Course Name: *Econometrics II*

Course code (Econ 2062)

A TEACHING MATERIAL FOR DISTANCE STUDENTS

Prepared By:

Gelagay Yeneneh

Mulatu Wondem

Teklemariam Tebabal

Department of Economics

College of Business and Economics

Debre Markos University

January, 2019

INTRODUCTION TO THE COURSE

The course econometrics II (Econ 2062) is a continuation of *Econometrics I*. It is designed principally to make students of economics familiar with the basics of the theory (and practice) of regression on qualitative information, time series and panel data econometrics as well as simultaneous equation modeling. It first makes an introduction to the basic concepts in qualitative information modeling such as dummy variable regression and binary choice models (LPM, Logit and Probit). Elementary time series models, estimations and tests for both stationary and non-stationary data will then be discussed. It also covers introduction to simultaneous equation modeling with alternative estimation methods. Introductory pooled cross-sectional and panel data models will finally be highlighted.

This course requires that the student be acquainted with the basic principles of macro and micro economics as some illustrations involve economic applications.

Course Objectives:

After the completion of this course, learners are expected to:

- Understand the basic concepts in regression involving dummy independent and dependent variables;
- Know the theory and practice of elementary time series econometrics;
- Understand the motivation and estimation methods of simultaneous equation modeling;
- Get introductory ideas on linear panel data models; and

Symbols

Dear learner! There are Symbols in this module to guide you in your study. Therefore, please use them properly.

 This indicates there is an objective to a section.

 This indicates there is a question to answer or think about in the text.

 This indicates to note and remember important points.

 This indicates there is a checklist of the main points.

 This indicates there is a self-check test for you to do.

 This indicates these are the answers to the activities and self-test Questions

Table of contents

Table of Contents

Chapter One

Regression Analysis with Qualitative Information: Binary or Dummy Variables

In this chapter you will be introduced with an important concept known as binary or dummy variable and non-linear regression. The need to deal with dummy variables and non-linear models is because we use them most often for specification and estimation of models involving of qualitative information.

The chapter starts by introducing the meaning of qualitative information or dummy variables. Section 1.2 presents the analysis of variance and analysis of covariance models. The distinction between the two depends of whether there are only dummy independent variables (Analysis of variance) or there are both qualitative and quantitative regressors (analysis of covariance models) Section 1.3 is devoted to the concept of qualitative dependent variable models. Three models are discussed in this chapter. These are the linear probability model (LPM), the logit model and the probit model. The last two have quite similar probability density function and similar probability estimations and used interchangeably quite often.

In addition, the truncated (censored) model – Tobit model (named after the Nobel price winner economist James Baumol Tobin) – is used before the end of the chapter.

Objective of the chapter

At the end of the chapter you are expected to:

- Know and define what qualitative information or dummy variable is,
- Understand how dummy/qualitative independent variables can be incorporated in econometric model
- Understand the purpose and use of slope dummy and intercept dummy
- Understand how interactions between a dummy and a qualitative independent variable can be incorporated in econometric model
- Know when and how to use binary response model
- undertake mathematical manipulations of logit and probit model

- Distinguish between linear probability model, logit model and probit model
- Be able to calculate marginal effect in logit model and probit model
- Understand extensions of logit and probit models (multinomial logit and multinomial probit, ordered logit and ordered probit)
- Know when to apply a Tobit model

What is qualitative information?

Have you ever heard of qualitative information and dummy variable? If yes, try to answer what does these mean? ----------
---

---.

If you try your own, read the following.

1.1. Describing Qualitative Information

In dealing with variables, there are four categories namely: ratio scale variable, ordinal scale variable, interval scale variable and nominal scale variable.

Ratio scale variables: these are variables that are ***quantitative*** and can be divided, subtracted and ordered for comparison. E.g. measures of income, consumption, wage, GDP, supply, price, etc. It is meaningful to ask how big this year's GDP is compared with the previous year's GDP.

Interval scale variable: If we are given different values of an interval scale variable, subtraction between any two values of a variable, ordering of values of a variable – such as for comparison – can give meaningful result, but dividing one value by another value is meaningless.

Ordinal scale variable: The values of an ordinal scale variable are measured using a certain categorical order, such as schooling (less than 8 years, 8 to 11 years, 12 years, and over 12 years. They can be ordered for comparison, but cannot be divided, nor subtracted.

Nominal scale variables: Values of a nominal scale variable cannot be divided, nor subtracted, nor ordered for comparison. Regressions involving ratio scale variables are covered in Econometrics I. Yet, using ratio scale variables is not the end of the story in regression analysis.

We may want to find the effect of nominal scale variables – such as variables which are ***qualitative*** in nature like gender, race, color, religion, nationality, geographical location, change in government policy, devaluation, war, draught, election, etc. – on the variable of our interest.

For example, we may be interested formulating an econometric model to deal with the following issues:

- ✓ Estimating the effect of ***gender***, and/or ***color*** on ***earning***? That is, we may want to answer a question; "Is there difference in average earning between *men* and *women*, or between the *white* and the *black* people"?
- ✓ Estimating the effect of ***gender, place of residence*** on ***consumption***? For example we may want to answer, "Is there difference in average consumption expenditure between *men* and *women*, or between *urban* and *rural* dwellers?"
- ✓ Measuring the effect of ***policy change***, for example *devaluation* of Birr, on Ethiopian balance of payments?
- ✓ What is the effect of ***war, or drought, natural hazard, etc.*** on a nation's ***GDP***?
- ✓ Etc.

Although the variables *gender, color, place of residence, de*valuation, *drought, natural hazard*, *war*, *etc*. may all significantly affect different economic variables, they are not quantitatively measurable. Hence, we need to have a mechanism of quantitatively analyzing the effect of such variables.

Consider another case, in which we model consumption as follows:

$$Consumption = \alpha_0 + \alpha_1 DispIncome + \alpha_2 FamilySize + \alpha_3 \boldsymbol{Gender} + u_i .. (1.1)$$

Consumption, DisposableIncome and FamilySize are measurable quantitatively; for example, a household's consumption could be 5000 birr per month, or 300 birr per day, its family size of could be 6, or 3, or 1, etc.

But, the variable "Gender" can only assume string of characters of "Male" and "Female". Basically, qualitative variables, such as gender, indicate only the *presence* or *absence* of an attribute. In other words, such types of qualitative variables indicate to which *category* a given observation is grouped to. This kind of variables is known *as* ***dummy variable***. For example, in equation (1.1) observations or individuals for whom we model consumption are grouped either to male, or to female category. But, how much is "Male"? Or, how much is "Female"? Can we measure such strings quantitatively? Obviously no! So, what should we do to estimate the impact of gender on consumption?

Fortunately, there are methods of "quantifying" such type of qualitative variables – by using *artificial variables* – which involves assigning binary numbers of 0 and 1 arbitrarily.

Variables that assume such 0 and 1 values are called *dummy variables* or *dichotomous variables* or *binary variables or categorical variables or indicator variables.*

In general, regression analysis with qualitative information tries to address such types of issues. In effect, qualitative variables can be introduced in a model as a dependent variable, or as independent variables, or both which are discussed in this chapter.

1.2. Dummy as Independent Variables

As a matter of fact, dummy variables can be incorporated in regression models just as easily as quantitative variables. A regression model may contain regressors that are all exclusively dummy, or qualitative, in nature. Such models are called *Analysis of Variance* **(*ANOVA*)** *models*. The significance of the difference between the means of two samples can be judged through either *z*-test or the *t*-test. But, when we want to examine the significance of the difference amongst more than two sample means at the same time, the ANOVA technique enables us to perform this simultaneous test.

On the other hand, regression models containing a mixture of quantitative and qualitative variables are called *analysis of covariance* **(ANCOVA)** models. The interpretation of dummy variables remains the same in both the **ANCOVA** and **ANOVA** models.

1.2.1. Regression with only qualitative regressors: The ANOVA Models

Consider the consumption model for a hypothetical town below where consumption is a function of only dummy variable gender having two ***categories or classes***: "Male" and "Female".

$$C_i = \alpha_0 + \alpha_1 D_i + u_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.2)$$

Where, C_i average monthly consumption

$D_i = 1$, if a person is male

$= 0$, Otherwise (i.e. if gender of the person is not male)

u_i = the error term satisfying the usual assumptions of classical linear regression model.

By doing so, equation (1.2) enables us to find out whether gender creates difference in average consumption among individuals, citrus paribus. Since all regressors are dummies, the intercept is equal to the expected value and the estimated model will be horizontal; hence, we can find the mean values of equation (1.2) for two different values of Di as follow:

- ✓ To estimate average consumption (intercept) of male people, use $D_i = 1$

$$E(C_i|D_i = 1) = E(\alpha_0 + \alpha_1 * 1 + u_i)$$

$$\alpha_0 + \alpha_1. \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.3)$$

- ✓ To estimate the average consumption (intercept) of female people, use $D_i = 0$

$$(C_i|D_i = 0) = E(\alpha_0 + \alpha_1 * 0 + u_i)$$

$$\alpha_0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.4)$$

Note that the difference between equation (1.3) and equation (1.4) equals α_1. This α_1 measures difference in average consumption between male and female people. That means:

- If the estimator of α_1 is **positive** and statistically significant, average consumption of male people **exceeds** average consumption of female people by the amount equal to α_1.
- On the other hand, if the estimator of α_1 is **negative** and statistically significant, it means average consumption of female people exceeds average consumption of male people by the amount equal to the estimator of α_1.
- If the estimator of α_1 is statistically insignificant, average consumption of male people does **not** have statistically significant difference with average consumption of female people.

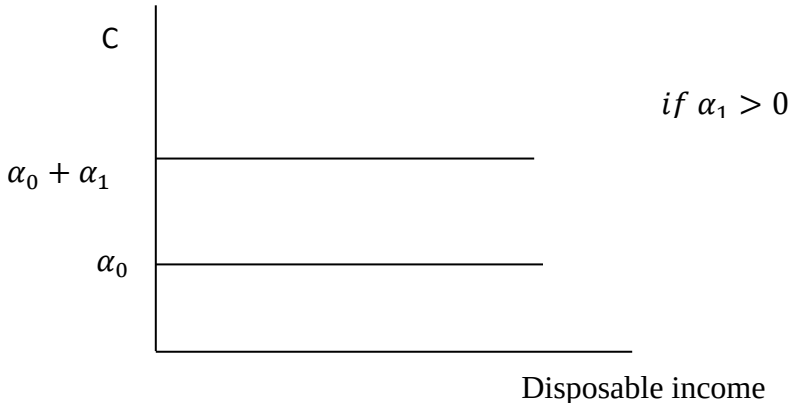


Figure 1.1: Average consumption as shown by dummy regressors

Remember equation (1.2) once again:

$$C_i = \alpha_0 + \alpha_1 D_i + u_i$$

$D_i = 1$, if a person is male

$= 0$, Otherwise

In the above case testing the existence of difference in average consumption between males and females requires setting the null and alternative hypothesis

If we assume that male people have higher consumption that female people, the hypothesis is as follows.

$$H_0: \alpha_1 = 0$$

$$H_1: \alpha_1 > 0$$

If, on the other hand, we assume that female people have higher consumption that male people, the hypothesis is as follows:

$H_0: \alpha_1 = 0$

$H_1: \alpha_1 < 0$ (if we assume that females have higher consumption)

Another thing worth mentioning here is that ANOVA models can have more than one dummy variables (see equation (1.6) below).

Example 1.1

Suppose estimating equation (1.2) for two different samples give the following, where values in braces are standard errors.

A. $\hat{C}_i = \underbrace{1500}_{(170.0)} - \underbrace{200}_{(13.5)} D_i$

B. $\hat{C}_i = \underbrace{3000}_{(880.0)} - \underbrace{500}_{(350.0)} D_i$

i. Find the average consumption of male?
ii. Find the average consumption of female?
iii. Find the difference in average consumption of males and females?

Solution

Remember that before dealing with problems of the above type, first check whether an estimator is statistically significant or not. If not significant, an estimator is regarded as zero.

Question A: Since the estimates divided by standard errors if greater than two, both estimators are statistically significant. Hence, we can estimate average consumption of all categories.

i. Average consumption for males

$$\hat{C}_i = 1500 - 200 * 1 = 1300$$

ii. Average consumption for females

$$\hat{C}_i = 1500 - 200 * 0 = 1500$$

iii. The difference between males and females is given by the coefficient of the dummy variable and it equals 200. Or, we can find it by taking the difference between the average values of the two categories. That is, 1500 -1300 = 200.

Question B: The dummy variable is not statistically significant, because 500/350 is less than two. Hence, average consumption for both male and female categories is equal to 3000.

1.2.2. Regression with a mixture of quantitative and qualitative regressors (The ANCOVA models)

Analysis of Covariance (ANCOVA) models contain a mixture of both qualitative (dummy) and quantitative regressors. ANCOVA models, which provide a method of statistically controlling the effects of quantitative regressors called **covariates** or **control** variables, are common than ANOVA models in economics. General form of the model**:**

$$C_i = \alpha_0 + \alpha_1 Yd_i + \alpha_2 D_i + \alpha_3 (D_i Yd_i) + u_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.5)$$

Where, C_i is consumption, Yd_i is disposable income, D_i is dummy for gender where,

D_i = 1, if gender is male

= 0, otherwise

α_2 = **differential intercep**t dummy

α_3 = **differential slope coefficient** or **slope** dummy.

As far as this model is concerned, there are four possibilities of modeling it.

A. Coincident Regressions: This is the case where both the intercept and the slope coefficients are the same. This means the coefficients of α_2 and α_{s3} are **insignificant** and equation (1.5) is reduced to:

$$C_i = \alpha_0 + \alpha_1 Yd_i + u_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.5a)$$

Because α_2 is insignificant both male and female have the same intercept which equals α_0 and because α_3 is insignificant both male and female have the same slope which equals α_1.

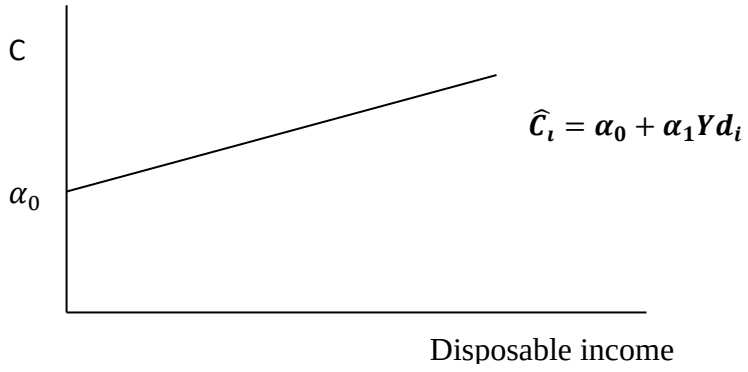


Figure 1.2: regression line with same intercepts and slope coefficients

B. Parallel Regression: This is the case where only the intercept regression is different but the slopes are the same. This means the coefficient of α_3 is **insignificant** and equation (1.5) is reduced to:

$$C_i = \alpha_0 + \alpha_1 Yd_i + \alpha_2 D_i + u_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.5b)$$

It is evident form equation (1.5b) that the model has one quantitative variable (disposable income) and one qualitative variable (gender). If individuals have the same average disposable and if the estimator for α_2 is statistically significant, we conclude that the average consumption between male and female individuals is different by the estimate equal to α_2.

