /*(*n*− *k*)*.*]

Step 3. Construct variables *pi* defined as

*pi* = ˆ*u*2*i /*˜*σ*2

which is simply each residual squared divided by ˜*σ*2*.*

Step 4. Regress *pi* thus constructed on the *Z*’s as

*pi* = *α*1 + *α*2*Z*2*i* +· · ·+*αmZmi* + *vi*

where *vi* is the residual term of this regression.

Step 5. Obtain the ESS (explained sum of squares) from the above regression and define =12 (ESS)

Assuming *ui* are normally distributed, one can show that if there is homoscedasticity and if the sample size *n* increases indefinitely, then ∼asy *χ*2*m* −1 that is, follows the chi-square distribution with (*m* − 1) degrees of freedom.

Step 6. Decision rule: Therefore, if in an application the computed (= *χ*2) exceeds the critical *χ*2 value at the chosen level of significance, one can reject the hypothesis of

homoscedasticity; otherwise one does not reject it.

1. White’s General Heteroscedasticity Test:

Unlike the Goldfeld– Quandt test, which requires reordering the observations with respect to the *X* variable that supposedly caused heteroscedasticity, or the BPG test, which is sensitive to the normality assumption, the general test of heteroscedasticity proposed by White does not rely on the normality assumption and is easy to implement.

This is a large sample test

Consider the following model,

Yi = β1 + β 2X2i + β 3X3i + ui

Step 1: Assumes the following pattern of heteroscedasticity

σ2i = α1 + α2X2 + α3X3 + α4X22 + α5X32 + α6X2X3

Because σ2i is not known, White suggests using squared residual from the original regression as a proxy

Step 2: So estimate the original model and retrieve the estimated residuals, u^.

Step3: Use the squared residuals as a proxy for the heteroscedastic error variance, σi2, and estimate the following auxiliary regression,

^

u2 = α1+ α2X2 + α3X3 + α4X22 + α5X32 + α6X1X2

Step 4: Now test the null of no heteroscedasticity, i.e.,

H0: α2 = α3 =α4 = α5 = α6 = 0

using nR2~χ2(k-1), where n is the sample size, R2 is the unadjusted R2 from the auxiliary regression in Step 3, and k is the number of parameters in the auxiliary regression

Step 5: . If the chi-square value obtained in auxiliary regression exceeds the critical chi-square value at the chosen level of significance, the conclusion is that there is heteroscedasticity. If it does not exceed the critical chi-square value, there is no heteroscedasticity, which is to say that in the auxiliary regression.

**Remedial Measures**

As we have seen, heteroscedasticity does not destroy the unbiasedness and consistency properties of the OLS estimators, but they are no longer efficient, not even asymptotically (i.e., large sample size). This lack of efficiency makes the usual hypothesis-testing procedure of dubious value. Therefore, remedial measures may be called for. There are two approaches to remediation:

1. σ2i is known and
2. When σ2i is not known.
3. σ2i is known: The Method of Weighted Least Squares

Recall that OLS estimators are derived from the least-squares principle of minimizing RSS. This assigns equal weights to all ui’s regardless of their size. This means that in the presence of heteroscedasticity RSS is dominated by large errors, which have larger variances than the smaller errors.

This suggests that an appropriate method for correcting heteroscedasticity is to assign smaller weights to errors with larger variance and larger weights to errors with smaller variances. But what should we use as weights?

If the heteroscedastic error variances, σi2, were known, then the square root of their reciprocal, i.e., 1/σi, would be the appropriate weights. Thus we would divide the data by σi and then minimize the resulting weighted RSS.

* Thus instead of estimating the original model with a hetroscedastic error variance

Yi = β1 + β2Xi+ ui

we estimated the following weighted regression:

Yi/σi = β1/σi + β2(Xi/σi) + ui/σi

Now the error term of the transformed (weighted) model in equation (2), i.e., ui/σi, is homoscedastic,

Var(ui/σi) = (1/σi2)Var(ui) = (1/σi2)σi2 = 1

So if the heteroscedastic variance of the error term, σi2, were known, we could remove heteroscedasticity from the model by dividing the data on the dependent and independent variables by σi. The resulting estimators, which are BLUE in the presence of heteroscedasticity are known as Weighted Least Squares (WLS).

1. When σ2i is not known.

But what about the fact that in practice the heteroscedastic error variance, σi2, is typically unknown? Then what should we use as weights to obtain WLS estimators?

In practice we must make an assumption about the pattern of heteroscedasticity and based on that choose the weights. One way to ascertain the pattern of heteroscedasticity is graphical examination of the sample residuals and their squared. Such an examination may suggest one of the following patterns:

If the residual plot against X3 is as following :

X3

ui

+

0

\_\_

^

X3

u2

**^**

These plots suggest a variance is increasing proportional to X3i2. The scattered plots spreading out as nonlinear trumpet pattern.

Therefore, we might expect σi2 = Zi2σ2 where Zi2 = X3i2 =>Zi =X3i

Hence, the transformed equation becomes

Yi 1 X2i X3i ui

X3i X3i X3i X3i X3i

= β1 + β2 + β3 +

We can write the above expression compactly as follows:

Yi\* = β1 X1\* + β2 X2\* + β3 + u\*

Where Yi\*  is Yi / X3i , X1\* is 1/ X3i . . . so on.

Further in the above equation, u\*i satisfies the assumptions of classical OLS

Since u\*i  satisfies the homoscedastic assumption of classical regression model, we can OLS to estimate the above transformed model and OLS estimators are BLUE.

Another possible pattern of the relation between σi2  and the explanatory variable Xi is

σi2 = σ2Xi2

In this case the appropriate weights for transforming the data would be 1/Xi.

Thus we should transform the original model by dividing all observations by Xi to get

Yi/Xi = β1/Xi + β2(Xi/Xi) + ui/Xi

= β2 + β1(1/Xi) + ui/Xi

So we regress Yi/Xi on a constant and (1/Xi) using the OLS method, in this case also OLS estimators are BLUE.

Because, this makes the error variance of the transformed model homoscedastic, as shown below. To see this, consider the error variance of the transformed model,

Var(ui/Xi) = (1/Xi2)Var(ui) = (1/X2)σ2Xi2 = σ2

The third possible relationship may be

σi2 = σ2Xi

In this case, the proper weights would be 1/√Xi, provided Xi is non-negative so that the square rot is defined and there is no division by zero.

Now, the equation of the transformed model is,

Yi/√Xi = ß1(1/√Xi) + ß2(Xi/√Xi) + ui/√Xi

Again, the transformation has made the new error variance homoscedastic,

Var(ui/√Xi) = (1/Xi)Var(ui) = (1/Xi)σ2Xi = σ2

This can be observed from the following figure:

If the residual plot against X3 is as following

X3

ui

+

0

-

^

X3

u2

**^**

These plots suggest a variance is increasing proportional to X3i. The scattered plots spreading out as a linear cone pattern.

Therefore, we might expect σi2 = *Z*iσ2 where σi2 = X3i  => σ= square root of X3i

The transformed equation is

= β1 + β2  + β3 +

Yi 1 X2i X3i ui

√X3i √X3i √X3i √ X3i √ X3i

=> Yi\* = β1 X1\* + β2 X2\* + β3 X3\* + u\*

Fourth possible relationship can assumed as

σi2 = σ2Yi2

In this case the proper weights would be 1/Yi. Thus we first ignore the fact that there is heteroscedasticity and estimate the original model and obtain the predicted values of Yi. Then we transform the data by dividing all observations on all variables, dependent and independent, by the predicted values of the dependent variable from the previous step.

When the transformed model is estimated by OLS, the resulting estimates, which are BLUE in the presence of heteroscedasticity, are called Two-Stage Weighted Least Squares (2SWLS). This works best in large samples. This transformation is valid if one believes that all independent variables are responsible for heteroscedasticity.

**III Autocorrelation or Serial Correlation**

Another important assumption of classical linear regression model is no autocorrelation or no serial correlation. As saw in the previous section heteroscedasticity problem is normally found in cross section data, because of scale difference. On the other hand the problem of autocorrelation is commonly found in time series type of data. It is because the time series data follow a natural ordering over time so that the successive observations are likely to have inter-relations, if the observed time is short, like daily, weekly or monthly. It does not mean there no presence of autocorrelation in cross section data, if it present it is called as special auto-correlation.