For example, if α_2 is positive, graphically it is shown as:

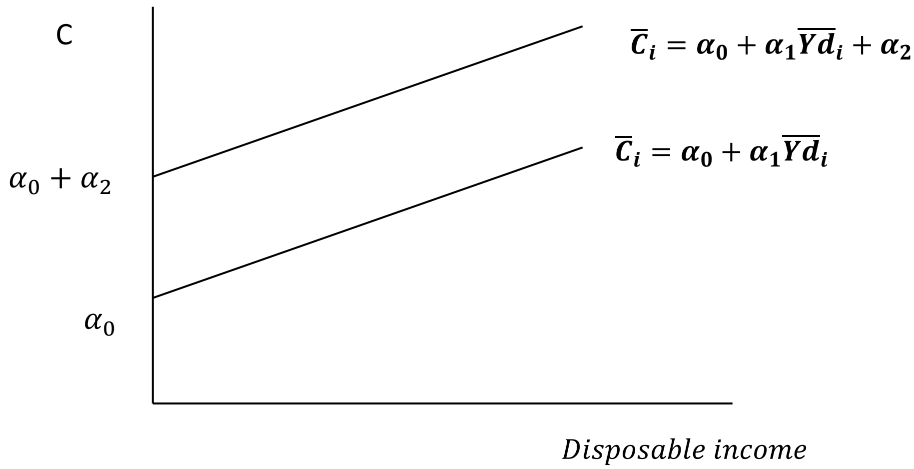


Figure 1.3: Regression line with different intercepts

As shown in figure 1.3, since the only difference is the intercept gap between the two regression lines shown above amounts to α_2 and the two regression lines are parallel.

C. Concurrent Regressions: The intercepts in the two regressions are the same, but the slopes are different. Returning, once again, to equation (1.5), the coefficient of α_2 is statistically **insignificant** and the equation is reduced to:

$$C_i = \alpha_0 + \alpha_1 Yd_i + \alpha_3 (D_i Yd_i) + u_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.5c)$$

From equation (1.5c), the intercept is equal to α_0 for both males and females, but the slope is different and it is equal to:

α_1 for females, and

$\alpha_1 + \alpha_3$ for males.

If, for example, the coefficient α_3 is negative and statistically significant then the slope of the consumption function is flatter for males than for females as shown under figure 1.4.

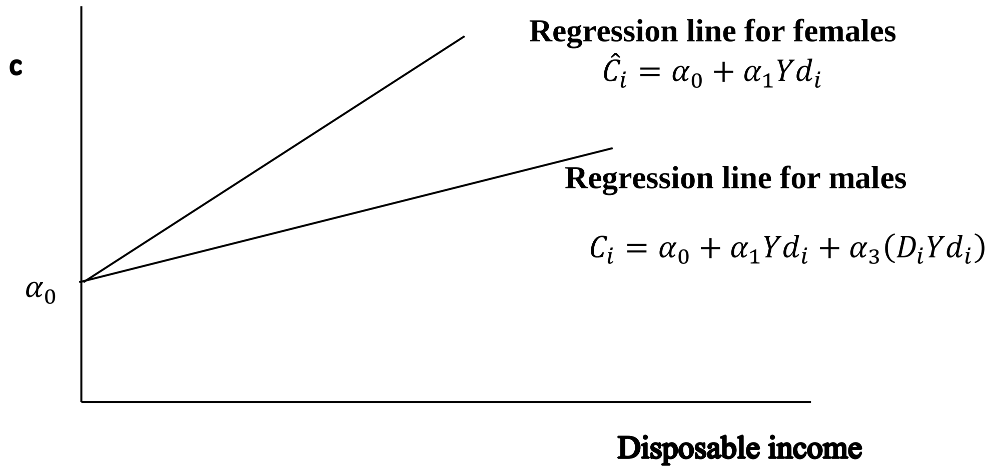


Figure *1.4: Regression line with different slope coefficients*

D. Dissimilar regressions: This is the case where both the intercepts and slopes in the two regressions are different. Considering equation (1.5), all parameters are statistically significant.

$$C_i = \alpha_0 + \alpha_1 Yd_i + \alpha_2 D_i + \alpha_3(D_i Yd_i) + u_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.5d)$$

Hence,

Intercept for females = α_0 Intercept for males = $\alpha_0 + \alpha_2$

Slope form females = α_1 Slope for males =$\alpha_0 + \alpha_3$

If α_2 is less than zero and if α_3 is greater than zero and if both coefficients are statistically significant.

The graph of the consumption for males and females becomes the following.

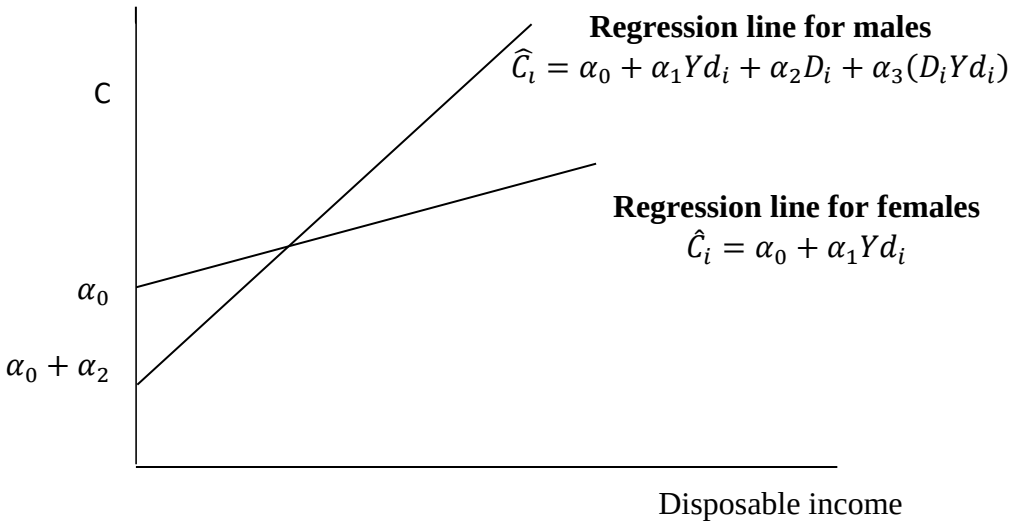


Figure *1.5: Regression line with different intercept and slope coefficients*

Note also that there are only two categories for gender: "Male" and "Female". In fact, a dummy variable can have more than two categories. For example, equation (1.5) can be extended by adding another dummy variable called "Education". Here, we incorporate four categories for education: ***illiterate***, ***primary***, ***Secondary***, and ***college and above*** shown by:

$$C_i = \alpha_0 + \alpha_1 Yd_i + \alpha_2 familySize_i + \alpha_3 D_{1i} + \alpha_4 D_{2i} + \alpha_5 D_{3i} + \alpha_6 D_{4i} + u_i \ldots\ldots (1.6)$$

$D_{1i} = 1$, if Gender is ***male***

=0, otherwise

$D_{2i} = 1$, if education is ***Primary***

=0, otherwise

$D_{3i} = 1$, if education is ***Secondary***

=0, otherwise

$D_{4i} = 1$, if education is ***College and above***

=0, otherwise

Note from equation (1.6) that:

- ✓ The assumption underlying equation(1.6) is that it is only in the intercept that changes for each group, but not the slope coefficients.
- ✓ Even though gender has two categories (***Female*** and ***Male),*** we ignore "***Female***" and considered only "***Male***". Also, we identified four categories for education (***illiterate***, ***primary***, ***Secondary***, and ***college and above***), but we ignore the category "***illiterate***", and incorporate only three categories.

  This is because if we incorporate all categories of a dummy variable, it results an exact linear relationship among regressors. This is known as **dummy variable trap.** If there is dummy variable trap, **perfect multicollinearity** problem arises and remember that we can't estimate the model under perfect multicollinearity, unless we drop at least the intercept in which case coefficients of each category becomes its own intercept. Therefore, if a qualitative variable has ***m*** categories, you have to introduce only (***m−1***) dummy variables.

- ✓ No dummy variables are assigned for the categories "***female***" and "***Illiterate***". Such categories for which no dummy variable is assigned are called the *base*, or *benchmark,* or *control*, or *comparison*, or *reference*, or *omitted* category. For example, in equation (1.2) "Female" is the benchmark since the category "Female" is omitted, from there. Also, in equation (1.6) the categories "***female***" and "***Illiterate***" are omitted, so the benchmark is the combination of illiterate and female which can be read as: ***illiterate people who are females***.
- ✓ The intercept term for each individual is obtained by substituting the appropriate values for D_1 through D_4. For example, for a male, with secondary education, we have, $D_1 = 1, D_2 = 0, D_3 = 1, D_4 = 0$

  Hence, the intercept is:

$$\alpha_0 + \alpha_3 * 1 + \alpha_4 * 0 + \alpha_5 * 1 + \alpha_6 * 0$$

$$= \alpha_0 + \alpha_3 + \alpha_5$$

Similarly, for females, with no education at all (illiterate), the intercept is:

$$\alpha_0 + \alpha_3 * 0 + \alpha_4 * 0 + \alpha_5 * 0 + \alpha_6 * 0$$

$$= \alpha_0$$

- ✓ You may choose any category as a benchmark, yet all comparisons are made in relation to the benchmark category. For example, in equation (1.6), the intercept of illiterate female people (=the bench mark), as obtained above, equals α_0.Whereas, the intercept of male who completed secondary education, equals $= \alpha_0 + \alpha_3 + \alpha_5$. Thus, keeping the effect of all other variables constant, the difference between the average consumption of male people who completed ***secondary*** education and the average consumption of ***illiterate*** **female** people (= the benchmark) equals to $\alpha_3 + \alpha_5$.

In addition, if the coefficient of the category for ***secondary*** education (= D_{3i}) which equals α_5 is *negative*, it means keeping the effect of all other variables constant, the average consumption of people (both male and female) who completed ***secondary*** education is *less than* the average consumption of ***illiterate*** **female** people (= the benchmark) by the amount equal to α_5.

The coefficients attached to each dummy variable are known as *the differential intercept coefficients*[1]**.** It tells by how much the value of the intercept term of the category to which the binary 1 is assigned differs from the intercept coefficient of the base category.

[1] This holds if the dummy is an intercept dummy. However, a dummy can be slope dummy or both slope and

Example: 1.2

Suppose estimation of equation (1.6) yields the following result. *Values in braces are standard errors.*

$$C_i = \underbrace{240}_{(16.0)} + \underbrace{0.4}_{(0.024)} Yd_i + \underbrace{160}_{((18.50)} familySize_i - \underbrace{60}_{(8.40)} D_{1i} + \underbrace{25}_{(2.42)} D_{2i} + \underbrace{80}_{(7.60)} D_{3i} + \underbrace{120}_{(12.40)} D_{4i}$$

Where, $E(Yd_i) = 500$, and $E(familySize_i) = 4$

1. What does 240 measure?
2. Compare the average consumption between male and females?
3. What is the difference in average consumption between primary education completed and secondary education completed people?
4. Interpret the coefficients of D_{3i}, D_{4i}, etc?
5. The average consumption of male people who completed primary school?

Solution

We can check that all coefficients are statistically significant. Hence, we can proceed to the answers directly .

1. 240 is the intercept for the bench mark. Therefore, the intercept for illiterate female people.
2. This is shown by the gender dummy which equals 60.
3. It equals the difference between the two categories. That is, 80-25=55
4. The coefficient of D3 equals 80. Since D3 represents the category "**secondary**", it means the average consumption of secondary school completed people is greater than the average consumption of illiterate female people (the bench mark) by 60.
   Similarly, the coefficient of D3 shows, the average consumption of college and above completed people is greater than the average consumption of illiterate female people (the bench mark) by 120.
5. Wait for a moment!

1.2.3. Interaction effects using dummy variables

The answers for questions (1)-(4) of example (1.2) can be answered without any problem. But, the answer to question (5) is a little bit different. Since it contains two categories ***male*** and ***primary***-school-completed, the answer is obtained by using a different method which is discussed below.

Look once again equation (1.6):

$$C_i = \alpha_0 + \alpha_1 YD_i + \alpha_2 familysize_i + \alpha_3 D_{1i} + \alpha_4 D_{2i} + \alpha_5 D_{3i} + \alpha_6 D_{4i} + u_i.$$

This model assumes that the *differential* effect of the gender dummy, D_{1i}, is constant across the four categories of education. Suppose the estimate of α_3 is ***negative***. This means the mean consumption of females is greater than males irrespective of the level of education of the latter. Succinctly speaking, it means

- the mean consumption of ***illiterate-females*** is *greater* than the mean consumption of ***illiterate-males*** by the amount equal to the estimator of $\boldsymbol{\alpha_3}$,
- the mean consumption of ***illiterate-females*** is *greater* than the mean consumption of ***primary-school-completed males*** by the amount equal to the estimator of $\boldsymbol{\alpha_3}$,
- the mean consumption of ***illiterate females*** is *greater* than the mean consumption of ***secondary-school-completed males*** by the amount equal to the estimator of $\boldsymbol{\alpha_3}$,
- the mean consumption of ***illiterate females*** is *greater* than the mean consumption of ***college and above -completed males*** by the amount equal to the estimator of $\boldsymbol{\alpha_3}$,

But, in many cases such an assumption may be invalid. For example, even if the mean consumption of ***illiterate-females*** is *greater* than the mean consumption of ***illiterate-males,*** it is more probable that the mean consumption of ***illiterate females*** to be *less than* the mean consumption of ***males*** who completed college and above as education favors for higher wage. Such types of occasions are accounted by using **interaction** between the coefficients of dummies.

Look at the following:

$$C_i = \alpha_0 + \alpha_1 Yd_i + \alpha_2 famSize_i + \alpha_3 D_{1i} + \alpha_4 D_{2i} + \alpha_5 D_{3i} + \alpha_6 D_{4i} + \alpha_7 D_{1i} D_{2i} + u_i \ldots (1.7)$$

Here, look the inclusion of the term "$\alpha_7 D_{1i} D_{2i}$" which is the product (***interaction***) of dummies D_{1i} and D_{2i}, and α_7 is the coefficient of interactions. This allows the gender to depend on level of education, just as it did in equation. It allows us to easily test the null hypothesis that the gender differential does not depend on education level.

Note form equation (1.6) that D_{1i} and D_{2i} represent dummies for ***male*** and ***primary*** education respectively. Equation (1.7), thus, gives the mean consumption function of male people who completed primary education. Observe that,

$$E(C_i | D_{1i} = 1, D_{2i} = 1, Yd_i, famSize_i) = \alpha_0 + \alpha_1 Yd_i + \alpha_2 famSize_{2i} + \alpha_3 + \alpha_4 + \alpha_7$$

α_3= differential effect of being a male person.

α_4= differential effect of being primary-education-completed person.

α_7 = differential effect of being primary-education-completed-male person.

Example 1.3

By working on equation (1.7)

i. Construct an equation showing the mean consumption of ***male*** people who completed **secondary** education?

Solution

This can be done by adding an interaction variable between male and secondary school completed people such as using the following.

$$C_i = \alpha_0 + \alpha_1 Yd_i + \alpha_2 famSize_i + \alpha_3 D_{1i} + \alpha_4 D_{2i} + \alpha_5 D_{3i} + \alpha_6 D_{4i} + \alpha_7 D_{1i} D_{2i} + \alpha_8 D_{1i} D_{3i} + u_i$$

Here, α_8 is the coefficient of the interaction between male and secondary school completed people

1.3. Dummy as Dependent Variable

So far in this chapter, we discussed models with quantitative dependent variables having qualitative or/and quantitative predictors. But, what if the dependent variable itself is qualitative?

Suppose you want to model "*Factors affecting house ownership of people*?" Here, the dependent variable is "ownership of a house" and the independent variables will be variables like income, family wealth, education, age, etc. If so, when you collect data about the ***dependent variable***, you will ask your samples, "***Do you have your own house?***", and the respondent's answer is either "Yes, I have", or "No, I don't have". Something comes strange here as the dependent variable is qualitative, and the "Yes" or "No" answers cannot be expressed quantitatively.

Nonetheless, if the ***dependent*** variable of the model is ***dummy***, the usual OLS technique will no more be useful. Instead, the ***maximum likelihood estimation*** technique is used. This is because when the dependent variable is dummy, the objective is finding the ***maximum probability (***whence the name ***maximum likelihood*** is derived) of something happening for the given values of regressors, and qualitative response regression models are often known as ***probability models*** *ipso facto.* If the *dummy dependent variable* has exactly two categories, it is called **binary, or dichotomous,** variable**.** Otherwise, it is called ***polychotomous*** *variable***.** Basically, three approaches to developing a probability model for a *binary* response variable are discussed in this chapter. These are:

i. The ***Linear Probability Model*** (LPM),
ii. The ***Logit*** model,
iii. The ***Probit*** Model.

1.3.1. The Linear Probability Model (LPM)

Consider a model on determinants of house ownership:

$$Y_i = \alpha_0 + \alpha_1 X_i + \alpha_2 D_i + u_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.8)$$

Where $Y_i =$ 1 if a person owns a house

$=0$ if he/she does not own a house

$X_i =$ is family income

$D_i = 1$ if the person is male, and 0 otherwise

Equation (1.8) looks like a linear regression model but because the regressand (Y_i) is binary, or dichotomous, it is called a linear probability model (LPM**).** The conditional expectation of Y_i given X_i and D_i, $E(Y_i|X_i, D_i)$, can be interpreted as the *conditional probability* that the event will ***occur*** given X_i and D_i, that is, $Pr(Y_i = 1|X_i, D_i)$. Thus, in our example, $E(Y_i|X_i, D_i)$ gives the probability of a person owning a house and whose income is given by X_i and whose gender is identified by D_i . Equation (1.8) can then be estimated using OLS although this has drawbacks which will bring us to either Logit or Probit models.

If we take the expected value of equation (1.8), we get

$$E(Y_i|X_i, D_i = 1) = \alpha_0 + \alpha_1 X_i + \alpha_2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.9)$$

Suppose the probability that the event will occur, that is,($Y_i = 1$) equals, P_i

Then, the probability that the event does not occur ($Y_i = 0$)equals $1 - P_i$.

(Remember this from the property of Bernoulli process of your statistics course).

Taking the expected value

$$E(Y_i|X_i, D_i) = E(\alpha_0 + \alpha_1 X_i + \alpha_2)$$

$$= 1(P_i) + 0(1 - P_i) = P_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.10)$$

Since the probability P_i must lie between 0 and 1, we have the restriction on the conditional expectation that

$$0 \leq E(Y_i|X_i, D_i) \leq 1 \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots . \ldots \ldots . (1.11)$$

1.3.1.1. **Drawbacks of LPM**

i. Non-Normality of the Disturbances

Since the dependent variable Y_i assumes only two values *(0 or 1)*, the disturbances u_i also takes only two values; that is, the error term follows the Bernoulli distribution. As a result, u_i is not normally distributed.

ii. The Variances of the Disturbances is Heteroscedastic

Remember that for a Bernoulli distribution that the theoretical mean and variance are, p and $p(1 - p)$, respectively where p is the probability of success (i.e., something happening), showing that the variance is a function of the mean. Thus, the error variance is heteroscedastic. But,

$$p_i = E(Y_i|X_i, D_i = 1) = \alpha_0 + \alpha_1 X_i + \alpha_2 \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots . (1.12)$$

Since, p_i itself depends on the regressors, $Var(u_i)$ also depends on regressors. Remember remedial measures if variances are heteroscedastic.

iii. The restriction is not fulfilled

OLS estimation of the LPM gives no guarantee for the probability to be between 0 and 1. This is because the probability increases linearly with regressors. In fact, we can restrict the LPM under OLS to be between 0 and 1 or use estimation techniques other than OLS that guarantee equation (1.11). (See figure 1.6)

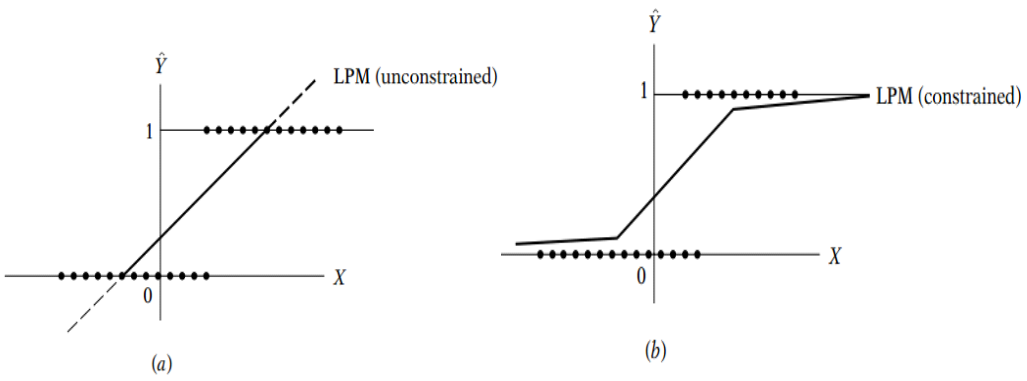


Figure 1.6: Linear probability models

iv. R^2 as a Measure of Goodness of Fit is Questionable

Corresponding to the value of regressors (X's), *the dependent variable (Y)* is either 0 or 1. Therefore, all the *Y* values will either lie along the *X* axis or along the line corresponding to Y equals 1. Therefore, generally no LPM is expected to fit such a scatter well. As a result, computed R^2 is of limited value in the dichotomous response models or in qualitative dependent variables be it constrained or unconstrained.

Example 1.4:

Suppose estimating equation (1.8) yields the following (values in brackets are Standard errors).

$$\hat{Y}_i = \underbrace{-0.084}_{(0.021)} + \underbrace{0.0017}_{(0.0004)} Income_i + \underbrace{0.008}_{(0.0032)} D_i$$

$$t = (-4.00) \quad (4.25) \quad (2.5) \qquad R^2 = 0.882$$

i. Interpret the intercept?

ii. Interpret the coefficient of X_i?

iii. Interpret the coefficient of D_i?

Solution

All t-values (coefficient divided by standard error) are greater than two; hence, all coefficients are statistically significant.

i. The intercept of $\hat{Y}_i$ gives the "probability'' that a female person (since $D_i = 0$ for female) with zero income will own a house. This value is negative; but since probability cannot be negative, we treat this value as zero, which is sensible in the present instance.

ii. The slope value of 0.0017 means that for a unit change in income, on the average the probability of owning a house increases by 0.0017 or about 0.17 percent. Of course, given a particular level of income, we can estimate the actual probability of owning a house.

Thus, if, for example, *income* = 500, the estimated probability of owning a house is

$$\left(\hat{Y}_i \middle| X_i = 500, D_i = 1\right) = -0.084 + 0.0017 * 500 + 0.008 * 1$$

$$= 0.846 = 84.6 \text{ percent.}$$

$$\left(\hat{Y}_i \middle| X_i = 500, D_i = 0\right) = -0.084 + 0.0017 * 500 + 0.008 * 0$$

$$= 0.766 = 76.6 \text{ percent.}$$

iii. The probability of owning a house for males is greater than for females on average by 0.8 percent.

Self Test : From example 1.4 above,

i. Estimate the probability of owning a house for a person having income of 1000?

ii. Relate your answer with drawback (iii)?

1.3.2. The Logit Model and Probit Model

As we saw above, the LPM has some drawbacks. For example, look once again exercise (ii) or figure 1.6 above. Even though a probability must be between the limits of 0 and 1, we can see from f*igure 1.6* (a) that the probability could be below 0 or above 1 which is not statistically plausible. Of course, as shown on *figure 1.6(b)*, using restricted least square, the probability under LPM can be made to be inside the limits of 0 and 1. Similarly, other drawbacks can also minimized using different methods.

Yet, the LPM model assumes that $Pi = E(Y = 1 \mid X)$ increases linearly with X, that is, the marginal or incremental effect of X remains constant throughout X. This means, for example, for a unit change in birr, the probability of owning a house for an individual earning monthly income of 100 birr per month and an individual earning monthly income of 30, 000 per month is equal. But, certainly this is unrealistic as one may reasonably argue that the probability of owning a house when monthly income is too low (e.g. 100) will be close to zero. To the opposite, the probability of owning a house if monthly income is sufficiently large (e.g. 30, 000) is close to 1.

Therefore, instead on depending on the LPM, we need a (probability) model that has these two features or conditions:

- As Xi increases, $Pi = E(Y = 1 \mid X)$ increases but never steps outside the 0–1 interval,
- the relationship between Pi and Xi is nonlinear, i.e., "one which approaches 0 at slower and slower rates as Xi gets small and approaches 1 at slower and slower rates as Xi gets very large.

Geometrically, this can be portrayed as:

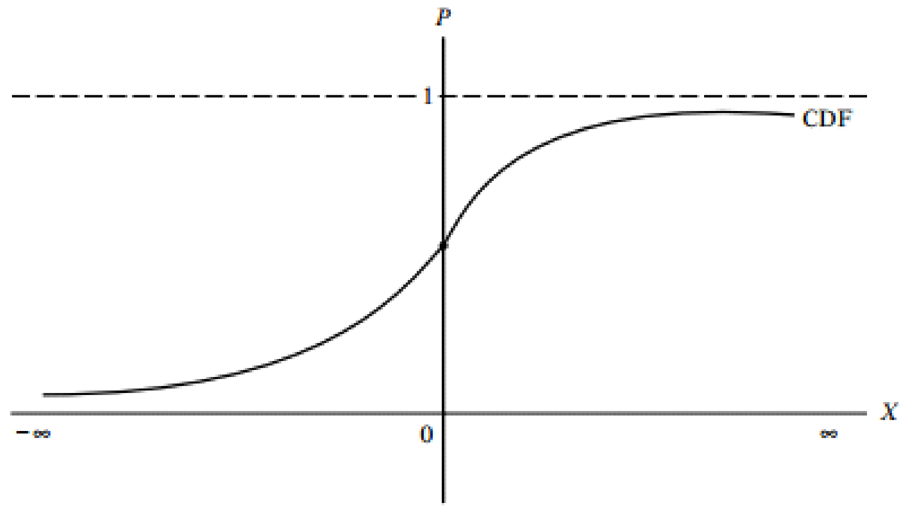


Figure 1.7: *cumulative distribution function (CDF)*

This sigmoid, or ***S****-shaped*, curve in the figure very much resembles the **cumulative distribution function** (CDF) of a random variable. Therefore, one can easily use the CDF to model regressions where the response variable is dichotomous, taking 0–1 values. For although all CDFs are ***S-shaped***, the CDFs commonly chosen to represent the 0–1 response models are (1) the logistic (**Logit)** model and (2) the normal (or **normit**) also called **(probit)** model.

1.3.2.1. The Logit Model

Consider the model for home ownership;

$$p_i = E(Y = 1|X_i) = \alpha_0 + \alpha_1 X_i \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (1.13)$$

Where, X is income and $Y = 1$ means the family owns a house.

Question:

As we said above, we need a model whose CDF will look like figure 1.7. But, what type of mathematical function can yield that shape? Consider the following model of home ownership;

$$p_i = E(Y = 1|X_i) = \frac{1}{1+e^{-(\alpha_0+\alpha_1 X_i)}} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (1.14)$$

If we let $Z_i = \alpha_0 + \alpha_1 X_i$, equation (1.14) can be re-expressed as;

$$p_i = E(Y = 1|X_i) = \frac{1}{1+e^{-Z_i}} = \frac{e^{Z_i}}{1+e^{Z_i}} \dots \dots \dots (1.15)$$

This is known as the (cumulative) **logistic distribution function** which satisfies the two conditions stated above. Since p_i gives the probability of owning a house, the probability of not owning a house, $(1 - p_i)$, is;

$$1 - p_i = \frac{1}{1+e^{Z_i}} \dots \dots \dots (1.16)$$

If we divide equation (1.15) by (1.16), we get;

$$\frac{p_i}{1-p_i} = \frac{1+e^{Z_i}}{1+e^{-Z_i}} = e^{Z_i} = e^{\alpha_0+\alpha_1 X_i} \dots \dots \dots (1.17)$$

This is called ***Odds Ratio*** in favor of owning a house. It is the ratio of the probability that a family will own a house to the probability that it will not own a house.

If we take the natural logarithm of equation (1.17), we get

$$L_i = \ln\left(\frac{p_i}{1-p_i}\right) = \ln\left(\frac{1+e^{Z_i}}{1+e^{-Z_i}}\right) = Z_i = \alpha_0 + \alpha_1 X_i \dots \dots \dots (1.18)$$

Equation (1.18) is an ***interesting*** result since it is linear in parameters *and* linear in variables, too.

L_i is called the **Logit** whence the name Logit model is derived.

Notes from equation (1.18)

- ✓ As *P* goes from 0 to 1 (i.e., as *Z* varies from −∞ to +∞), the Logit, *L*, goes from −∞ to +∞ which is unrestricted.
- ✓ Although *L* is linear in *X*, the probabilities themselves are not, unlike the LPM.
- ✓ We can incorporate as many regressors as may be dictated by the underlying theory.
- ✓ If *L*, the logit, is positive, it means that when the value of the regressor(s) increases, the odds that the regressand equals 1 (meaning some event of interest happens) increases. If *L* is negative, the odds that the regressand equals 1decreases as the value of *X* increases.

- ✓ The slope coefficients measure the change in *L* for a unit change in *X; it* tells how the log-odds in favor of owning a house change as income changes by a unit.
- ✓ The LPM assumes that *Pi* is linearly related to *Xi*, whereas the Logit model assumes that the log of the odds ratio is linearly related to *Xi*

Estimation of the Logit model

Rewriting equation (1.18);

$$L_i = \ln\left(\frac{p_i}{1-p_i}\right) = \alpha_0 + \alpha_1 X_i + u_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.19)$$

To estimate equation (1.19), we need data of X_i and Logit, L_i. This depends on the type of data we have for analysis. We distinguish two types of data:

(1). D*ata at the individual, or micro, level,* and

(2). G*rouped or replicated data.*

i. Data at the individual or micro, level

From equation (1.19), *Pi* = 1 if a family owns a house and *Pi* = 0 if it does not own a house. Thus,

$L_i = \ln\left(\frac{1}{1-1}\right) = \ln\left(\frac{1}{0}\right)$, if a person owns a house

$L_i = \ln\left(\frac{0}{1-0}\right) = \ln\left(\frac{0}{1}\right)$, if a person does not owns a house

These imply, if we have data at the micro, or individual, level, we cannot estimate (1.19) by the standard OLS nor WLS routine. In this situation, we may have to resort to non-linear estimating
procedures using the **maximum likelihood (ML)** method.

Note on estimation of individual Logit model

- ✓ Are estimated using only MLM
- ✓ SE are asymptotic hence we have to use Z statistic instead of t-statistic.
- ✓ R-square is not meaningful in binary response models.
- ✓ LR test, which is chi-square test with df equal to number of ***regressors***, in Logit is equivalent the use of F-test for joint test of multiple regression model.

ii. Grouped or Replicated Data

Here, for a given regressor, observations which have equal values are grouped together and the logit is called Grouped Logit or GLogit ipso facto. For example, in the model represented by equation (1.19), people who earn the same disposable income are grouped together. Corresponding to each group of income level X_i, there are N_i families, among whom n_i are home owners ($n_i \leq N_i$). Therefore;

$$\hat{p}_i = \frac{n_i}{N_i} \qquad (1.20)$$

gives the *relative frequency* house owners, and can be used as an estimate of the true *Pi* corresponding to each *Xi*. This true *Pi* is used to estimate the logit, $\hat{L}_i$.

Thus, $$\hat{L}_i = \ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \hat{\alpha}_0 + \hat{\alpha}_1 X_i \qquad (1.21)$$

gives fairly good estimate of the true logit, *Li*, if the number of observations *Ni* at each *Xi* is reasonably large.

But, since the variance of error term, although it could be normally distributed for large *Ni as* $u_i \sim N\left(0, \frac{1}{N_i\hat{p}_i(1-\hat{P}_i)}\right)$, is heteroscedastic, we have to use WLS instead of OLS.