In this section we will discuss the nature, causes , consequences, detection and remedial measures of autocorrelation problem.

1. **Nature of Autocorrelation**

The assumption of no autocorrelation required observations on the error term, u, not be correlated with one another. In time series and cross sectional data it defined as follows:

In time-series samples, this means…

Cov(ut, ut-1) = 0 for all t

In cross-sectional samples, it means

Cov(ui, uj) = 0 for all i ≠ j

When this assumption is violated, we have the problem of autocorrelation or serial correlation

This means *Cov(ut, ut-1)≠ 0* in time series samples

Or *Cov(ui, uj)≠ 0* in cross-sectional samples

Based on some criterion we can classify serial correlation as follows,

* First order and higher order autocorrelation

First-order autocorrelation means,

ut = ρut-1 + vt

where -1≤ ρ ≤ 1

The above relation is known as first-order Markov process and is denoted AR(1).

Note that vt is assumed to be a classical error, i.e.,

* + E(vt) = 0
  + Var(vt) = constant
  + Cov(vt, vt-1) = 0

Higher orders of autocorrelation involve processes in which the error term depend on its own one-, two-, three-, …period lagged values.

For example second-order autocorrelation denoted AR(2) has the following form,

ut = ρ1ut-1 + ρ2ut-2 + vt

Where ρ1 and ρ2 are coefficients of first‑and second-order autocorrelation.

* Positive and Negative autocorrelation

Positive autocorrelation refers to the case where ρ is positive indicating errors of a given sign are generally followed by errors of that same sign.

Negative autocorrelation means ρ is negative so that errors of a given sign are generally followed by errors of the opposite sign.

In economic time-series, autocorrelation is almost always positive. Presence of negative autocorrelation is time-series macroeconomic data is typically a sign of model misspecification.

Perfect Positive and Perfect Negative Autocorrelation

The case of ρ = -1 is called perfect negative first-order autocorrelation.

The special case of ρ = 1 is referred to as perfect positive first-order autocorrelation.

In the case of ρ = 1, the first-order Markov process or AR(1) becomes

ut = ut-1 + vt

1. **Causes for Autocorrelation**

* Inertia or Business Cycle Phenomenon: Economic time series often exhibit cyclical patterns in the sense that if they rise (fall) in one period, they are likely to rise (fall) in the next period as well. That is economic time series such as GDP, Employment, Export, etc. exhibit business cycles pattern in their behavior. Because of this the successive values of the variables are high related and this lead to serial correlation problem.
* Specification Bias: Excluded variables case and functional form

Some times in specifying regression model the econometricians exclude some important variable from the model for some reasons. This is called as excluded variable specification bias. Due to this specification error the autocorrelation problem may present in regression model. Serial correlation may also arise in the model because of fitting a wrong function form.

* Lagged response or Cobweb Effect

Lag in the response of the dependent variable to changes in independent variables can cause pure autocorrelation. 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 take time to implement (the gestation period). This can represented as follows:

Supplyt = β1 + β2Pt−1 + ut

* Autoregression

This means inclusion of the lagged value of the dependent variable in the regression equation as an independent variable,

Yt = ß1 + ß2Xt + ß3Yt-1 + ut

Such models suffer from a number of problems, including autocorrelation.

* Manipulation of Data

Data manipulations prior to estimation can increase the likelihood of autocorrelation. These include: smoothing, seasonal adjustments and moving averages.

**Consequences of Autocorrelation:**

What would happen if we ignore autocorrelation when it exists and continue using OLS formula we derived under the assumption of no autocorrelation?

* To begin with, the OLS estimators of the β’s remain unbiased and consistent.
* In the case of positive autocorrelation, the true variances of β’s become larger than when there is no autocorrelation.
* The variances of βwould be underestimated, even though the true variance of β’sis now larger as a result of positive autocorrelation.

^

* σ2 = RSS/(n-k) underestimates the true σ2.
* Because of the above results, t and F tests would be unreliable in the following sense:
  + Because SE(β) is underestimated and the βis unbiased, t ratios are overestimated yielding more confidence than warranted.
  + Similarly, all F statistics (as well as R2 and adjusted R2) will be overestimated.
  + Thus the probability of committing Type I Error will be higher when OLS is used in the presence of positive autocorrelation.

**Detection Autocorrelation**

1. Graphical method

Visual examination of the estimated residuals, uˆ’s gives some clues about the existence of autocorrelation in the model like heteroscedasticity examined in the previous section. Hence we can get some idea by plotting uˆ’s against the time.

1. The Runs Test

If errors are correlated they follow some systematic pattern. Like several residuals that are negative, then there is a series of positive residuals, and then there are several residuals that are negative. This intuition can be checked by the so-called runs test, sometimes also known as the Geary test, a nonparametric test.

For example, the sequence of positive and negative are shown below.

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

There are 5 positive residuals, followed by 8 negative residuals, followed by 12 residuals, then by 5 residuals and followed by 5 positives residuals for a total of 35 observations.

Now we can define a run as an uninterrupted sequence of one attribute or sign here and length of a run as number of elements in it. In the previous sequence there are 5 runs. To test whether runs are random or not run test has to be used.

1. The Durbin-Watson Test

This is the most popular test of first-order pure autocorrelation, *AR(1)*.

* This test is applicable only if all five of the following conditions are met:
  + Regression equation contains an intercept
  + The independent variables are nonstochastic
  + The model is not autoregressive
  + The pattern of autocorrelation is that of AR(1)
  + The error term is normally distributed

The following is the DW test statistic:

^ ^ ^

d = ∑(ut ‑ut-1)2/∑u2t

^

where ut is the OLS sample residuals.

It can be shown that in large samples,

^

d ≈ 2(1 ‑ρ)

^

where ρ is an estimator of ρ.

So suppose you want to test the null of NO first- order autocorrelation, H0:ρ = 0, against the alternative of H1:ρ ≠ 0.

Then the closer d is to 2 (i.e., ρ is close to 0 ) the less likely you are to reject H0, i.e., the more likely you are to conclude that there is no statistically significant evidence of AR(1) in the data.

On the other hand, the closer d is to 4 (i.e., ρ is close to -1 ), the more likely it is that there is negative first-order autocorrelation (i.e., we reject H0: ρ ≥ 0 in favor of H1: ρ < 0).

Finally, the closer d is to 0 (i.e., ρ is close to 1 ), the more likely it is that there is positive first- order autocorrelation (i.e. we reject H0:ρ≤ 0 in favor of H1: ρ > 0).

But how close is “close?”

Unfortunately, the DW statistic does not have unique critical values.

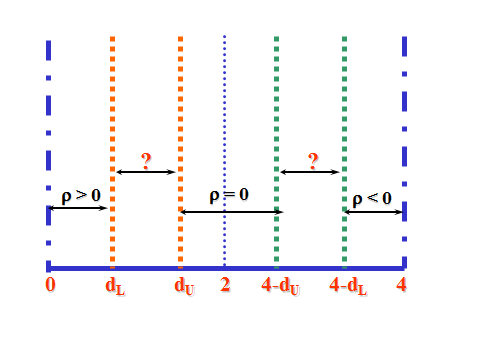
The test procedure is as follows:

* + Estimate the regression equation and calculate the d-statistic.
  + Choose a level of significance and look up du and dl in the D-W table, given the sample size, n, and the number of slope parameters in the model, k’.
  + DW Test Decision Rule : Two Sided Test

Reject H0:ρ = 0 in favor of H1: ρ ≠ 0 if d < dl or if d > (4 - dl)

Do not reject H0: ρ = 0 if du< d < 4 - du

Test is inconclusive if (4 ‑du) < d < (4 - dl) or if dl ≤ d ≤ du



* One-Tail Test

Reject H0:ρ ≤ 0 in favor of H1: ρ > 0 if d < dl

Do not reject H0:ρ ≤ 0 if d > du,

Test is inconclusive if dl ≤ d ≤ du

Reject H0:ρ ≥ 0 in favor of H1:ρ<0 if d > 4-dl

Do not reject H0: ρ ≥ 0 if d < (4 - du)

Test is inconclusive if (4 - du) < d < (4 - dl)

Limitations of the DW Test

* It is strictly a test of *first-order* autocorrelation
* It does not apply to autoregressive models
* It can be inconclusive

IV The Breusch-Godfrey or Lagrange Multiplier Test

This is a test of any order of autocorrelation, it is never inconclusive and it is applicable to both non-autoregressive and autoregressive models.

Consider the following regression model,

Yt = β1 + β2X2t + β3X3t + … + βkXkt + ut

Note that the independent variables may include lagged values of the dependent variable.

Here is the steps for testing ut for Pth-order autocorrelation, where P is a positive integer:

* Estimate the model & retrieve the residuals, ut

^

* Regress ut on the independent variables of the ^

original model (X2t, X3t, …, Xkt) *and* the ut lagged

^ ^ ^

once through P periods (ut-1, ut-2, …, ut-p), where P is the order of autocorrelation you are testing for (e.g., P = 1 if you are testing for AR(1), and would equal 2 if you are testing for AR(2), etc.).

* Test the null of “no pth-order autocorrelation,” H0:α1=α2=...=αp= 0 using the (n-p)R2 as test statistic

(n-p)R2 has aχ2 distribution withP degrees of freedom and The decision rule is as usual. That is rejecting the null hypothesis of no autocorrelation if the calculated value is greater than the table or critical value. Otherwise do not reject.

**Remedial Measures:**

If one or the few of above discussed detection methods identified the presence of autocorrelation in a regression model we must correct it, otherwise the results obtained are dubious of value. Hence we need to develop some remedial measures. This section is devoted to that. As we noted earlier, autocorrelation in majority of cases related to time series data, the rules of thumbs discussed below are pertaining correction of serial correlation for time series data.

The remedy depends on the knowledge one has about the nature of interdependence among the disturbances, that is, knowledge about the structure of autocorrelation.

1. Generalised Least Squares Method ( GLS)

This method is used when ρ is known and the procedure is given below:

Consider the following simple regression

Yt = ß1 + ß2Xt + ut

Assume ut is AR(1), i.e,

ut = ρut-1 + vt

where -1≤ρ≤1 and vt is a classical error term (i.e., it is not autocorrelated).

The idea is to remove the autocorrelation term, ρut-1 from the error term, ut

To see how this is done, rewrite the AR(1) relation as follows,

ut - ρut-1 = vt

What this suggests is that by removing (subtracting) ρut-1 from the autocorrelated error, ut,we obtain the unautocorrelated error, vt.

So to correct for autocorrelation, we transform the data so as to remove ρut-1 from ut.

To do this, lag the regression equation one period

Yt-1 = ß1 + ß2Xt-1 + ut-1

Now multiply it through by ρ,

ρYt-1 = ρß1 + ρß2Xt-1 + ρut-1

Next, subtract this from the original (unlagged) regression equation, i.e.

Yt = ß1 + ß2Xt + ut

Yt - ρYt-1= ß1- ρß1+ ß 2Xt - ρß 2Xt-1+ut -ρut-1

Collect terms and we get,

Yt- ρYt-1= ß 1(1 - ρ) + ß2(Xt - ρXt-1) + (ut - ρut-1)

Remember, ut - ρut-1= vt

Use this to write the previous equation as

Yt ‑ρYt-1 = ß1(1 ‑ρ) + ß2(Xt ‑ρXt-1) + vt

Now, by assumption vt is not autocorrelated, so the above transformed model in is free of autocorrelation.

In simple words, we have to estimate ρ, multiply it times the lagged values of the dependent and independent variables and then apply OLS to the “transformed” data.

The resulting estimators are called Generalized Least Squares (GLS). Unlike OLS estimators, which are not BLUE when the model suffers from autocorrelation, the GLS estimators are BLUE

In particular, GLS estimators are minimum variance, and thus are more efficient than OLS estimators in the presence of first-order true autocorrelation.

In this differencing procedure we lose one observation because the first observation has no antecedent. To avoid this loss of one observation the Prais–Winsten have developed a method and it is known as Prais–Winsten transformation.

When ρ Is Not Known

Although GLS method is conceptually straightforward, is it difficult use in practice because ρ is rarely known. Therefore we need some ways to find it. They are as follows:

1. First Difference Method

Since the value of ρ lies between 1- to +1, we can use this information as starting values. ρ = 0 represents no autocorrelation and extreme values - 1 and + 1 represent perfect negative and perfect positive autocorrelation respectively. Further the perfect negative autocorrelation ruled out in majority of cases in aggregate time series, the possible type of autocorrelation is positive autocorrelation. If we assume ρ = +1 and express the model in first difference form as follows:

Original model - Yt = β1 + β2 Xt + ut

In first difference

Yt − Yt−1 = β2(Xt − Xt−1) + (ut − ut−1)

or

ΔYt = β2ΔXt + εt

Where Δ is the first-difference operator.

(ut − ut−1) or εt  transformed error term which satisfies no autocorrelation. Hence the first difference model corrects the problem of autocorrelation.

1. ρ based on Durbin–Watson d Statistic:

To use the first difference form to correct autocorrelation problem ρ should be +1 or close one. But if it is not sufficiently close to one the another way to approximate the value of ρ is from Durbin–Watson d Statistic. Which can be estimate as follows:

ρˆ ≈ 1 − d 2

the value derived from the above relationship in applicable to large samples and may not holds true in case of small sample.

1. Iterative Methods of Estimating ρ

All the methods the discussed above provide us with only a single estimate of ρ. But there are some iterative methods estimate ρ iteratively, that is, by successive approximation, starting with some initial value of ρ. They are the Cochrane–Orcutt iterative procedure, the Cochrane– Orcutt two-step procedure, the Durbin two–step procedure, and the Hildreth–Lu scanning or search procedure. Of these, the most popular is the Cochran–Orcutt iterative method.

The following procedure is used to estimate ρ from Cochran–Orcutt iterative method.

Consider Yt=β1+β2Xt+ut and assume ut are AR(1)

STEP 1: Ignore the autocorrelation problem and estimate the original model using

OLS. From this, obtain sample residuals, u^t

STEP 2: Assume the AR(1) pattern of autocorrelation in the errors, ut, is reflected

in the sample residuals u^t from Step 1 that is, u^t=ρu^t-1+vt

^ ^

Thus if you regress ut on ut-1, the estimatedcoefficient of u^t-1 will be an

estimate of ρ.

STEP 3: But, the estimate of ρ from Step 2 is biased since it is from an autoregressive model.

To reduce this bias, use the estimated value of ρ from Step 2 to transform

the data into the generalize-differenced form, estimate the resulting model

(see below), and obtain the residuals, which we can call wt,

^ ^ ^

Yt - ρYt-1 = β1(1 - ρ) + β2(Xt - ρXt-1) + wt

STEP 4: Assume the residuals from Step 3, i.e., wt, are AR(1) as well, and estimate

the second-iteration ρ by estimating the following

^

wt = ρwt‑1 + mt

STEP 5: Continue iterating by repeating Steps 3 and 4 until the difference between

the estimates of ρ from two consecutive iterations is smaller than a certain

tolerance level (say, 0.001). This would then be the final estimate of ρ.

This method is known in literature as Feasible Generalised Least Squares (FGLS).

1. The Newey–West Method

Another method to correct autocorrelation problem is Newey–West method. In this method we can continue to use OLS but we can correct standard errors for autocorrelation. This is an extension of White’s heteroscedasticity-consistent standard errors and the corrected standard errors are known as HAC (heteroscedasticity- and autocorrelation-consistent) standard errors or simply Newey–West standard errors. This method also true only in large sample.

**Module - V  
DUMMY VARIABLE AND DYNAMIC REGRESSION MODELS**

**I: Dummy Variable Model**

One of the serious limitations of simple and multiple-regression analysis, as discussed in modules II and III, is that it accommodates only quantitative response, Y and explanatory variables, Xs. That is we have considered regression models in which both the dependent and independent variables have been quantitative, e.g., income, consumption interest rate, etc. In practice, sometimes we need to control for variables that are not quantifiable in the sense that they are not quantitative but qualitative, e.g., strikes, wars, gender, race, religion, etc.

Since qualitative variables indicate the existence or absence of an event or attribute, they can be represented by variables that take on but two distinct values, 0 and 1. Zero indicating absence of a quality, and one indicating presence of a quality. Such variables are called dummy, qualitative, indicator, binary, dichotomous, or categorical. Thus, dummy variables are artificial variables used to control for qualitative effects and are able classify into mutually exclusive categories.