Steps:

i. For each income level X, compute the probability of owning a house using equation (1.20)

ii. For each X_i, obtain the logit as $\hat{L}_i = l\,n\left(\frac{\hat{p}_i}{1-\hat{P}_i}\right)$

iii. Then to resolve the problem of heteroscedasticity, transform equation (1.19) using

$$\sqrt{w_i}L_i = \alpha_0\sqrt{w_i} + \alpha_1\sqrt{w_i}X_i + \sqrt{w_i}u_i$$

Or,

$$L^*{}_i = \alpha_0\sqrt{w_i} + \alpha_1 X^*{}_i + v_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.22)$$

Where, $w_i = N_i\hat{p}_i(1-\hat{P}_i)$, $L^*{}_i$, $X^*{}_i$ and v_i are transformed L_i, X_i and u_i respectively.

iv. Then estimate equation (1.22) using OLS. Note that is no intercept term introduced explicitly and also test of hypothesis should be made at reasonably large sample.

Example 1.5

Suppose estimating equation (1.22) yields the following result

$$L^*{}_i = \underbrace{-1.5}_{(0.250)}\sqrt{w_i} + \underbrace{0.06X^*{}_i}_{(0.016)}$$

$$t = \quad (-6.00) \qquad (3.75) \qquad\qquad R^2 = 0.9242 \ldots\ldots\ldots (1.23)$$

Where $L^*{}_i$ and$X^*{}_i$ are weighted $\hat{L}_i$and X_i of equation (1.21) respectively.

Interpret coefficients?

In both logit and probit models, we interpret the *sign* of the coefficient but not the *magnitude*. The magnitude cannot be interpreted using the coefficient because different models have different scales of coefficients.

There are various ways of interpreting estimated logit model.

i. Logit Interpretation

The coefficient of X^*_i shows that for a unit increase in weighted income, the weighted log of the odds in favor of owning a house goes up by 0.06 units.

ii. Odds Interpretation

An odds ratio of 2 means that the outcome y=1 is twice as likely as the outcome of y=0.

Since $L_i = \ln\left(\frac{p_i}{1-p_i}\right)$, its antilog gives the odds ratio which equals,

$$\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = e^{-1.5\sqrt{w_i}+0.06X^*_i} = e^{-1.5\sqrt{w_i}} * e^{0.06X^*_i} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.24)$$

Hence, $e^{0.06} = 1.0618$

This is interpreted as that for a unit increase in weighted income, the (weighted) odds in favor of owing a house increases by 1.0618 or about 6.1 percent.

Note:

If you want to carry the analysis in terms of unweighted logit, all you have to do is to divide the estimated *Li** by $\sqrt{w_i}$.

iii. Computing Probabilities

The predicted probability indicate the likelihood of y=1. If the predicted probability is greater than 0.5 we can predict that y=1, otherwise y=0

To compute the probability of owning a house, for example, at X=22, given value of $\sqrt{w_i}$ say, 4, which corresponds to the given value of X, substitute this in equation (1.24). That is,

$$L^*_i = -1.5\sqrt{w_i} + 0.06X^*_i = -1.5 * 4 + 0.06 * 22 * 4 = -0.72$$

Dividing this by 4 gives the value of $-\frac{0.72}{4} = -0.18$

Therefore, at income level of 22, we have, $ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -0.18$

Taking the antilog, $\frac{\hat{p}_i}{1-\hat{P}_i} = e^{-0.18} = 1$

Solving for $\hat{p}_i$ we get, $\hat{p}_i = \frac{e^{-0.18}}{1+e^{-0.18}} = \frac{0.835}{1+0.835} = 0.454$

This is interpreted as given the income of birr 22, the probability of a family owning a house is about 45.4 percent.

1.3.2.2. Probit Models

As it is said earlier, the probit model is usually used if the Cumulative Density Function (CDF) is normal.

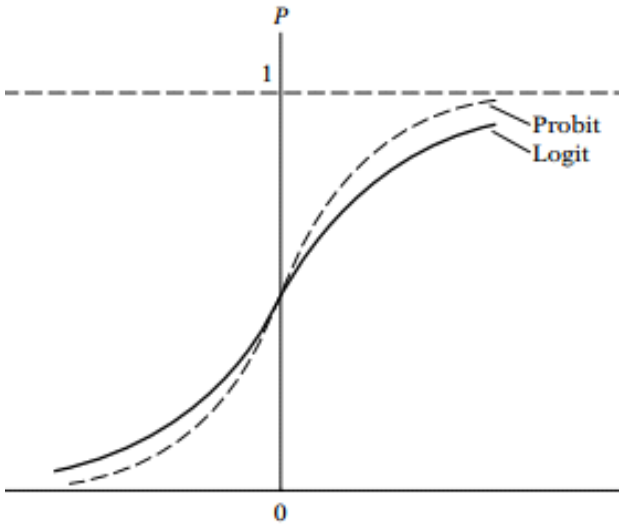


Figure1.8: Logit and probit cumulative distributions

Figure (1.8) portrays the difference between CDF of Logit and Probit models. As the value of the regressor (X) increases, the probability of occurrence of the dependent variable increases faster (gets closer to 1) under probit than under logit. Similarly, as the value of X gets smaller and smaller, the probability of the dependent variable not-to-occur decreases faster (gets closer to 0) under probit than under logit. Beyond this, they are quite similar. Yet, in principle the normal CDF can be substituted in place of the logistic CDF because of this pretty resemblance between them.

Motivation

Suppose the decision of ownership of a house by a person or family depends on the utility obtained from owning the house. Mathematically,

$$I_i = \alpha_0 + \alpha_1 X_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.25)$$

Where, I_i is *unobservable* the utility index, also called ***latent variable***, whose higher values are associated with higher probability of owning a house, X_i is income of i^th person or family.

Now, assume that there is a threshold or critical, say I_i^* which itself is unobservable, such that if I_i exceeds I_i^*, the family will own a house, otherwise it will not. If we assume further that I_i^*, like I_i, is ***normally*** distributed with the same mean and variance, it is possible not only to estimate the parameters of the index given in equation (1.25) but also to get some information about the unobservable index itself.

The probability that I_i^* is less than or equal to Ii can be computed from the standardized normal CDF as:

$$P_i = P(Y = 1 \mid X) = P(I_i^* \leq I_i) = P(Z_i \leq \alpha_0 + \alpha_1 X_i) = F(\alpha_0 + \alpha_1 X_i) \ldots\ldots\ldots (1.26)$$

Where, $P(Y = 1 \mid X)$ means the probability that an event occurs given the value(s) of the regressors, Z_i is the standard normal variable, i.e., $Z_i \sim N(0, \sigma^2$. F is the standard normal CDF, which is:

$$F(I_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{I_i} e^{\frac{-z^2}{2}} dz$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha_0 + \alpha_1 X_i} e^{\frac{-z^2}{2}} dz \ldots\ldots\ldots\ldots\ldots (1.27)$$

The area between $-\infty$ and I_i measures the probability of owning a house.

If we take the inverse of equation (1.26), we get information on I_i, α_0, α_1. That is,

$$I_i = F^{-1}(I_i) = F^{-1}(P_i) = \alpha_0 + \alpha_1 X_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.28)$$

Where, F^{-1} is the inverse of the normal CDF.

Estimation of Probit Model

Estimation of parameters and I_i depends on whether we have grouped data or ungrouped data.

Probit Estimation with Grouped Data: gprobit

As in the case of logit model, we get the relative frequency, (the empirical measure of probability) of owning a house at various income level, and I_i can be obtained from normal CDF. Obtaining I_i makes estimating parameters relatively straightforward; it is even easier than LPM and Logit.

Interpretation of the Probit Estimates

Suppose the estimated model is

Table 1.1: Probit estimation results for probability of owning a house

Variable	coefficient	Std. error	t-statistic	Probability
Consumption	-1.0166	0.0.72	-17.7473	1.0397E-07
Disp_income	0.04846	0.00247	19.5585	4.8547E-08
	R^2=0.97951	Durbin-Watson statistic =0.91384		

To find out the effect of a unit change in income measured in birr on the probability that Y = 1, i.e., a family purchases a house, we want to take the derivative of equation (1.26) with respect to *X:*

$$\frac{dP_i}{dX_i} = f(\alpha_0 + \alpha_1 X_i)\alpha_1 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.29)$$

Where, $f(\alpha_0 + \alpha_1 X_i)$ is the standard normal PDF evaluated at $\alpha_0 + \alpha_1 X_i$.

If say, X=6, we want to find NDF at $(-1.0166 + 0.04846(6)) = f(-0.72548)$. From normal distribution table, for $Z = -0.72548$, the normal density is about 0.2358. Thus,

$$= 0.2358 * 0.04846 = 0.01142$$

This means, starting with an income level of birr 6, if the income goes up by 1, the probability of a family purchasing a house goes up by about $0.0128 * 100 = 1.142$ percent

Probit Model for Ungrouped or Individual Data

The problem we face in Logit regression also presents here. For this reason, will have to use a nonlinear estimating procedure based on the method of maximum likelihood; we will go for this in lab sessions.

Marginal Effects in Logit and Probit models

Most papers report marginal effects at the mean. A problem is that there may not be such a person in the sample.

For dummy independent variables, the marginal effect is expressed in comparison to the base category (x=0). For continuous independent variables, the marginal effect is expressed for a one-unit change in x.

a) In the ***linear regression model***, the slope coefficient measures the change in the average value of the regressand for a unit change in the value of a regressor, with all other variables held constant.
b) In the ***LPM***, the slope coefficient measures directly the change in the probability of an event occurring as the result of a unit change in the value of a regressor, with the effect of all other variables held constant.
c) In the ***logit model*** the slope coefficient of a variable gives the change in the log of the odds associated with a unit change in that variable, again holding all other variables constant. Moreover, the rate of change in the probability of an event happening is given by $\alpha_i P_i(1 - P_i)$.
d) In the ***probit model***, the rate of change in the probability is given by $\alpha_i f(Z_i)$, where Z_i is the density function of the standard normal variable and $Z_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \cdots + \alpha_k X_{ki}$, that is, the regression model used in the analysis.

Thus, in both the logit and probit models all the regressors are involved in computing the changes in probability, whereas in the LPM only the j^{th} regressor is involved. This difference may be one reason for the early popularity of the LPM model

We interpret both the sign and the magnitude of the marginal effects. The probit and logit models produce almost identical marginal effects.

Further reading

The course outline for this chapter is limited to the binary choice models which we have discussed so far. Nonetheless, below you are introduced to some further concepts which you can proceed in different books if you want.

I. Extensions of the Logit and Probit Models

Earlier we say that in logit and probit models the dependent variable contains ***two*** categories and is called ***binary*** choice model ipso facto. But, they have the following varieties.

a) Ordinal logit and probit models

Usually, a qualitative dependent variable, or regressand, can have more than two outcomes and very often these outcomes are **ordinal** in nature; that is, they cannot be expressed on an interval scale. For example, suppose you want to study about "Determinants of Satisfaction of Debre Markos University Students from Library Service". Hopefully, the dependent variable-level of Satisfaction-may assume categories like, "highly satisfied", "satisfied", "dissatisfied", "highly dissatisfied". In this case, it consists of 4 categories which require some kind of ordering.

Ordinal logit and probit models are employed to model such types of phenomena.

b) Multinomial logit and probit models

Such models are similar to ordinal logit and probit models because in both cases to the dependent variable has more than two categories. The difference, however, is that ordering the categories of the dependent variable is not required, here. For example, if you want to research, "The determinants of choice of modes of transportation", the dependent variable will be the type of transport mode-which assumes one of bicycle, motorbike, car, bus, or train-chosen by a certain observation. In this case, the dependent variable does not require ordering, yet has more than two categories. Such types of phenomenon are modeled using ***multinomial logit and probit models.***

1.3.2.3. The Tobit model

This is an extension of the probit model. It was originally developed by the Nobel laureate economist James Tobin.

To explain this model, let's take our previous home ownership example,

- ✓ In the ***probit model*** our concern is estimating the probability of owning a house as a function of some socioeconomic variables.
- ✓ In the ***Tobit model***, however, our interest shifts to finding out the amount of money a person spends (measured quantitatively) on a house in relation to socioeconomic variables.

Yet, the consumer may not own a house due to two reasons. First, the consumer may not have sufficient money to purchase a house. Second, even if the consumer may have sufficient money, he/she may not want to purchase a house.

We face a dilemma here: If a consumer does not purchase a house, obviously we have no data on housing expenditure for such consumers. In other words, we have such data only on consumers who actually purchase a house.

Hence, samples from which data is collected will have the following features.

- First, consumers about whom we have information on the regressors as well as the regressand.
- Second, consumers about whom we have information only on the regressors but not on the regressand. Basically, such type of observation is known as a **censored sample.**

Mathematically,

$$Y_i = \alpha_0 + \alpha_1 X_i + u_i \qquad if\ RHS > 0 \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(1.30)$$

$$= 0 \qquad otherwise$$

Where, Y_i is expenditure on house, $RHS = right - hand\ side$

Since Y_i is expenditure on house it can be measured quantitatively and can be estimated using OLS. The problem, however, is due to the presence of people with no spending on housing (censored observations), OLS estimation of equation (1.30) by neglecting data from people who do not own a house-because there is no data for them-will make the estimators biased and inconsistent. Instead, maximum likelihood method can still be in action.

Check Lists

Analysis of covariance model

Analysis of variance model

Benchmark

Categorical variable

Dummy variable

Interaction variable

Linear probability model

Logit model

Odds ratio

Omitted category

Probit model

Qualitative information

Slope dummy intercept dummy

Tobit model /censored regression

Chapter Two

Introduction to Basic Regression Analysis with Time Series Data

From your lessons of econometrics I and the first chapter of econometrics II, you have covered the skills and methods understanding of how to use the multiple regression models for cross-sectional applications. Now, we turn to the econometric analysis of time series data. Since we will rely heavily on the method of ordinary least squares, most of the work concerning mechanics and inference has already been done. However, as you might have noted, time series data have certain characteristics that cross-sectional data do not, and these can require special attention when applying OLS to time series data.

The chapter starts by introducing the nature of time series data under section 2.1. Here, we discussed two types of data: deterministic trend and stochastic trend. Section 2.2 presents the stationary and non-stationary stochastic processes. This section covers the criteria of stationary and non-stationary time series. In addition, the two classic or common types of non-stationary data are discussed here. These are the random walk model with drift and the random walk model without drift. Trend stationary and difference stationary stochastic processes and the methods of making such stochastic processes are discussed in section 2.3. In section 2.4, we discuss the Integrated stochastic process and explored the properties of linear combinations of stationary and non-stationary stochastic processes. Under section 2.5, we present the tests of stationary, and principally introduced about the unit root test.

Objective of the chapter

At the end of the chapter you are expected to:

- Define the meaning and nature of time series data
- Differentiate between stationary and non-stationary data.
- Identify trend stationary stochastic process (TSP) and difference stationary stochastic process (DSP).
- Be able to transform TSP and DSP in to stationary
- Understand the properties of linear combination of stationary and non-stationary stochastic process.

- Perform tests of stationary using unit root test

What is time series data?

What do you know about time series data? Have you ever seen data ordered in time such as GDP, exchange rate, price of coffee, price of petroleum, etc.? If yes, try to answer what does these mean? ---
----

----------------------------------.

One of the basic points we make in econometrics is that the properties of the estimators and their usefulness for point estimation and hypothesis testing depends on how the data behave. For instance, in a linear regression model where errors are correlated with regressors, least squares won't be consistent and consequently it should not be used for either estimation or subsequent testing. In this chapter, we begin to study the properties of OLS for estimating linear regression models using time series data.

While considering the standard regression model, we did not pay attention to the timing of the explanatory variable(s) on the dependent variable. The standard linear regression implies that change in one of the explanatory variables causes a change in the dependent variable during the same time period and during that period alone. But in economics, such specification is scarcely found. In economic phenomenon, generally, a cause often produces its effect only after a lapse of time; this lapse of time (between cause and its effect) is called a lag. Therefore, realistic formulations of economic relations often require the insertion of lapped values of the explanatory or insertion of lagged dependent variables.

2.1. The nature of Time Series Data

More often than not, economists study time series data. For example, economists might study import-export time trends in Ethiopian GDP, consumption, investment, unemployment, inflation, interest rates and so on. Time series data are often the only window we have into

important economic processes. Many data are collected or analyzed only at the national level. Unfortunately, time series data hold their own challenges.