Dummy variables indicate a “quality” or an attribute,

such as “male” or “female”,

“black” or “white”,

“urban” or non-urban”

“before” or “after”

“North” or “south”, “east” or “west”

………..etc.

While either the dependent variable(s) or the independent variable(s) or both can be dummy, we only consider the case where some or all of the independent variables are dummy variables, the dependent variable being quantitative in this module.

Dummy variables are introduced to regression model just as easily as quantitative variables. If all the independent variables in a regression model are dummy or qualitative variables those models are called Analysis of Variance (ANOVA) models. But in a regression model includes both quantitative and qualitative variables as independent variables it is called as Analysis of Co-Variance (ANCOVA) models.

**Caution in the Use of Dummy Variables**

Although incorporating dummy variables into a regression model easy, but one must use it carefully. Follows are the guidelines one cloud follow:

* 1. The quality that is assigned the value of zero is called the base, control, reference, comparison, or omitted quality or category. Note that the common intercept term, ß1, is that of the base quality
  2. The choice of the base quality is arbitrary
  3. The coefficient of the dummy variable, ß2, is called the differential intercept term, since it is the value by which the common intercept, ß1, is changed due to the dummy variable.
  4. When there are only two mutually exclusive qualities (e.g., male and female), all we need is one dummy variable, as long as the model already includes an intercept term.

In general, if there are a total of q qualities, include q - 1 dummy variables in the model, provided the model includes an intercept

Why? Because if you include q dummy variables, you cause perfect multicollinearity, a situation known dummy-variable trap

For example,

If we introduce two dummy variables in one model to identify two categories of one qualitative variable such as

Yi = β1+ β\*1 D1i + β\*\*1 D2i + β2 Xi + ui

where D1i = 1 if female

= 0 otherwise

where D2i = 1 if male

= 0 otherwise

This model cannot be estimated because of perfect collinearity between D1 and D2,  Shown below

D1  = 1 – D2 or

D2  = 1 – D1  or

D1  + D2 = 1 ( Perfect collinearity )

Use two dummy variables to identify two different qualitative categories in one model

will be fall into the “Trap of perfect multi-collinearity”. Hence the General rule is

To avoid the perfect multicollinearity, If a qualitative variable has “q” categories,

introduce only “q-1” dummy variables.

**Dummy Variable Model: A Case of Single Qualitative Variable With Two Categories**

This model includes a quantitative dependent variable and only one qualitative or dummy independent variable. The purpose of this model is to calculate the mean difference in the dependent variable because of qualitative independent variable. This is also known as factor variable in this context. For example, suppose our purpose is to identify wage discrimination between males and females, we can express the dummy variable model as follows:

Yi = β1 + β2D1 + ui

Where D1 = male, D0 = female.

In this example we considered female as reference or base category, its value is obtained

from β1

Take the conditional mean of Y making use of the fact that E(ui | Di) = 0,

E(Yi|Di=0) = ß1 = average wage of females

E(Yi|Di=1) = (ß1+ß2) = average wage of males

Here ß2 is the difference between average wage of female workers and that of male workers. Further we can use the OLS to estimate unknown parameters and after estimating the parameters, namely β1 and β2, we have to test whether the difference in the average wage between males and females statistically significant or not by suing the t test.

Suppose we want to find out if the average salary of 25 workers differs among males and females. We can use dummy variable with single qualitative variable for this purpose.

Data on gender and salary of twenty five workers

|  |  |  |  |
| --- | --- | --- | --- |
|  | male | female | Annual Salary(k) |
| obs | D2 | D1 | Y |
| 1 | 0 | 1 | 23 |
| 2 | 1 | 0 | 19.5 |
| 3 | 0 | 1 | 24 |
| 4 | 1 | 0 | 21 |
| 5 | 0 | 1 | 25 |
| 6 | 1 | 0 | 22 |
| 7 | 0 | 1 | 26.5 |
| 8 | 1 | 0 | 23.1 |
| 9 | 1 | 0 | 25 |
| 10 | 0 | 1 | 28 |
| 11 | 0 | 1 | 29.5 |
| 12 | 1 | 0 | 26 |
| 13 | 1 | 0 | 27.5 |
| 14 | 0 | 1 | 31.5 |
| 15 | 1 | 0 | 29 |
| 16 | 0 | 1 | 22 |
| 17 | 1 | 0 | 19 |
| 18 | 0 | 1 | 18 |
| 19 | 1 | 0 | 21.7 |
| 20 | 1 | 0 | 18.5 |
| 21 | 0 | 1 | 21 |
| 22 | 0 | 1 | 20.5 |
| 23 | 1 | 0 | 17 |
| 24 | 1 | 0 | 17.5 |
| 25 | 0 | 1 | 21.2 |

The estimated regression results are as follows:

Salary = 16.66 + 7.78 D1

Se = (1.459) (2.023)

t = (11.419) (3.847)

P. value = (0.000) (0.001)

As these regression results shows, the mean salary of female worker is about Rs.16.66

thousands whereas for males it is about 24.44 Rs. thousand (16.66 + 7.78) that is (ß1+ß2). It indicates that there is a difference in the salary between man and women or wage discrimination. Male workers salary is higher by an amount of Rs. 7.78 thousands (ß2) compare to females. It is about Rs.16.66 thousands (ß1). The result also show that the difference in the salary is also statistically significant at 1 percent level, since the p. value is 0.001.

**Dummy Variable Model: A Case of Single Qualitative Variable With More Than Two Categories**

A dummy variable may also have more than two categories. We can include all the categories into the dummy variable regression by adding the dummies. Let consider a dummy with three categories.

Yi = ß1 + ß2D2 + ß1D3 + Ui

Where Yi is health care expenditure,

D1 = 1 if less than high school education

= 0 otherwise

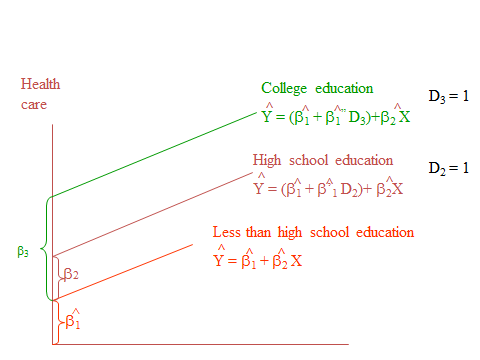
D12 = 1 if high school education

= 0 otherwise

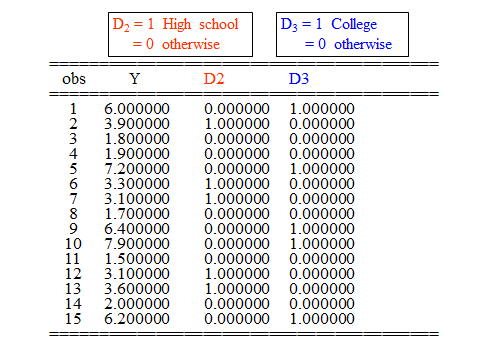
D3 = 1 if college education

= 0 otherwise

The above dummy variable model can be diagrammatically given below:



suppose we data on 15 persons on health care expenditure and the level of education and it is given in the following table.



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dependent Variable: HEALTHCARE | | | |  |  |
| Method: Least Squares | | |  |  |  |
| Date: 12/25/20 Time: 16:44 | | |  |  |  |
| Sample: 1 15 | |  |  |  |  |
| Included observations: 15 | | |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| Variable | Coefficient | Std. Error | t-Statistic | Prob. |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| C | 1.780000 | 0.228765 | 7.780917 | 0.0000 |  |
| HIGH\_SCHOOL | 1.620000 | 0.323522 | 5.007383 | 0.0003 |  |
| COLLEGE | 4.960000 | 0.323522 | 15.33125 | 0.0000 |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| R-squared | 0.953211 | Mean dependent var | | 3.973333 |  |
| Adjusted R-squared | 0.945412 | S.D. dependent var | | 2.189412 |  |
| S.E. of regression | 0.511534 | Akaike info criterion | | 1.674050 |  |
| Sum squared resid | 3.140000 | Schwarz criterion | | 1.815660 |  |
| Log likelihood | -9.555372 | Hannan-Quinn criter. | | 1.672541 |  |
| F-statistic | 122.2344 | Durbin-Watson stat | | 2.393376 |  |
| Prob(F-statistic) | 0.000000 |  |  |  |  |

Measuring the estimated results of different education level

Less than high school: health care = 1.780000

High school: health care = 1.780000 + 1.620000

College health care = 1.780000 + 4.960000

Since the t values for both D2 and D3 dummy slopes are statistically significant at 1 percent level, we can conclude that there is a significant difference in the health expenditure among respondents with different level of education.

**Dummy Variable Regression With Two Qualitative Variables**

A quantitative dependent variable is not just influenced by a single dummy variable. In practice it can be influenced several qualitative variables. In such situations we extend our model by incorporating them in to the model. let us consider a model with two dummy or qualitative variables.

Salary = β1 + β2D2 + β”1 D3 + u

Sex D1 = 1 if male

= 0 otherwise

Race D2 = 1 if white

= 0 otherwise

Regression with a Mixture of Quantitative and Qualitative Regressors: The ANCOVA Models:

ANOVA models of the type discussed in the preceding only include the qualitative or dummy variables as independent variables. But in practice, both quantitative and qualitative variables have exert influence on the quantitative dependent in economics.. Regression models containing an admixture of quantitative and qualitative variables are called analysis of covariance (ANCOVA) models. These models are an extension of the ANOVA models in that they provide a method of statistically controlling the effects of quantitative regressors, called covariates or control variables, in a model that includes both quantitative and qualitative, or dummy, explanatory variables.

Model: Yi = β1 + β2 D1 + β3Xi + ui

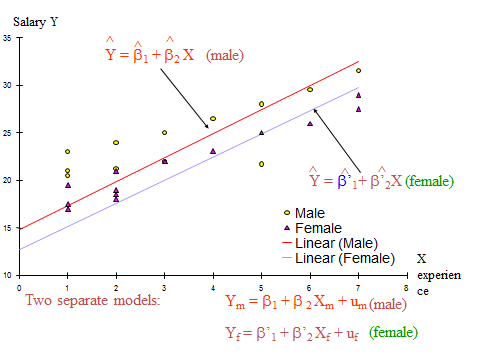
Where Yi = annual salary

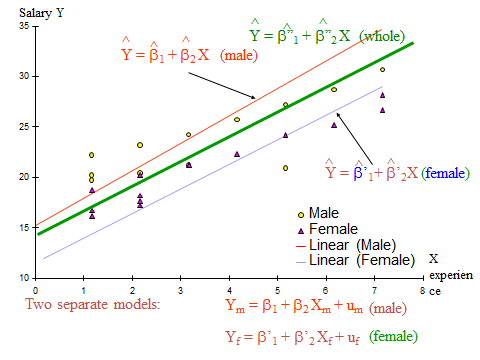
D1 = 1 if male ( qualitative Variable)

= 0 otherwise, (i.e. female)

Xi = years of experience ( covariate or quantitative variable)

The results are shown in the table below (Eviews Output):

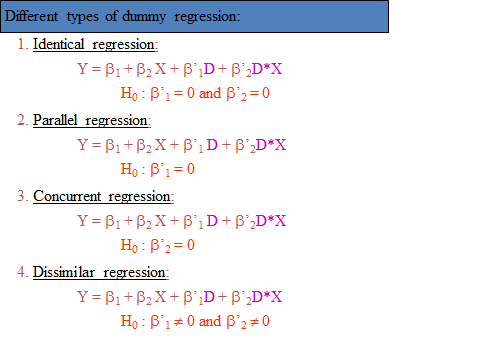


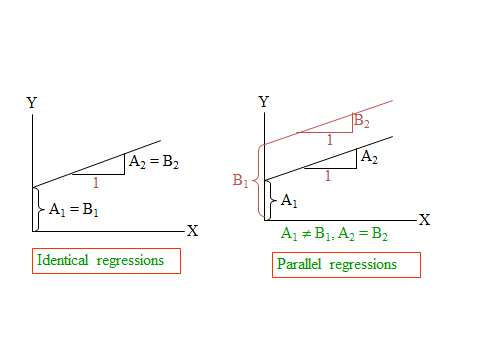


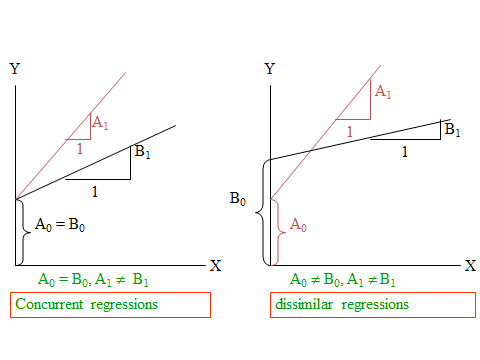
The results are shown in the table below (Eviews Output):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dependent Variable: ANNUAL\_SALARY | | | |  |
| Method: Least Squares | | |  |  |
| Date: 12/25/20 Time: 19:26 | | |  |  |
| Sample: 1 25 | |  |  |  |
| Included observations: 25 | | |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|  |  |  |  |  |
|  |  |  |  |  |
| C | 17.40129 | 1.262157 | 13.78695 | 0.0000 |
| MALE | 5.296868 | 1.110518 | 4.769725 | 0.0001 |
| YEARS\_OF\_EXPERIENCE | 1.346293 | 0.290738 | 4.630605 | 0.0001 |
|  |  |  |  |  |
|  |  |  |  |  |
| R-squared | 0.700445 | Mean dependent var | | 24.95200 |
| Adjusted R-squared | 0.673213 | S.D. dependent var | | 4.804262 |
| S.E. of regression | 2.746371 | Akaike info criterion | | 4.970605 |
| Sum squared resid | 165.9362 | Schwarz criterion | | 5.116870 |
| Log likelihood | -59.13256 | Hannan-Quinn criter. | | 5.011172 |
| F-statistic | 25.72114 | Durbin-Watson stat | | 1.546624 |
| Prob(F-statistic) | 0.000002 |  |  |  |

The estimated results reveal the mean salary of females is about Rs. 17.4 thousands and for mens’ it is about Rs. 22.7 thousands. Further, the difference in the annual salary between males and females is about Rs. 5.3 thousands and this observed difference is statistically significant at 1 percent level. As expected the influence of experience on the dependent variable, annual salary is found to be positive and it is also significant at 1 percent level. Estimated coefficient is 1.35, which indicates that for every extra year of service on the average the annual salary for both males and females increases by Rs.1.35 thousands.





  
 The dummy variable models have wide applications in practice. To mention few important application are like

* + Testing Stability of the Regression Equation
  + Seasonal Analysis
  + Piecewise Linear Regression (Jack-Knifing)
  + Estimating Pooled Time-series/Cross-section Models

**II Distributed Lag and Dynamic Econometric Models**

Economic theory may suggest that the relationship between the dependent and the independent variables has a “dynamic” component:

* + Past values of the independent variables may determine the current value of the dependent variable (Models with lagged independent variables). This is a Distributed Lag model.

For example, today's supply is a function of prices in past periods, today's

price is a function of money supply in past periods

* + Past values of the dependent variable may influence the current value of the dependent variable. This is a Autoregressive or dynamic model.
    - In some situations, economists have even shown that the past value of a dependent variable was the best predictor of the current value. Stock index, Money supply etc

Therefore in time series type of data we have two type of model where the past or previous values of the variable influences on the current values of another variable. They may represent as follows:

A) Distributed-lag Models: If a model contains both the current and past (lagged) values of the explanatory variables.

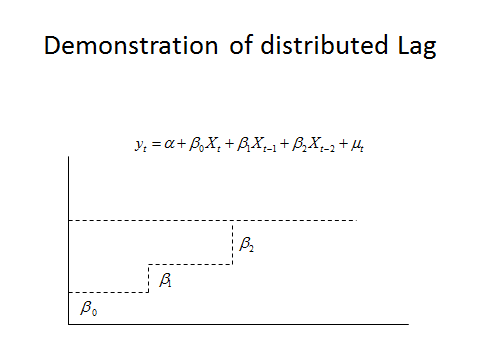
Yt = α + β0Xt + β1Xt-1 + … + βkXt-k + ut

B) Autoregressive or Dynamic Models: If a model includes one or more past values of the dependent variable among its explanatory variables.