Trends: persistent upward or downward movements of variables over time. It can be very difficult to disentangle trends over time. Trends can threaten the consistency and asymptotic normality of OLS. Many macroeconomic variables have long-term trends: Real GDP per capita, real consumption per capita, Real investment per capita, Inflation (the CPI).

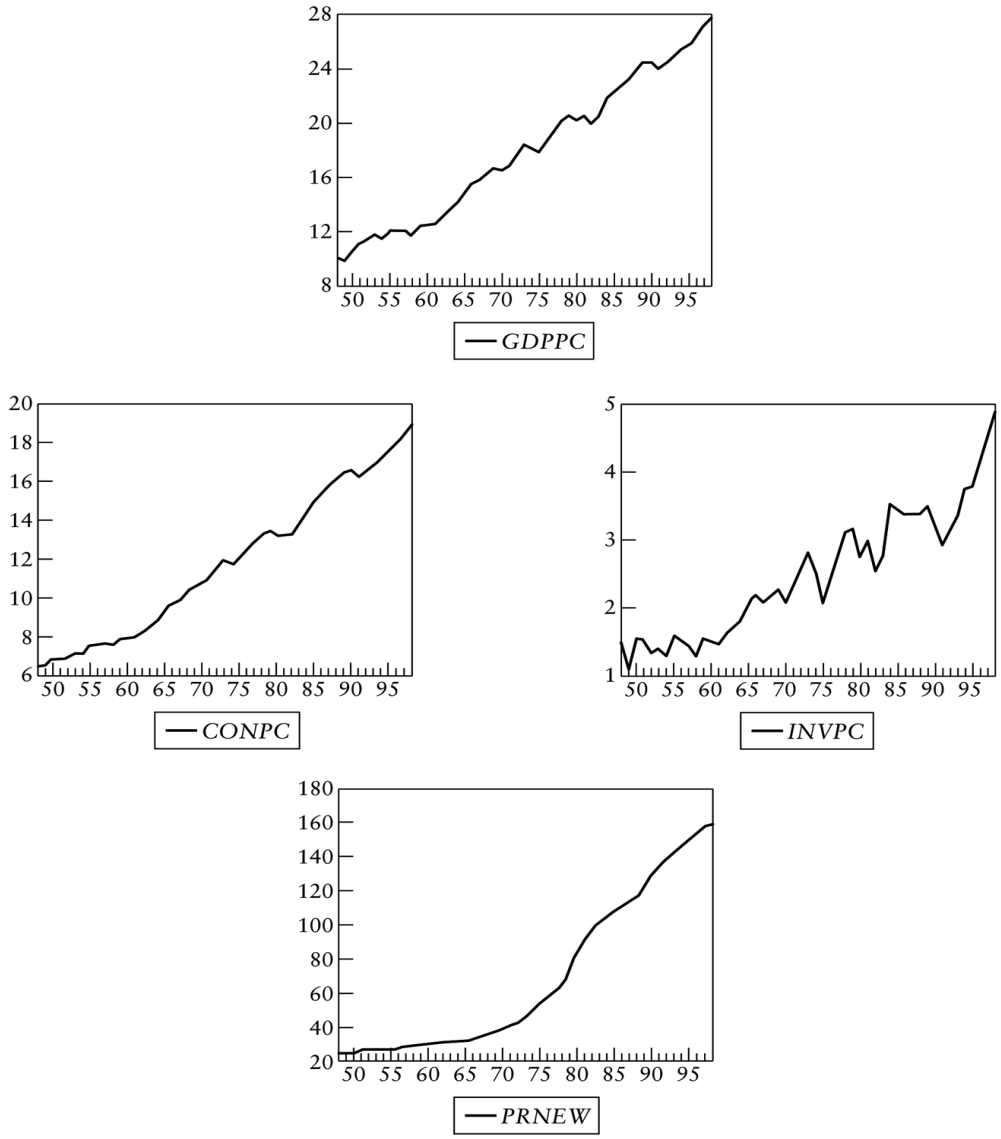


Figure 2.1: GDP, Consumption, Investment, and the consumer price index (CPI), 1948–1998

When we talk about trends, there are two common types of trends:

***Deterministic Trends*: $E(y_t) - E(y_{t-1}) = a$.** The trending variable changes by a constant amount each period

***Stochastic Trends*: $E(y_t) - E(y_{t-1}) = b + v_t$.** The trending variable changes by a random amount each period (v_t).

We will discuss these two trends in detail on the coming sections.

An obvious characteristic of time series data that distinguishes them from cross-sectional data is temporal ordering. A time series y_t is a process observed in sequence over time, $t = 1, \ldots, T$. To indicate the dependence on time, we adopt new notation, and use the subscript t to denote the individual observation, and T to denote the number of observations. Because of the sequential nature of time series, we expect that y_t and y_{t-1} to be not independent. So, classical assumptions are not valid.

For analyzing time series data, we must recognize that the past can affect the future, but not vice versa. To emphasize the proper ordering of time series data, Table 2.1 gives a partial listing of the data on U.S. inflation and unemployment rates from various editions of the *Economic Report of the President*.

Year	Inflation	Unemployment
1998	**1.6**	**4.5**
1999	**2.2**	**4.2**
2000	**3.4**	**4.0**
2001	**2.8**	**4.7**
2002	**1.6**	**5.8**
2003	**2.3**	**6.0**

Table 2.1: U.S. Inflation and Unemployment Rates, 1998 - 2003

Another difference between cross-sectional and time series data is more subtle. In Econometrics I, we studied statistical properties of the OLS estimators based on the notion

that samples were randomly drawn from the appropriate population. Understanding why cross-sectional data should be viewed as random outcomes is fairly straightforward: a different sample, drawn from the population, will generally yield different values of the independent and dependent variables (such as education, experience, wage, and so on). Therefore, the OLS estimates computed from different random samples will generally differ, and this is why we consider the OLS estimators to be random variables.

How should we think about randomness in time series data? Certainly, economic time series satisfy the intuitive requirements for being outcomes of random variables. For example, today we do not know what the Real Estate Industrial Average will be at the close of the next trading day. We do not know what the annual growth in output will be in Ethiopia during the coming year. Since the outcomes of these variables are not foreknown, they should clearly be viewed as random variables.

Formally, a sequence of random variables indexed by time is called a **stochastic Process** or a **time series process**. ("Stochastic" is a synonym for random.) A random or stochastic process is a collection of random variables ordered in time. When we collect a time series data set, we obtain one possible outcome, or *realization*, of the stochastic process.

We can only see a single realization, because we cannot go back in time and start the process all over again. (This is analogous to cross-sectional analysis where we can collect only one random sample.) However, if certain conditions in history had been different, we would generally obtain a different realization for the stochastic process, and this is why we think of time series data as the outcome of random variables. The set of all possible realizations of a time series process plays the role of the population in cross-sectional analysis. The sample size for a time series data set is the number of time periods over which we observe the variables of interest.

2.2. Stationary and non-stationary stochastic Processes

2.2.1. Stationary Stochastic Processes

A type of stochastic process that has received a great deal of attention and scrutiny by time series analysts is the so-called **stationary stochastic process.** Broadly speaking, a stochastic process is said to be stationary if its mean and variance are constant over time and the value of the covariance between the two time periods depends only on the distance or gap or lag between the two time periods and not the actual time at which the covariance is computed.

In the time series literature, such a stochastic process is known as a **weakly stationary,** or **covariance stationary,** or **second-order stationary, stochastic process.** To explain weak stationarity, let Y_t be a stochastic time series with these properties:

$$\textit{Mean:} \quad E(Y_t) = \mu \tag{2.1}$$

$$\textit{Variance:} \quad \text{var}\,(Y_t) = E(Y_t - \mu)^2 = \sigma^2 \tag{2.2}$$

$$\textit{Covariance:} \quad \gamma_k = E[(Y_t - \mu)(Y_{t+k} - \mu)] \tag{2.3}$$

where γ_k, the covariance (or autocovariance) at lag k, is the covariance between the values of Y_t and Y_{t+k}, that is, between two Y values k periods apart. If $k = 0$, we obtain γ_0, which is simply the variance of Y ($= \sigma^2$); if $k = 1$, γ_1 is the covariance between two adjacent values of Y.

If a time series is stationary, its mean, variance, and autocovariance (at various lags) remain the same no matter at what point we measure them; that is, they are time invariant. Such a time series will tend to return to its mean (called **mean reversion**) and fluctuations around this mean (measured by its variance) will have a broadly constant amplitude.

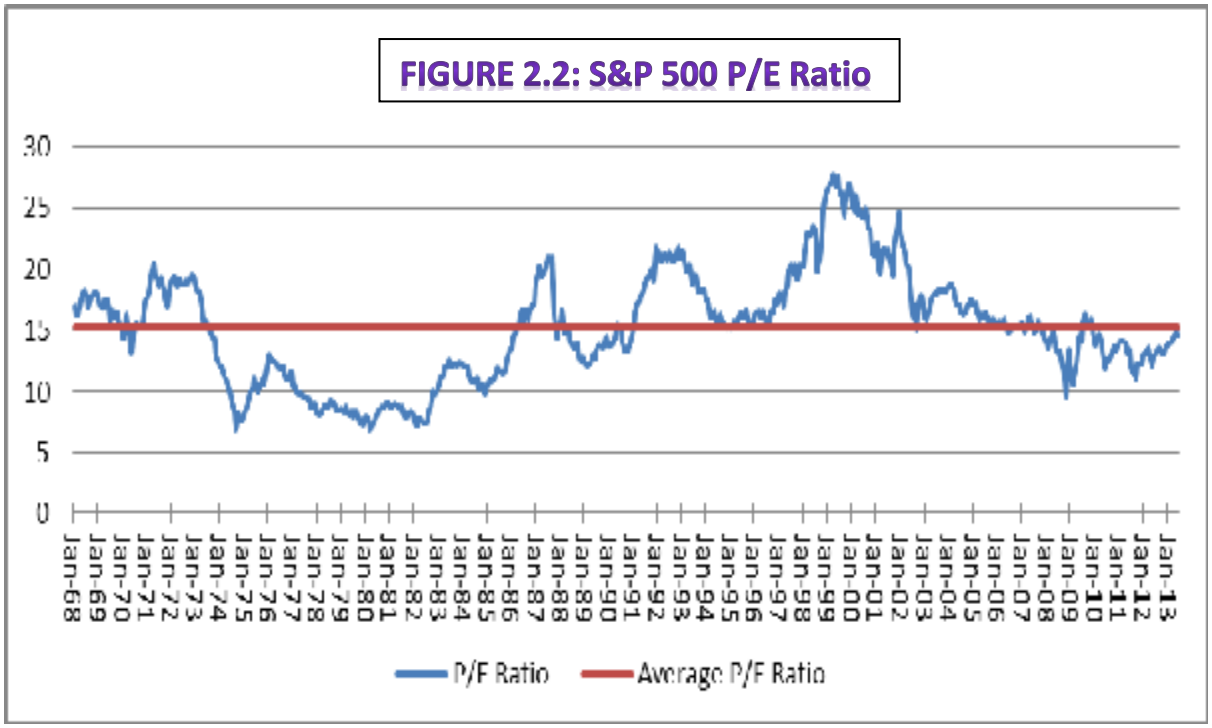
If a time series is not stationary in the sense just defined, it is called a **non-stationary time series** (keep in mind we are talking only about weak stationarity). In other words, a non-stationary time series will have a time varying mean or a time-varying variance or both.

Why are stationary time series so important? Because if a time series is non-stationary, we can study its behavior only for the time period under consideration. Each set of time series data will therefore be for a particular episode. As a consequence, it is not possible to

generalize it to other time periods. Therefore, for the purpose of forecasting, such (non-stationary) time series may be of little practical value.

A stationary time-series' statistical properties like mean & variance will be constant over time. They can (and will) move around but revert to the mean over time.

For example, **Price to Earning ratio** of a stock market index, say The Standard & Poor's 500 (often abbreviated as S&P 500 which is an American stock market index) is likely to be stationary (see figure 2.2).



2.2.2. Finite Sample Properties of Ordinary Least Squares Estimators

In analysing time series data, we need to alter some of our assumptions in the standard OLS regression to take into account the fact that we no longer have the usual random sample of individual items.

I. Linear in parameters: the stochastic process $x_{t1}, x_{t2}, \ldots, x_{tk}, y_t$: follows the linear model

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots + \beta_n x_{tk} + u_t$$

where t=1, 2, …, n; and n=the number of observations (number of time periods)

II. Zero conditional mean: for each t, the expected value of the error term (u_t), given the explanatory variables for all time periods, is zero.

$$E(u_t / x_{ti}) = E(u_t/x_{t1}, x_{t2}, \ldots, x_{tk}) = 0; \text{ where } t=1, 2, \ldots, n$$

This assumption implies that the error term at time t is uncorrelated with each explanatory variable in every time period. If u_t *i*s independent of *x's* and $E(u_t) = 0$, this assumption automatically holds.

III. No Perfect collinearity: like in cross-sectional regression, in the sample (in the underlying stochastic process), no independent variable is a perfect linear combination of another independent variable.

Theorem 1: under assumptions I, II, III, in other words, if these three assumptions are satisfied, the Ordinary Least Squares Estimators (OLSEs) are unbiased: i.e. $E(\hat{\beta}_i) = \beta_i$ *; for all* $i=0, 1, \ldots, k$

IV. Homoscedasticity: conditional on *x's*, the variance of u_t is the same for all *t*.

$$Var(u_t /x)= Var(u_t) = \sigma^2 \text{ ; where } t=1, 2, \ldots, n$$

If this does not hold true, the errors are heteroskedastic.

V. No serial correlation: conditional on *x's*, the errors in two time periods are uncorrelated.

$$Corr(u_t u_s /x) = Corr(u_t u_s) = 0; \text{ for all } t \neq s$$

Gauss-Markov Theorem: given the asumptions I through V, the OLSEs are BLUE.

Hypothesis Testing

In order to use the usual OLS standard errors in hypothesis testing, t-statistics and F-statistics, we need the normality assumption.

VI. Normality of the error term: the error u_t is independently and identically distributed as normal with is zero mean and constant variance (σ^2).

$$u_t = IIDN(0, \sigma^2)$$

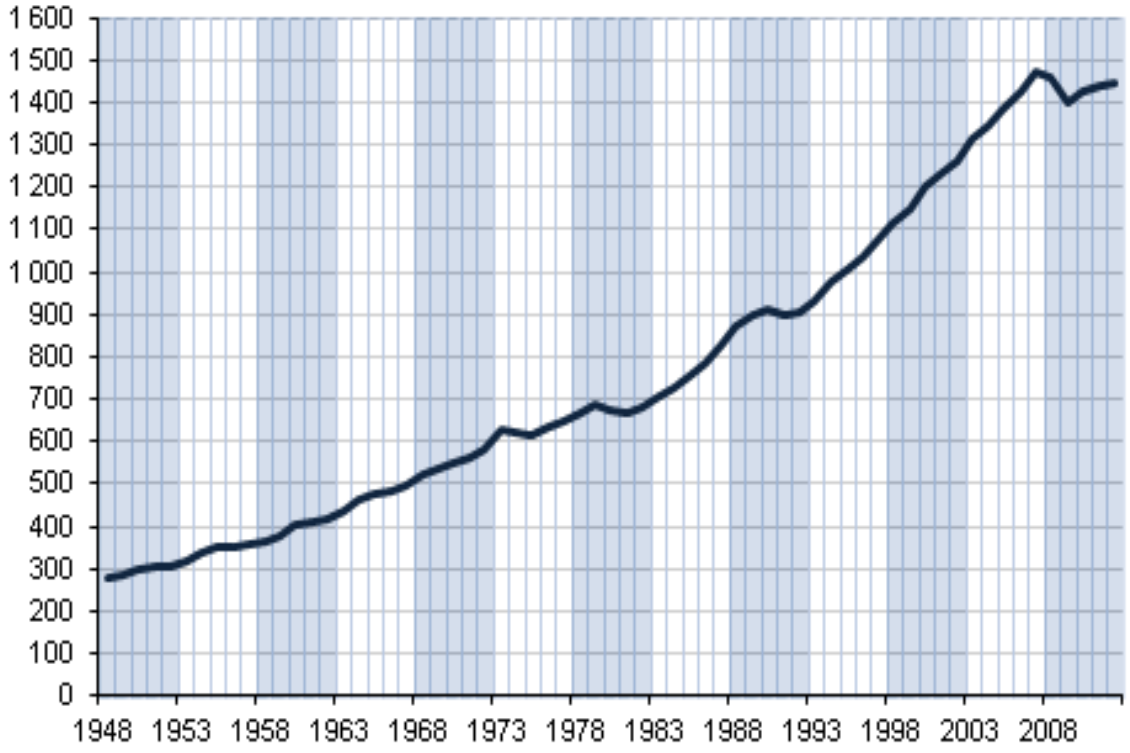
When assumptions I through VI hold true, everything that applies to estimation and inference for cross-sectional regression applies directly to time series regressions. t-statistic tests the statistical significance of individual explanatory variables; whereas F-statistic tests joint significance.