Yt = α0 + α 1Xt + α 2Yt-1 + vt

The Role of “Time,’’ or “Lag,’’ in Economics

In economics the response a variable to changes in the other related variable is not instantaneous or immediate. The dependent or a response variable needs some time to adjust for the changes in independent or explained variable. Such a lapse or delay of time is called ‘lag’ in econometrics.



The above figure shows the how the total effect of an independent variable spread or distributed over a period of time in terms of βs. By using distributed model we are able to estimate short run, intermediate and long run effects or multipliers. We can show that by the following model.

Yt = α + β0Xt + β1Xt-1 + … + βkXt-k + ut

β0 = short run multiplier

(β0 + β1) or (β0 + β1  + β2 ) are example of interim or intermediate multipliers.

∑ βi = β0 + β1 + β2 + ···+ βk = β is known as the long-run, or total, distributed-lag multiplier.

The Reasons for Lags

Psychological reason: due to habit or inertia nature, people will not react fully to changing factors, e.g. Income, price level, money supply etc.

Information reason: because imperfect information makes people hesitate on their full response to changing factors.

Institutional reason: people cannot react to change because of contractual obligation.

Estimation of Distributed-Lag Models

Once a distributed lag model is specified, next task is to estimate the model. following are the methods used in practice.

1. **Ad Hoc Estimation of Distributed-Lag Models**

First regress Yt on Xt, then regress Yt on Xt and Xt-1, then regress Yt on Xt, Xt-1 and Xt-2, and so on.

Y = α + β0Xt

Y = α + β0Xt  + β0Xt-1

Y = α + β0Xt  + β0Xt-1 + β0Xt-2 so on.

This sequential procedure stops when the regression coefficients of the lagged variables start becoming statistically insignificant and / or the coefficient of at least one of the variables change signs, which deviate from our expectation.

Problems of Ad-hoc Estimation

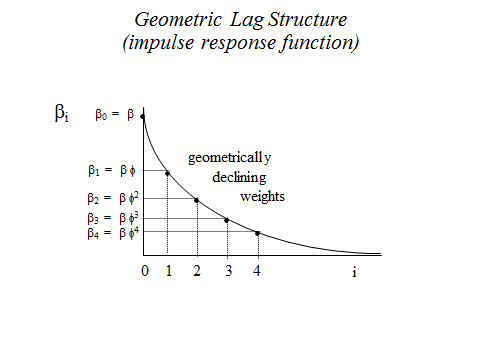
* No a priori guide to length of lag.
* Longer lags => less degrees of freedom
* Multicollinearity
* Need long enough data to construct the distributed-lag model.

2. **The Koyck Approach to Distributed-Lag Models**

Koyck has proposed an ingenious method of estimating distributed-lag models. Suppose we start with the infinite lag distributed-lag model and assuming all βs coefficients are all of same sign they decline geometrically.

Koyck model has following important feature :

* Converts infinite d-l model to estimatable model
* Assume all *b’s* have the same sign and that the lag structure is infinite
* βk = β0λk , k = 0, 1, 2…, 0 < λ < 1
* λ = rate of decay; 1 - λ = speed of adjustment.
* i.e. effect of lagged X’s is geometrically damped as lag increases
* β*’s* can’t change sign (0 < l)
* less weight to distant β*’s* (l < 1)
* Sum of βk up to infinity = β0(1/(1-l))
* Such a lag structure (with geometrically declining weights) is known as a Koyck distribution.



Koyck Transformation

* Yt = α + β0Xt + β 0λ Xt-1 + β 0λ 2Xt-2 +… + ut (i)

lag (i) by one period and multiply by λ:

* λYt-1 = α + β0Xt + β 0λ Xt-1 + β 0λ 2Xt-2 + β 0λ 3Xt-3 … + λ u-t (ii)

subtract (ii) from (i):

* Yt - λYt-1 = α(1 - λ) + β0Xt + (ut - λut-1) (iii)
* or
* Yt = α(1 - λ) + β0Xt + λYt-1 + vt

This procedure transformed the infinite distributed lag model to a one period Autoregressive model. It only has one extra parameter (λ) than to the static model. Coefficient of any period distributed lag can be now estimated by using the βk = β0λk .

**Problems with Koyck Model**

* Equation Yt = a(1 - l) + b0Xt + lYt-1 + vt  is an AR (autoregressive) model
* Yt-1 is a stochastic regressor and may be correlated with error term ( estimation requires Instrument variable in OLS or MLE)
* Problem of autocorrelation
* Durbin Watson d test statistic is not appropriate to test for the presence of first order autocorrelation instead Durbin H is used for large samples. h statistic is asymptotically normally distributed with zero mean and unit variance; recall P(-1.96 < h < 1.96) = 0.95. Ho=No autocorrelation



**Adaptive Expectations Model**

* + “Expected” value of the independent variable explains the movement in dependent variable.
  + Example: relationship between prices and wages.
  + Wages are negotiated at the beginning of the period. Unions form expectations about prices
  + “Expected” Price not observable data
  + Economists have proposed several ways of modeling expectations. The “adaptive expectations model” assumes that agents use information from the past and *learn from their mistakes*



Adjustment Parameter Past errors in expectations

( lies between 0 and 1)



This method has similar problems as Koyck in using OLS. OLS may not appropriate if assumptions are violated so model can be estimated by maximum likelihood. Since maximum likelihood method is complex, simple method of estimation is instrumental variables.

**Partial Adjustment Model**



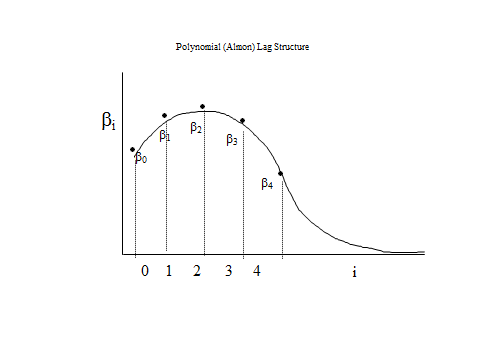
**Properties of Partial Adjustment Model**

* Estimating equation looks like Koyck but is different as far as estimation is concerned
* Error term is well behaved
* In the limit the lagged dependent variable is uncorrelated with the error term
* model can be estimated consistently by OLS

**Polynomial Distributed Lag - Almon Approach**

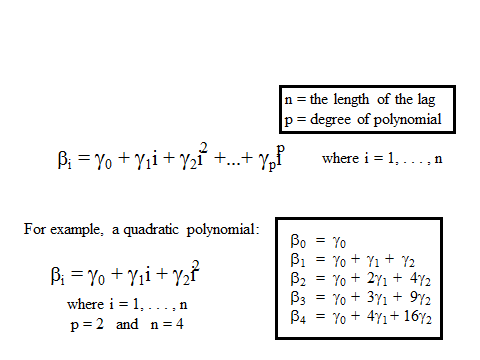
Another method of estimation of distributed model was developed by Shirley Almon approach. This method has following properties:

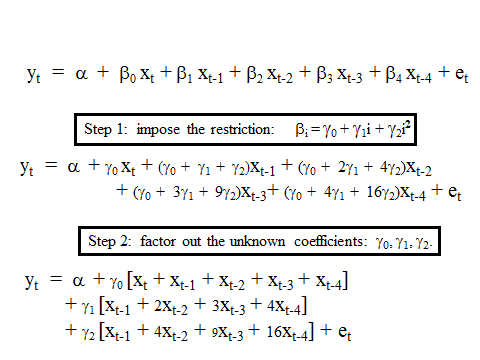
* “hump” shape
* Polynomial --- quadratic or higher
* Still Finite: the effect of X eventually goes to zero
* The coefficients are related to each other
  + the effect of each lag will not necessarily be less than previous one i.e. not uniform decline

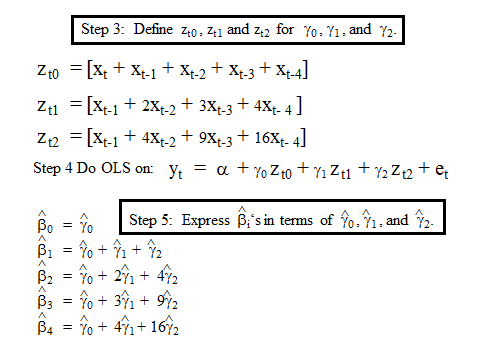


**Estimation**

* Estimate using OLS
* only need to estimate p parameters: g
  + number of parameters is equal to degree of polynomial
* Have to do some algebra to rewrite the model in form that can be estimated.
  + model reduces to arithmetic model if polynomial is of degree 1
* Do OLS on transformed model







**Advantages/Disadvantages**

Fewer parameters to be estimated (only the degree of polynomial) than in the unrestricted lag

structure

* + more precise

What if the restriction is untrue?