2.2.3. Non-stationary Stochastic Processes

A **non-stationary** time series' statistical properties like mean, variance etc will not be constant over time An example of a non-stationary time series is a series with a trend - something that grows over time, for instance. The sample mean and variance of such a series will grow as you increase the size of the sample.

Many economic and financial variables are non-stationary. Nominal GDP is one such. Below (in figure 2.3) is UK's GDP over the years. There is a trend as you can see.

FIGURE 2.3: Annual GDP, in billion pounds



Although our interest is in stationary time series, one often encounters non-stationary time series, the classic example being the **random walk model** (RWM). It is often said that asset prices, such as stock prices or exchange rates, follow a random walk; that is, they are non-stationary. We distinguish two types of random walks: (1) random walk without drift (i.e., no constant or intercept term) and (2) random walk with drift (i.e., a constant term is present).

Random Walk without Drift: Suppose u_t is a white noise error term with mean 0 and variance σ^2. Then the series Y_t is said to be a random walk if

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

In the above random walk model, the value of Y at time t is equal to its value at time $(t - 1)$ plus a random shock. We can think of random walk without a drift as a regression of Y at time t on its value lagged one period.

Now from $Y_t = Y_{t-1} + u_t$, we can write

$$Y_1 = Y_0 + u_1$$

$$Y_2 = Y_1 + u_2 = Y_0 + u_1 + u_2$$

$$Y_3 = Y_2 + u_3 = Y_0 + u_1 + u_2 + u_3$$

In general, if the process started at some time *0* with a value of Y_0, we have

$$Y_t = Y_0 + \sum u_t$$

$$E(Y_t) = E(Y_0 + \sum u_t)$$

$$= E(Y_0) + E(\sum u_t)$$

$$= Y_0, \text{ since } E(\sum u_t) = 0$$

When you calculate the variance, you will find that

$$Var\ (Y_t) = E(Y_t - E(Y_t))^2$$

$$= E(Y_0 + \sum u_t - Y_0)^2$$

$$= E(\sum u_t)^2$$

$$= E(u_1 + u_2 + u_3 + \ldots + u_t)^2$$

$$= E[(u_1 + u_2 + u_3 + \ldots + u_t)(u_1 + u_2 + u_3 + \ldots + u_t)]$$

$$= E[(u_1)^2 + (u_2)^2 + (u_3)^2 + \ldots + (u_t)^2 + (u_1u_2) + (u_2u_1) + (u_1u_3) + \ldots + (u_iu_j)$$

$$= E(u_1)^2 + E(u_2)^2 + E(u_3)^2 + \ldots + E(u_t)^2 + E(\sum(u_iu_j))$$

$$= \sigma^2 + \sigma^2 + \sigma^2 + \ldots + \sigma^2; \text{ since } E(\sum(u_iu_j) = 0 \text{ for all } i \neq j$$

$$= \sum_{n=1}^{t} \sigma_n^2$$

$$= t\sigma^2$$

As the preceding expression shows, the mean of Y is equal to its initial or starting value, which is constant, but as t increases, its variance increases indefinitely, thus violating a condition of stationarity. In short, the RWM without drift is a non-stationary stochastic process. In practice Y_0 is often set at zero, in which case $E(Y_t) = 0$.

An interesting feature of RWM is the *persistence of random shocks* (i.e., random errors). Y_t is the sum of initial Y_0 plus the sum of random shocks. As a result, the impact of a particular shock does not vanish. For example, if $u_2 = 2$ rather than $u_2 = 0$, then all Y_t 's from Y_2 onward will be 2 units higher and the effect of this shock never dies out. That is why random walk is said to have an *infinite memory.*

If we take the first difference of a random walk without a drift, we get

$$(Y_t - Y_{t-1}) = \Delta Y_t$$

$$= (Y_t + u_t) - Y_t$$

$$= u_t$$

where Δ is the first difference operator. It is easy to show that, while Y_t is non-stationary, its first difference is stationary. In other words, the first differences of a random walk time series are stationary.

Random Walk with Drift: Let us modify the random walk without a drift a little bit as follows:

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

where α is known as the **drift parameter.** The name drift comes from the fact that if we write the preceding equation as

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

It shows that Y_t drifts upward or downward, depending on α being positive or negative. Following the procedure discussed for random walk without drift, it can be shown that for the random walk with drift model:

$$Y_1 = \alpha + Y_0 + u_1$$

$$Y_2 = Y_1 + u_2 = \alpha + \alpha + Y_0 + u_1 + u_2$$

$$Y_3 = Y_2 + u_3 = \alpha + \alpha + \alpha + Y_0 + u_1 + u_2 + u_3$$

$$Y_t = t \cdot \alpha + Y_0 + \sum u_t$$

$$E(Y_t) = \mathrm{E}(t \cdot \alpha + Y_0 + \sum u_t)$$

$$= Y_0 + t \cdot \alpha$$

And the variance will be:

$$Var\,(Y_t) = E(Y_t - E(Y_t))^2$$

$$= E(\alpha + Y_0 + \sum u_t - (t\alpha + Y_0))^2$$

$$= t\sigma^2$$

As you can see, for random walk model (RWM) with a drift, the mean as well as the variance increases over time. Again it violates the conditions of (weak) stationarity. In short, RWM, with or without drift, is a non-stationary stochastic process. Figure 2.4 may be an illustration of RWM with and without a drift. In the graph, RWM with a drift is slightly above the RWM without a drift which shows the drift is positive.

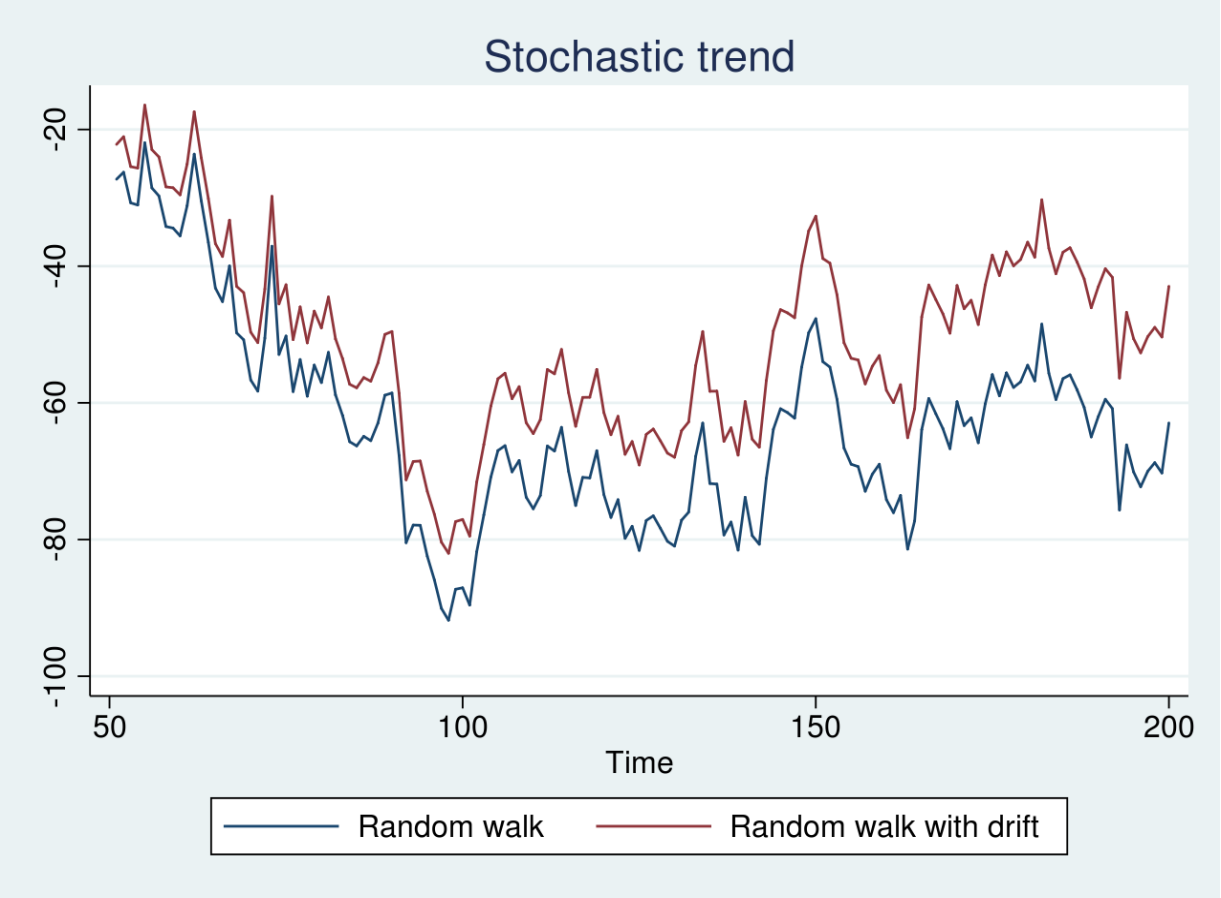


Figure 2.4: Random Walk Models

Deterministic trend: if the stochastic trend is expressed as:

$$y_t = \beta_0 + \beta_1 t + u_t$$

This is called a Trend Stationary Process (TSP).

$$E(y_t) = E(\beta_0 + \beta_1 t + u_t)$$

$$= E(\beta_0) + E(\beta_1 t) + E(u_t)$$

$$= \beta_0 + \beta_1 t$$

Although the mean of y_t is not constant, its variance, indeed, is.

$$Var(y_t) = E(y_t - \mu)^2 =$$

$$= E[(\beta_0 + \beta_1 t + u_t) - (\beta_0 + \beta_1 t)]^2 = E(u_t)^2 = \sigma^2$$

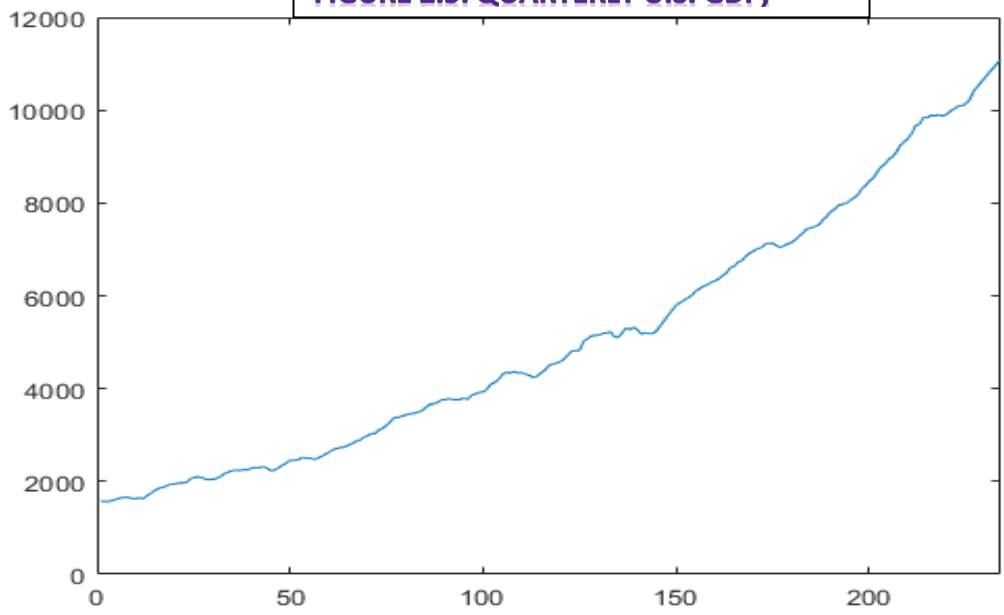
Once the values of β_0 & β_1 are known, the mean can be forecast perfectly since $\mu = \beta_0 + \beta_1 t$. Therefore, if we subtract the mean of y_t from y_t, the resulting series will be stationary. That is why, it is called trend stationary. This procedure of removing the trend is called ***detrending***.

2.3. Trend Stationary and Difference Stationary Stochastic Processes

Non-stationary data, as a rule, are unpredictable and cannot be modeled or forecasted. The results obtained by using non-stationary time series may be spurious in that they may indicate a relationship between two variables where one does not exist. In order to receive consistent, reliable results, the non-stationary data needs to be somehow transformed into stationary data. In contrast to the non-stationary process that has a variable variance and a mean that does not remain near, or returns to a long-run mean over time, the stationary process reverts around a constant long-term mean and has a constant variance independent of time.

The stationary stochastic process is a building block of many econometric time series models. Many observed time series, however, have empirical features that are inconsistent with the assumptions of stationarity. For example, the following plot shows quarterly U.S. GDP measured from 1947 to 2005. There is a very obvious upward trend in this series that one should incorporate into any model for the process.

FIGURE 2.5: QUARTERLY U.S. GDP,



Before we get to the point of transformation for the non-stationary financial time series data, we should distinguish between the different types of the non-stationary processes discussed above. This will provide us with a better understanding of the processes and allow us to apply the correct transformation. A trending mean is a common violation of stationarity. Examples of non-stationary processes are **stochastic trends** being either random walk with or without a drift (a slow steady change) and **deterministic trends** (trends that are constant, positive or negative, independent of time for the whole life of the series).

- *Trend stationary*: The mean trend is deterministic. Once the trend is estimated and removed from the data, the residual series is a stationary stochastic process.
- *Difference stationary*: The mean trend is stochastic. Differencing the series *d* times yields a stationary stochastic process.

The distinction between a deterministic and stochastic trend has important implications for the long-term behavior of a process:

- Time series with a deterministic trend always revert to the trend in the long run (the effects of shocks are eventually eliminated). Forecast intervals have constant width.
- Time series with a stochastic trend never recover from shocks to the system (the effects of shocks are permanent). Forecast intervals grow over time.