* + - biased and inconsistent

**MODULE-VI: SIMULTANEOUS EQUATION MODELS**

So far we only discussed regression model in which a dependent variable Y is explained by one or a set of explanatory variables with assumption that there one way causation from independent variables to the dependent variables. However, many variables in economics are interdependent. Consider a market model with demand and supply. How much price can a firm charge for a particular product from its customers depends on the quantity sold in the market and how much quantity is demanded by customers depend on the market price. Price determines quantity and quantity determines price. Same is true in national income determination model. Consumption is a component of income that determines income, but income is the major determinant of consumption. Both quantities and prices and income and consumption are determined simultaneously.

So the basic assumption of a linear regression model that the explanatory variable and disturbance are uncorrelated or explanatory variables are fixed is violated and consequently ordinary least squares estimator becomes inconsistent.

We need to estimates a system of equations, not a single equation, in order to be able to capture this interdependency among variables.

The main features of a simultaneous equation model are:

1. two or more dependent (endogenous) variables
2. A set of equations
3. Computationally cumbersome, highly non-linearity in parameters and errors in one equation transmitted through the whole system

Similar to the classification of variables as explanatory variable and study variable in linear regression model, the variables in simultaneous equation models are classified as endogenous variables and exogenous variables.

# Endogenous variables (Jointly determined variables)

The variables which are explained by the functioning of system and values of which are determined by the simultaneous interaction of the relations in the model are endogenous variables or jointly determined variables.

# Exogenous variables (Predetermined variables)

The variables that contribute to provide explanations for the endogenous variables and values of which are determined from outside the model are exogenous variables or predetermined variables.

The classification of variables as endogenous and exogenous is important because a necessary condition for uniquely estimating all the parameters is that the number of endogenous variables is equal to the number of independent equations in the system. Moreover, the main distinction of predetermined variable in estimation of parameters is that they are uncorrelated with disturbance term in the equations in which they appear.

An Example

Consider the demand and supply models

*q* = β1 + β2*p* + β3 *y* + *u*1  demand function

*q* = α1 + α2*p* + α3*R* + *u*2  supply function

q is the quantity, p the price, y the income, R the rainfall, and u1 and u2 are the error terms

Here p and q are the endogenous variables and y and R are the exogenous variables

**Simultaneity Problem -An Example**

# Consider a relation between quantity and price



A priory it is impossible to say whether this a demand or supply model, both of them have same variables. If we estimate a regression model like this how can we be sure whether the parameters belong to a demand or supply model? We need extra information. Economic theory suggests that demand is related with income of individual and supply may be respond lagged price or state of technology.

Demand



Supply



where is quantity demanded and is quantity supplied, is the price of commodity, is price lagged by one period, is income of an individual, and are independently and identically distributed (iid) error terms with a zero mean and a constant variance.

and are endogenous variables and  and are exogenous variables, , ,, ,  andare six parameters defining the system.

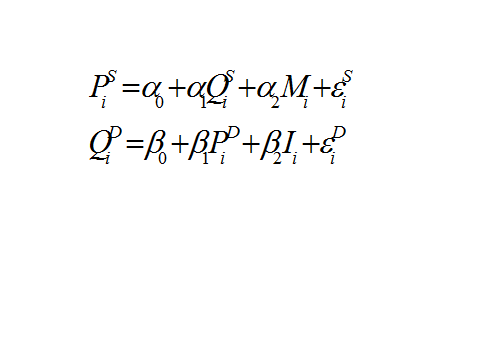
In equilibrium =, however note that quantities bought and sold (Q) depends upon market prices (P) and the equilibrium prices is determined by quantity supplied and demanded.

The random terms of quantity and price equations above,  and , are not independent of each other. There is simultaneity problem.

Thus in the above equations the error term u is correlated with right hand variables. Hence an estimation of the equation by ordinary least squares produces inconsistent estimates of the parameters.

Structural and Reduced form equations.

Structural equations: The underlying equations derived from economic behavior are called the structural equations. We cannot estimate the structural equations directly.

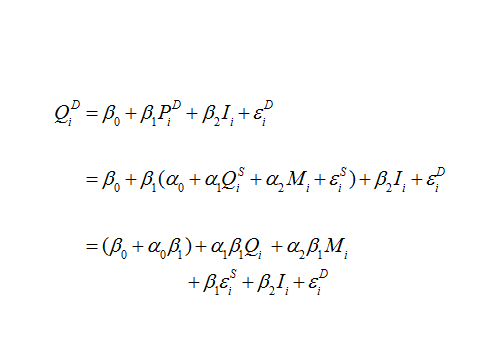


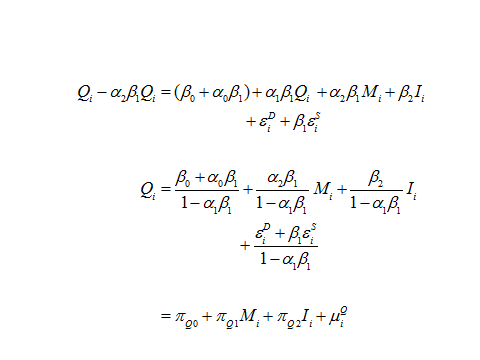
These structural equations cannot be estimated directly, because they both contain endogenous variables on the RHS.

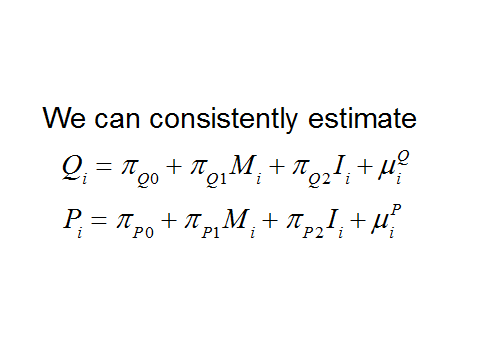
To estimate an equation (consistently), we need all explanators to be exogenous.

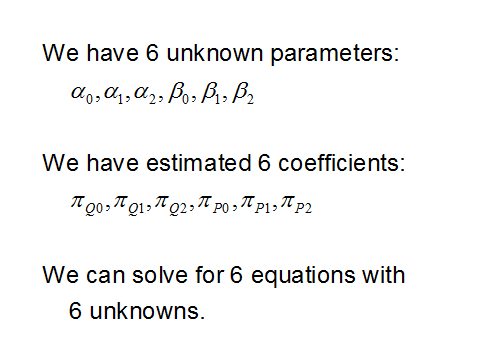
Reduced Form Equations: equations that express endogenous variables as functions only of exogenous variables and disturbances.

We can estimate reduced form equations consistently.









**The Identification Problem**

Before you estimate a structural equation that is part of a simultaneous equation system, you must first determine whether the equation is identified. If the equation is not identified, then estimating its parameters is meaningless. This is because the estimates you obtain will have no interpretation, and therefore will not provide any useful information.

Classifying Structural Equations

Every structural equation can be placed in one of the following three categories.

1. *Unidentified equation* – The parameters of an unidentified equation have no interpretation, because we do not have enough information to obtain meaningful estimates. In other words an equation is under-identified (or not identified or unidentified) if its structural (behavioural) parameters cannot be expressed in terms of the reduced form parameters.
2. *Exactly identified equation* – The parameters of an exactly identified equation have an interpretation, because we have just enough information to obtain meaningful estimates. It means, an equation is exactly identified if its structural (behavioural) parameters can be uniquely expressed in terms of the reduced form parameters.
3. *Overidentified equation* – The parameters of an overidentified equation have an interpretation, because we have more than enough information to obtain meaningful estimates. Also defined as an equation is over-identified if there is more than one solution for expressing its structural (behavioural) parameters in terms of the reduced form parameters.

**Rank and Order Condition for Identification**

Exclusion restrictions are most often used to identify a structural equation in a simultaneous equations model. When using exclusion restrictions, you can use two general rules to check if identification is achieved. These are the rank condition and the order condition. The order condition is a necessary but not sufficient condition for identification. The rank condition is both a necessary and sufficient condition for identification. Because the rank condition is more difficult to apply, many economists only check the order condition and gamble that the rank condition is satisfied. This is usually, but not always the case.

**Order Condition**

The order condition is a simple counting rule that you can use to determine if one structural equation in a system of linear simultaneous equations is identified. Define the following:

G = total number of endogenous variables in the model (i.e., in all equations that comprise the model).

K = total number of variables (endogenous and exogenous) excluded in the equation being checked for identification.

The order condition is as follows.

If K = G – 1 the equation is exactly identified

If K > G – 1 the equation is overidentified

If K < G – 1 the equation is unidentified

Example



Equation 1: Y1= 3Y2 -2X1 –X2+U1

Order Condition:

|  |
| --- |
| G = total number of endogenous variables in the model (i.e., in all equations that comprise the model). |
|  |
| K = total number of variables (endogenous and exogenous) excluded in the equation being checked for identification. |

K = 2

G = 3

K = G -1

2 = 3 -1

2 = 2

Exactly identified

Equation 2: Y2=Y3+X3+U2 Order Condition

|  |
| --- |
| G = total number of endogenous variables in the model (i.e., in all equations that comprise the model). |
|  |
| K = total number of variables (endogenous and exogenous) excluded in the equation being checked for identification. |

K = 3

G = 3

K = G -1

3 = 3 -1

3 = 2

Over-identified

Equation 3: Y3 = Y1  - Y2 + 2X3 + U3 Order Condition:

|  |
| --- |
| G = total number of endogenous variables in the model (i.e., in all equations that comprise the model). |
|  |
| K = total number of variables (endogenous and exogenous) excluded in the equation being checked for identification. |

K = 2

G = 3

K = G -1

2 = 3 -1

2 = 2

Exactly identified

**Rank Condition**

The rank condition tells you whether the structural equation you are checking for identification can be distinguished from a linear combination of all structural equations in the simultaneous equation system. The procedure is as follows.

1. Construct a matrix for which each row represents one equation and each column represents one variable in the simultaneous equations model.
2. If a variable occurs in an equation, mark it with an X. If a variable does not occur in an equation, market it with a 0.
3. Delete the row for the equation you are checking for identification.
4. Form a new matrix from the columns that correspond to the elements that have zeros in the row that you deleted.
5. For this new matrix, if you can find at least (G – 1) rows and columns that are not all zeros, then the equation is identified. If you cannot, the equation is unidentified.

Example:







Equation 1



As this is non-zero, equation 1 is exactly identified.

Equation 2





As this is non-zero, equation 2 is exactly identified.

Equation 3:



As this is zero, equation 3 is not identified.

**Estimation of Simultaneous Equation Model**

They are two approaches to estimate the simultaneous equations, namely, single-equation methods, also known as limited information methods, and system methods, also known as full information methods. In the single-equation methods we estimate each equation in the system (of simultaneous equations) individually, taking into account any restrictions placed on that equation (such as exclusion of some variables) without worrying about the restrictions on the other equations in the system, hence the name limited information methods. In the system methods, on the other hand, we estimate all the equations in the model simultaneously, taking due account of all restrictions on such equations by the omission or absence of some variables (recall that for identification such restrictions are essential), hence the name full information methods.

**Single Equation Estimation**

Single equation estimation involves estimating either one equation in the model, or two or more equations in the model separately. For example, suppose we have a simultaneous equation regression model that consists of two equations: a demand equation and a supply equation. Suppose the objective is to obtain an estimate of the price elasticity of demand. In this case, we might estimate the demand equation only. Suppose our objective is to obtain estimates of price elasticity of demand and price elasticity of supply. In this case, we might estimate the demand equation by itself and the supply equation by itself.

**System Estimation**

System estimation involves estimating two or more equations in the model jointly. For instance, in the above example we might estimate the demand and supply equations together. We might do this even if the objective is to obtain an estimate of the price elasticity of demand only.

Advantages and Disadvantages of the Two Approaches

The major advantage of system estimation is that it uses more information, and therefore results in more precise parameter estimates. The major disadvantages are that it requires more data and is sensitive to model specification errors. The opposite is true for single equation estimation.

**Single Equation Methods**

1. Recursive Model and OLS Estimation

Consider the following system of equations:



Assume that the error terms are not correlated with each other. Can we estimate the equations individually using OLS?

Equation 1: Contains no endogenous variables, so *X*1 and *X*2 are not correlated with *u*1. So we can use OLS.

Equation 2: Contains endogenous *Y*1 together with exogenous *X*1 and *X*2. We can use OLS on if all the RHS variables in this equation are uncorrelated with that equation’s error term. In fact, *Y*1 is not correlated with *u*2 because there is no *Y*2 term in equation. So we can use OLS on.

Equation 3: Contains both *Y*1 and *Y*2; we require these to be uncorrelated with *u*3. By similar arguments to the above, equations (1) and (2) do not contain *Y*3, so we can use OLS on (3).

This is known as a RECURSIVE or TRIANGULAR system. We do not have a simultaneity problem here.

But in practice not many systems of equations will be recursive.

1. Indirect Least Squares (ILS)

Cannot use OLS on structural equations, but we can validly apply it to the reduced form equations. If the system is just identified, ILS involves estimating the reduced form equations using OLS, and then using them to substitute back to obtain the structural parameters.

 However, ILS is not used much because

a. Solving back to get the structural parameters can be tedious.

b. Most simultaneous equations systems are over-identified.

Steps to Indirect Least Squares:

* 1. Algebraically rewrite the structural equations in reduced form.
  2. Estimate the reduced form equations using OLS.
  3. Algebraically calculate formulas for each structural parameter in terms of the reduced form parameters.
  4. Compute the formulas using the reduced form estimates.

1. Two-Stage Least Squares

In fact, we can use this technique for just-identified and over-identified systems.

 Two stage least squares (2SLS or TSLS) is done in two stages:

Stage 1:

Obtain and estimate the reduced form equations using OLS. Save the fitted values for the dependent variables.

Stage 2:

Estimate the structural equations, but replace any RHS endogenous variables with their stage 1 fitted values.

Example:



Stage 1:

Estimate the reduced form equations individually by OLS and obtain the fitted values, 

Stage 2:

Replace the RHS endogenous variables with their stage 1 estimated values:



Now and  will not be correlated with *u*1, will not be correlated with *u*2 , and will not be correlated with *u*3 .

If the disturbances in the structural equations are autocorrelated, the 2SLS estimator is not even consistent.

The standard error estimates also need to be modified compared with their OLS counterparts, but once this has been done, we can use the usual *t*- and *F*-tests to test hypotheses about the structural form coefficients.

1. Instrumental Variable Method

Recall that the reason we cannot use OLS directly on the structural equations is that the endogenous variables are correlated with the errors. One solution to this would be not to use *Y*2 or *Y*3 , but rather to use some other variables instead. We want these other variables to be (highly) correlated with *Y*2 and *Y*3, but not correlated with the errors - they are called instruments. Say we found suitable instruments for *Y*2 and *Y*3, *z*2 and *z*3 respectively. We do not use the instruments directly, but run regressions of the form



Obtain the fitted values from the above model and replace *Y*2 and *Y*3 with  and  respectively in to structural equation.

If the instruments are the variables in the reduced form equations, then IV is equivalent to 2SLS. If we use IV and 2SLS methods unnecessarily, the coefficient estimates will still be consistent, but will be inefficient compared to those that just used OLS directly. But the most important limitation of IV method is identification of correct instruments.

System Method of Estimation

Three-Stage Least Squares (3SLS) Estimator

The I3SLS estimator involves the following 3 stage procedure.

1. Same as 2SLS.
2. Same as 2SLS.
3. Apply the SUR estimator.

Properties of the 3SLS Estimator

If the error term is correlated with one or more explanatory variables, then the 3SLS estimator is biased in small samples. However it is consistent and asymptotically more efficient than single equation estimators. Thus, it has desirable large sample properties.

**References**

1. Damodar N Gujarati, and Dawn C Porter, Basic Econometrics Fifth edition, McGraw Hill, 2009
2. A.H. Studenmund, Using Econometrics – A Practical Guide, Pearson education, 2017.