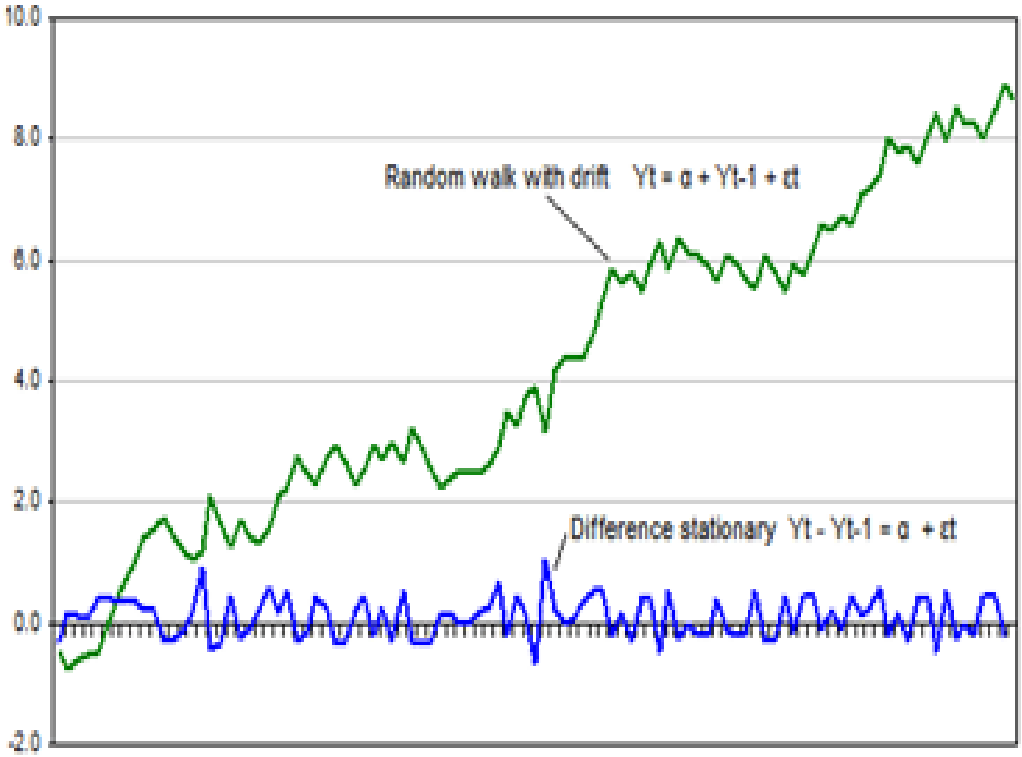
Unit root tests are a tool for assessing the presence of a stochastic trend in an observed series.

The distinction between stationary and non-stationary stochastic processes (or time series) has a crucial bearing on whether the trend is **deterministic** or **stochastic.** Broadly speaking, if the trend in a time series is completely predictable and not variable, we call it a deterministic trend, whereas if it is not predictable, we call it a stochastic trend.

2.3.1. Difference Stationary

A random walk with or without a drift can be transformed to a stationary process by differencing (subtracting Y_{t-1} from Y_t, taking the difference $Y_t - Y_{t-1}$) correspondingly to $Y_t - Y_{t-1} = \varepsilon_t$ or $Y_t - Y_{t-1} = \alpha + \varepsilon_t$ and then the process becomes difference-stationary. The disadvantage of differencing is that the process loses one observation each time the difference is taken.

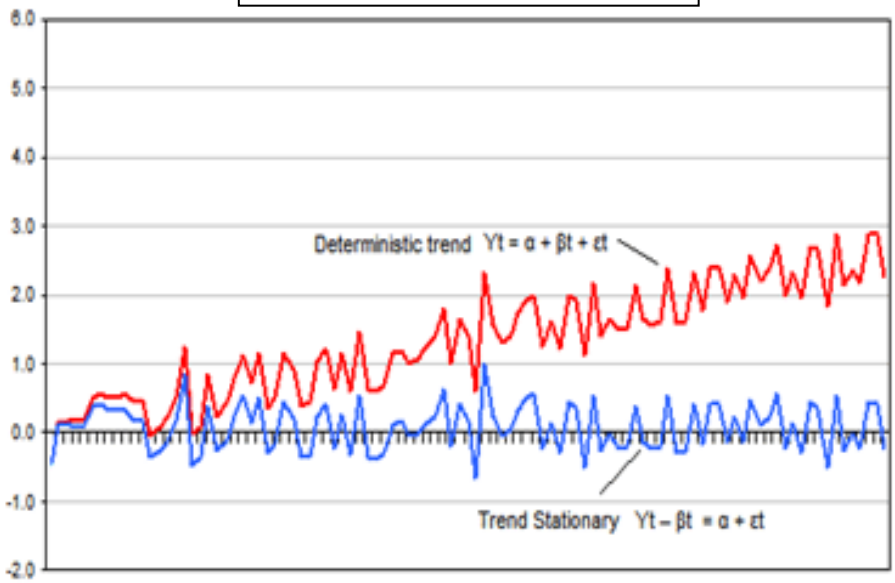
FIGURE 2.6: Differencing



2.3.2. Trend Stationary

A non-stationary process with a deterministic trend becomes stationary after removing the trend, or detrending. For example, $Y_t = \alpha + \beta_t + u_t$ is transformed into a stationary process by subtracting the trend β_t: $Y_t - \beta_t = \alpha + u_t$, as shown in the Figure below. No observation is lost when detrending is used to transform a non-stationary process to a stationary one.

FIGURE 2.7: Detrending



2.4. Integrated Stochastic Process

Time series that can be made stationary by differencing is called integrated stochastic process. Recall that the RWM without a drift is non-stationary but its first difference is stationary. Thus we call RWM without a drift integrated of order 1, denoted as $y_t \sim I(1)$.

Similarly, if a time series has to be differenced twice to make it stationary, such a time series is called integrated of order 2, denoted as $y_t \sim I(2)$. In general, if a non-stationary time series has to be differenced d times to make it stationary, that time series is said to be integrated of order d, $y_t \sim I(d)$.

If a time series y_t is stationary from the start, it is called integrated of order 0, $y_t \sim I(0)$. We often use the terms *'stationary time series'* and *'time series integrated of order zero'* to say the same thing.

2.4.1. Properties of integrated series

Let x_t, y_t, & z_t be three time series:

a. If $x_t \sim I(0)$ and $y_t \sim I(1)$, then $z_t = (x_t + y_t)$ is $I(1)$.

The sum of stationary and non-stationary time series is non-stationary.

b. If $x_t \sim I(d)$, then $y_t = (a + bx_t) \sim I(d)$; where a and b are constants.

The linear combination of I(d) series is also I(d).

c. If $x_t \sim I(d_1)$ and $y_t \sim I(d_2)$, then $z_t = (ax_t + by_t) \sim I(d_1)$, where $d_1 > d_2$.

d. If $x_t \sim I(d)$ and $y_t \sim I(d)$, then $z_t = (ax_t + by_t) \sim I(d')$, *where* $d^* = d$, but sometimes $d^* < d$.

2.5. Tests of Stationarity: The Unit Root Test

Recall that stationary time series is what we most care about mainly because non-stationary time series gives spurious results. So the question is '*how do we know whether a given time series is stationary or not*?' To find out the stationarity of a time series, it is always important and advisable to plot the time series under study graphically as a starting point of more formal tests of stationarity. There are several tests of stationarity. But we will focus on a test which has become popular in recent past, which is *the unit root test*.

2.5.1. The Unit Root Test

The starting point to unit root test is the following autoregressive process.

$$y_t = \rho y_{t-1} + u_t$$

When $\rho=1$, we have a unit root, and thereby a RWM without a drift. The general essence behind the unit root test of stationarity is, therefore, to find out if the estimated rho(ρ) is statistically equal to one. In principle, we can run this regression ($y_t = \rho y_{t-1} + u_t$) and see if $\rho=1$, but we cannot estimate model regressing the series on its lagged value to find out if the estimated $\rho=1$ because in the presence of a unit root, the t-statistic for rho(ρ) coefficient is severely biased.

Therefore, we manipulate this equation ($y_t = \rho y_{t-1} + u_t$) as follows:

$$y_t - y_{t-1} = \rho y_{t-1} - y_{t-1} + u_t$$

$$\Delta y_t = (\rho-1) y_{t-1} + u_t$$

If we let $(\rho-1) = \delta$, then

$$\Delta y_t = \delta y_{t-1} + u_t$$

Now it is a matter of testing if δ is zero or less than zero.

- If $\delta=0$, $\rho-1=0$ ➔ $\rho = 1$; *implying a unit root (non-stationary)*
- $\delta < 0$, $\rho-1<0$ ➔ $\rho < 1$; *implying stationary*
- we exclude a situation $\delta > 0$, $\rho>1$

But the problem is that we cannot rely on the usual t-test on the significance of $\widehat{\delta}$. It is because the null hypothesis is $\delta=0$ *(i.e.* $\rho=1$*)*, and the t-value of the estimated coefficient of y_{t-1} does not follow the t-distribution. The alternative is the Dickey-Fuller (DF) test.

Dicky and Fuller have shown that under the null hypothesis that $\delta=0$, the standard t-value of the coefficient of y_{t-1} follows τ *(tau)* statistic. These authors have computed the critical values of the τ-statistics. In principle, three specifications can be tried, depending on whether or not the series show a trend.

1. $\Delta y_t = \delta y_{t-1} + u_t$ ➔ random walk (no drift, no trend)……..*eq. 1*
2. $\Delta y_t = \beta_0 + \delta y_{t-1} + u_t$ ➔ random walk (with drift, no trend) …….*eq. 2*
3. $\Delta y_t = \beta_0 + \beta_1 t + \delta y_{t-1} + u_t$ ➔ random walk (with drift, with trend)…….*eq. 3*

Note that for each case, H_0: $\delta=0$ (i.e. there is a unit root, and the series is non-stationary or it has a stochastic trend) against H_1: $\delta<0$ (i.e. there is no unit root and the series is stationary, possibly around a deterministic trend).

If the null hypothesis is rejected, it means that y_t is a stationary time series with zero mean in the case of *eq. 1*; that y_t is stationary with a nonzero mean $[= \beta_1/(1-\rho)]$ in the case of *eq. 2*; and that y_t is stationary around a deterministic trend in *eq. 3*.

It is extremely important to note that the critical values of the tau test to test the hypothesis that $\delta = 0$, are different for each of the preceding three specifications of the DF test. Moreover, if, say, specification in *eq. 2* is correct, but we estimate *eq. 1*, we will be committing a specification error. The same is true if we estimate *eq. 3* rather than the true *eq. 2*. Of course, there is no way of knowing which specification is correct to begin with. Some trial and error is inevitable, data mining, nonetheless. The actual estimation procedure is as follows:

Estimate *eq. 1*, or *eq. 2*, or *eq. 2* by OLS; divide the estimated coefficient of y_{t-1} in each case by its standard error to compute the (τ) tau statistic; and refer to the DF tables (or any statistical package).

STATA Commands for DF-test

1. dfuller *y*, noconstant regress
2. dfuller *y*, regress
3. dfuller *y*, trend regress

The results report a Mackinnon p-value.

- If the p-value is less than the significance level, reject the null hypothesis ($\delta=0$).
- If the p-value is greater than the significance level, there is a unit root.

Or alternatively, check the tau-statistic of the lagged y_{t-1} or its coefficient.

- If |computed Ƭ-statistic| > |critical Ƭ-value|, reject the null, ($\delta=0$), hypothesis which implies the time series is stationary
- If the reverse is true, do not reject the null hypothesis which implies the time series is non-stationary.

Critical/ table value of Ƭ-statistic in the three cases

	1%	5%	10%
1. $\Delta y_t = \delta y_{t-1} + u_t$	-2.59	-1.94	-1.62
2. $\Delta y_t = \beta_0 + \delta y_{t-1} + u_t$	-3.51	-2.89	-2.58
3. $\Delta y_t = \beta_0 + \beta_1 t + \delta y_{t-1} + u_t$	-4.07	-3.46	-3.16

Example: the DF-test results of time series variable, y_t, is given as follows.

$\Delta\hat{y}_t = -0.5y_{t-1}$ …… *Eq. 4*

(-6.03)

$\Delta\hat{y}_t = 15 + 0.8y_{t-1}$…… *Eq. 5*

(0.09) (0.97)

$\Delta\hat{y}_t = 27 - 0.6y_{t-1} + 2.317t$…… *Eq. 6*

(15.03) (-4.78) (0.876)

Our primary interest here is in the τ value of the y_{t-1} coefficient. The critical, as given above, 1 percent, 5 percent, and 10 percent τ values for model (*Eq. 4)* are −2.59, −1.94, and −1.62, respectively; and are −3.51, −2.89, and −2.58 for model (*Eq. 5),* respectively; and −4.07, −3.46, and −3.16 for model (*Eq. 6),* respectively.

As noted before, these critical values are different for the three models. Before we examine the results, we have to decide which of the three models may be appropriate. We should rule out model (*Eq. 5)* because the coefficient of y_{t-1}, which is equal to δ is positive. Because $\delta = (\rho - 1)$, a positive δ would imply that $\rho > 1$. Although it is a theoretical possibility, we rule this case out because in this case the time series *y* would be explosive. More technically, the so-called stability condition requires that $|\rho| < 1$.

That leaves us with models (*Eq. 4*) and (*Eq. 6*). In both cases the estimated δ coefficient is negative, implying that the estimated ρ is less than 1. For these two models, the estimated ρ values are 0.5 (1 – 0.5) and 0.4 (1 – 0.6), respectively. The only question now is if these values are statistically significantly below 1 for us to declare that the *y* time series is stationary.

For model *(Eq. 4)* the estimated τ value is −6.03, which in absolute value is above even the 1 percent critical value of −2.59. Since, in absolute terms, the former is larger than the latter, our conclusion is that the *y* time series is stationary. The story is the same for model (*Eq. 6*). The computed τ value of *−4.78* is greater than even the 1 percent critical τ value of *−4.07* in

absolute terms. Therefore, on the basis of the Dickey–Fuller test, the conclusion is that the given *y* time series does not contain a unit root.

2.6. Co-integration and Error Correction Mechanism

The discussion of spurious regression in the previous section certainly makes one wary of using the levels of I(1) variables in regression analysis. In earlier chapters, we suggested that I(1) variables should be differenced before they are used in linear regression models, whether they are estimated by OLS or instrumental variables. This is certainly a safe course to follow. Unfortunately, always differencing I(1) variables limits the scope of the questions that we can answer.

There is a unique case where a regression of a non-stationary series on another non-stationary series does not result in spurious regression. Recall one of the properties of integrated stochastic processes in section 2.4; that if $x_t \sim I(d)$ and$y_t \sim I(d)$, then $z_t = (ax_t + by_t) \sim I(d')$, *where* $d' = d$, but sometimes $d' < d$. In other words, if two non-stationary stochastic processes are integrated of order one, the linear combination of the two stochastic processes can produce stationary stochastic process. This is the situation of cointegration.

If two time series have stochastic trends (i.e. they are non-stationary), a regression of one on the other may cancel out the stochastic trends, which may suggest that there is a long-run, or equilibrium, relationship between them even though individually the two series are non-stationary.

Let us suppose that we consider an income (Y_t) and consumption (C_t) two stochastic processes. Subjecting these time series individually to unit root analysis, suppose that they both are $I(1)$; that is, they contain a unit root. Suppose, then, that we regress C_t on Y_t as follows:

$$C_t = \alpha_1 + \alpha_2 Y_t + u_t \ldots\ldots\ldots \quad (1)$$

Let us rewrite this model as*:*

$$u_t = C_t - \alpha_1 - \alpha_2 Y_t \ldots\ldots\ldots \quad (2)$$

Suppose we now subject u_t to unit root analysis and find that it is stationary; that is, it is $I(0)$. This is an interesting situation, for although C_t and Y_t are individually $I(1)$, that is, they have

stochastic trends, their linear combination equation *(2)* is $I(0)$. So to speak, the linear combination cancels out the stochastic trends in the two series. If you take consumption and income as two $I(1)$ variables, savings defined as (income − consumption) could be $I(0)$.

As a result, a regression of consumption on income as in equation *(1)* would be meaningful (i.e., not spurious). In this case, we say that the two variables are **cointegrated.** Economically speaking, two variables will be cointegrated if they have a long-term, or equilibrium, relationship between them.

In short, provided we check that the residuals from regressions like equation *(1)* are $I(0)$ or stationary, the traditional regression methodology (including the *t* and *F* tests) that we have considered extensively is applicable to data involving (non-stationary) time series. The valuable contribution of the concepts of unit root, cointegration, etc. is to force us to find out if the regression residuals are stationary. As Granger notes, "A test for cointegration can be thought of as a pre-test to avoid 'spurious regression' situations."

In the language of cointegration theory, a regression such as equation *(1)* is known as a **cointegrating regression** and the slope parameter α_2 is known as the **cointegrating parameter.** The concept of cointegration can be extended to a regression model containing *k* regressors. In this case we will have *k* cointegrating parameters.

2.6.1. Test cointegration:

A. Engle Granger (EG) test

We already know how to apply the DF or ADF unit root tests. All we have to do is:

→ estimate a regression like equation (1)***;*** $C_t = \hat{\alpha_1} + \hat{\alpha_2} Y_t$

→ obtain the estimated residuals $\hat{u}_t$

→ use the DF test to see whether or not $\hat{u}_t$ has a unit root

There is one precaution to exercise, however. Since the estimated $\hat{u}_t$ are based on the *estimated* cointegrating parameter α_2, the DF critical significance values are not quite appropriate. Engle and Granger have calculated these values, which can be found in the references. Therefore, the Dickey Fuller (DF) and Augumented Dickey Fuller (ADF) tests in the present context are known as **Engle–Granger (EG)** and **augmented Engle–Granger**

(AEG) tests. However, several software packages now present these critical values along with other outputs.

Let us illustrate these tests. We first regressed PCE on PDI and obtained the following regression:

$$C_t = 278.5 + 0.92Y_t$$

$t = (3.58) \quad (12.8)$

Since C_t and Y_t are individually non-stationary, there is the possibility that this regression is spurious. But suppose that when we perform a unit root test on the residuals obtained from the above regression result, we obtain the following results:

$$\widehat{\Delta u}_t = -0.312\hat{u}_{t-1}$$

$t = (-4.321)$

The Engle–Granger 1 percent critical τ value is −2.5899. Since the computed $\tau(= t)$ value is much more negative than this, our conclusion is that the residuals from the regression of C_t on Y_t are $I(0)$; that is, they are stationary. Hence, the above regression result of C_t on Y_t is a cointegrating regression and this regression is not spurious, even though individually the two variables are non-stationary.

One can call this cointegrating regression the **static** or **long run** consumption function and interpret its parameters as long run parameters. Thus, 0.92 represents the long-run, or equilibrium, Marginal Propensity to Consumer (MPC).

2.6.2. Error Correction Mechanism

We just showed that C_t and Y_t are cointegrated; that is, there is a long term, or equilibrium, relationship between the two. Of course, in the short run, there may be disequilibrium. Therefore, one can treat the error term in equation (2) as the "equilibrium error." And we can use this error term to tie the short-run behavior of C_t to its long-run value. The **error correction mechanism (ECM)** first used by Sargan (J. D. Sargan, "Wages and Prices in the United Kingdom: A Study in Econometric Methodology," In K. F. Wallis and D. F. Hendry, eds., *Quantitative Economics and Econometric Analysis*, Basil Blackwell, Oxford, U.K., 1984.) and later popularized by Engle and Granger corrects for disequilibrium. An important

theorem, known as the **Granger representation theorem,** states that *if two variables Y and X are cointegrated, then the relationship between the two can be expressed as ECM.* To see what this means, let us revert to our consumption–income example. Now consider the following model:

$$\Delta C_t = \alpha_1 + \alpha_2 \Delta Y_t + \alpha_3 u_{t-1} + e_t \ldots\ldots\ldots\ldots \quad (3)$$

Where Δas usual denotes the first difference operator, e_t is a random error term, and $u_{t-1} = (C_{t-1} - \alpha_1 - \alpha_2 Y_t)$, that is, the one-period lagged value of the error from the cointegrating regression of equation (1).

ECM equation (3) states that ΔC depends on ΔY and also on the equilibrium error term. If the equilibrium error term (u_{t-1}) is nonzero, then the model is out of equilibrium. Suppose ΔY is zero and $u_{t-1}$is positive. This means C_{t-1} is too high to be in equilibrium, that is, C_{t-1} is above its equilibrium value of ($\alpha_1 + \alpha_2 Y_{t-1}$). Since α_3 is expected to be negative, the term $\alpha_3 u_{t-1}$ is negative and, therefore, ΔC_t will be negative to restore the equilibrium. That is, if C_t is above its equilibrium value, it will start falling in the next period to correct the equilibrium error; hence the name error correction mechanism (ECM).

By the same token, if u_{t-1} is negative (i.e., C is below its equilibrium value), $\alpha_3 u_{t-1}$ will be positive, which will cause ΔC_t to be positive, leading C_t to rise in period t. Thus, the absolute value of α_3 decides how quickly the equilibrium is restored. In practice, we estimate u_{t-1} by $\hat{u}_{t-1} = (C_t - \hat{\alpha_1} - \hat{\alpha_2} Y_t)$.

Returning to our illustrative example, the empirical counterpart of equation (3) is:

$$\widehat{\Delta C}_t = 57.45 + 0.19 \Delta Y_t - 0.12 \hat{u}_{t-1}$$
$$t = (4.69) \qquad (6.12) \qquad (-3.59)$$

Statistically, the equilibrium error term is zero, suggesting that C adjusts to changes in Y in the same time period. As the above result shows, short-run changes in Y have a positive impact on short-run changes in C. One can interpret 0.19 as the short-run marginal propensity to consume (MPC); the long-run MPC is given by the estimated (static) equilibrium relation in the original OLS regression result as 0.92.

Check list of terms

Cointegration

Deterministic trend

Detrending

Difference stationary stochastic process

Drift

Engle Granger Test

Error correction mechanism

Granger representation theorem

Integrated stochastic process

Non stationary data

Random walk with drift

Random walk without drift

Stationary data

Stochastic process

Temporal ordering

Trend stationary stochastic process

Unit root test

Chapter Three

Introduction to Simultaneous Equation models

In the previous sections, you have learned regressions involving single equation models. In such models, the dependent variable is expressed as a function of one or more independent variables. In this chapter, you are going learn about models involving two or more equations. These models are known as simultaneous equation models.

The nature of simultaneous equation model is presented in section 3.1. In this section, you will look at the endogenous variable and exogenous variable. Section 3.2 then presents the simultaneity bias and the inconsistency of OLS estimation of simultaneous equations. In section 3.3, you will see identification and estimation of structural equations in simultaneous equation models. In this section you will know about two methods of identification: the order condition for identification and the rank condition for identification.

Finally, in section 3.4, you will be familiar with methods of estimation of simultaneous equations. These include, the Indirect Least Squares (ILS), Instrumental Variable (IV) and Two-Stage Least Squares (2SLS) estimation of structural equations.

Objective of the chapter

At the end of the chapter you are expected to:

- Know the meaning and nature of simultaneous equations,
- Identify endogenous and exogenous variables of a simultaneous equation
- Understand what does it mean by simultaneity bias
- Know the problems of estimating simultaneous equation using ordinary least squares method
- Understand what does it mean by structural equation of simultaneous equation
- Understand what does it mean by reduced form equation of simultaneous equation
- Determine identification of a simultaneous equation using order condition
- Determine identification of a simultaneous equation using rank condition
- Know how to estimate simultaneous equation using indirect least squares method

- Know how to estimate simultaneous equation using instrumental variables method
- Know how to estimate simultaneous equation using two-stage least squares method
- Know how to drive structural equation parameters from estimates of reduced form equation.

What is simultaneous equation model?

Dear learner, what do you know about simultaneous equations from your linear algebra lessons and from your high school math classes? Write what comes to your mind on the following spaces ----------
---

---.

Hopefully, you answered that a simultaneous equation is an equation system which has more than one equation and more than one variables or unknowns. Usually, the number of equations is equals the number of variables. Otherwise, it would be difficult to find a unique solution or may not be able to find a solution at all. Simultaneous equation models quite resemble simultaneous equations you know before. Try to read and understand the following.

3.1. The Nature of Simultaneous Equation Models

All discussions we have made so far involve single dependent variable - called most of the time *Y* - and one or more explanatory variables, usually called X's. In such models the emphasis was on estimating and/or predicting the average value of the dependent variable(*Y)*conditional upon the fixed values of the explanatory variables (*X*'s). In such models it is, thus, assumed that X's cause/determine Y, or Y is caused/determined by X's. For example, if you are given two variables consumption (Y) and income (X), and asked about the cause-effect relationship, you may reasonably respond that income determines consumption or X causes Y.

In any regression modeling, generally an equation is considered to represent a relationship describing a phenomenon. Many situations involve a **set of relationships** which explain the

behavior of certain variables. For example, in analyzing the market conditions for a particular commodity, there can be a demand equation and supply equations which explain the price and quantity of commodity exchanged in the market at market equilibrium. So there are two equations to explain the whole phenomenon - one for demand and another for supply. In such cases, it is not necessary that all the variables should appear in all the equations. So estimation of parameters under this type of situation has those features that are not present when a model involves only a single relationship. In particular, when a relationship is a part of a system, then some explanatory variables are stochastic and are correlated with the disturbances. 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.

Question: Can it always be the case that one variable causes the other? Obviously No!

We may believe for whatever reason that the dependent variable (Y) is not only a function of explanatory variables (X's) but also all or some of the X's, in turn, are function of the dependent variable (Y) itself. To identify such situation, we may reasonably use counterfactual reasoning that there is two way relationship between two or more variables. The existence of this ***two-way flow of influence*** between Y and the X's makes the distinction between dependent and independent variables a little dubious or doubtful. Therefore, to understand the multi-flow of influence among the variables, we need to consider more than one regression equations for each variable and this is what ***simultaneous equation*** **models** (SEM) deal about.

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.

3.1.1. 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.

3.1.2. 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.

Exogenous variables help is explaining the variations in endogenous variables. It is customary to include past values of endogenous variables in the predetermined group. Since exogenous variables are predetermined, so they are independent of disturbance term in the model. They satisfy those assumptions which explanatory variables satisfy in the usual regression model. Exogenous variables influence then endogenous variables but are not themselves influenced by them. One variable which is endogenous for one model can be exogenous variable for the other model.

Note that: *In linear regression model, the explanatory variables influence study variable but not vice versa. So relationship is one sided*.

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.

In general simultaneous equation models have the following form:

$$\left.\begin{aligned} Y_i &= \alpha_0 + \alpha_1 X_i + \alpha_2 Z_{1i} + u_i \\ X_i &= \beta_0 + \beta_1 Y_i + \beta_2 Z_{2i} + \varepsilon_i \end{aligned}\right\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.1)$$

Where, Z_{1i} is a variable which affects Y_i but not X_i and the variable Z_{2i} affects only X_i but not Y_i. u_i and ε_i are the stochastic disturbance terms. We will discuss these in more detail later.

Look at the first line of equation (3.1) that Y_i is a function of X_i and Z_{1i}. On the other hand, in the second equation, X_i -which was serving as an independent variable in the first equation- becomes a function of Y_i-which was treated as a dependent variable in the first equation- and Z_{2i}. So, we can't identify whether the variables Y_i and X_i are dependent or independent in this SEM. Can we?

A model constitutes a system of simultaneous equations if all the relationships involved are needed for determining the value of at least one of the endogenous variables included in the model. This implies that at least one of the relationships includes more them one endogenous variable.

Since Z_{1i} and Z_{2i} are observable, they are called **observed shifters** in the SEM. On the other hand u_i and ε_i are not observed but affect the two equations respectively and are, thus, called **unobserved shifters** ipso facto.

Several important features can be observed from equation (3.1).

- ✓ One, if the equations are derived from economic theory and have causal interpretation, they are called **structural or behavioral equations.**
- ✓ Two, given the values of Z_{1i}, Z_{2i}, u_i, ε_i, these two equations determine Y_i and X_i. For this reason,Y_i and X_i are known **as endogenous variables** in the SEM.
- ✓ Three, because Z_{1i} and Z_{2i} are determined *outside* the model we view them as **exogenous variables.** From a statistical standpoint, the key assumption concerning Z_{1i} and Z_{2i} is that they are both *uncorrelated with error terms*.
- ✓ Four, without including Z_{1i} and Z_{2i} in the model, there is no way to ***identify***[2] each equation.

Example 3.1:

If you want to specify an econometric model containing the variables asset and income, then which one will you make the dependent and which one will you make the independent variable?

To answer this, ask yourself: Does asset determine income? Or, does income determine asset?

Here, if you think intelligently, you may reason out that a person who receives higher income will have higher asset. Also, large asset can be source of higher income. If so,you will answer that we cannot distinguish which one is the dependent variable and which one is the independent variable.

[2] Refer to section 3.3.1 to make yourself clear with the concept of ***identification***.

The model for the above example may, therefore, have the following type of appearance:

$$\left.\begin{aligned} Asset &= \alpha_0 + \alpha_1 Income + \alpha_2 Z_{1i} + u_i \\ Income &= \beta_0 + \beta_1 Asset + \beta_2 Z_{2i} + \varepsilon_i \end{aligned}\right\} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (3.2)$$

Where Z_{1i} is a variable which affects asset but not income so that income model can be ***identified*** and the variable Z_{2i} affects only income but not asset so that asset model can be ***identified***.

Example 3.2:

From economic theory we know that quantity demanded is determined by price and other factors that affect demand. Also, quantity supplied is determined by price and other factors which affect supply. So, will finding equilibrium point require formulating SEM? Why?

Answer: (See equation (3.9))

3.2. Simultaneity bias (Inconsistency of OLS Estimators under SEM)

In a simultaneous equation model, if an explanatory variable is determined simultaneously with the dependent variable, it is generally correlated with the error term and applying OLS will result in biased and inconsistent estimates. That is, the least squares estimator of parameters in a structural simultaneous equation is biased and inconsistent because of the correlation between the random error and the endogenous variables on the right-hand side of the equation.

Take once again equation (3.1) specified above. For simplicity, suppress/ignore the constant term.

$$\left.\begin{aligned} Y_i &= \alpha_0 + \alpha_1 X_i + \alpha_2 Z_{1i} + u_i \\ X_i &= \beta_0 + \beta_1 Y_i + \beta_2 Z_{2i} + \varepsilon_i \end{aligned}\right\} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (3.3)$$

Equation (3.3) is a simultaneous equation model in which Y_i and X_i are endogenous variables, and Z_{1i} and Z_{2i} are exogenous variables. By assumption (which emanates from reasoning) Z_{1i} and u_i are uncorrelated and also Z_{2i} and ε_i are uncorrelated. If we estimate, for example, the first equation $Y_i = \alpha_1 X_i + \alpha_2 Z_{1i} + u_i$ alone using OLS, the estimates for α_1 and α_2 will be biased and inconsistent. This is because X_i and u_i are correlated which creates ***endogeniety***

problem (the most serious assumption of linear regression model). This can be shown mathematically as follows.

Let's solve for X_i in terms of exogenous variables ($Z_{1i}$and Z_{2i}) and error term. To doso, plug in the value of Y_i from the first equation in to the second equation which becomes:

$$X_i = \beta_1(\alpha_0 + \alpha_1 X_i + \alpha_2 Z_{1i} + u_i) + \beta_2 Z_{2i} + \varepsilon_i$$

$$X_i - \beta_1\alpha_1 X_i = \beta_1\alpha_0 + \beta_1\alpha_2 Z_{1i} + \beta_1 u_i + \beta_2 Z_{2i} + \varepsilon_i$$

$X_i(1 - \beta_1\alpha_1) = \beta_1\alpha_0 + \beta_1\alpha_2 Z_{1i} + \beta_1 u_i + \beta_2 Z_{2i} + \varepsilon_i$, where $\beta_1\alpha_1 \neq 1$ ……..(3.4)

Dividing both sides of the equation by $(1 - \beta_1\alpha_1)$, gives:

$$X_i = \frac{\beta_1\alpha_0}{1 - \beta_1\alpha_1} + \frac{\beta_1\alpha_2}{1 - \beta_1\alpha_1} Z_{1i} + \frac{\beta_2}{1 - \beta_1\alpha_1} Z_{2i} + \frac{\beta_1 u_i + \varepsilon_i}{1 - \beta_1\alpha_1} \text{ ……..(3.5)}$$

If we let,

$\pi_0 = \frac{\beta_1\alpha_0}{1-\beta_1\alpha_1}$, $\pi_1 = \frac{\beta_1\alpha_2}{1-\beta_1\alpha_1}$, $\pi_2 = \frac{\beta_2}{1-\beta_1\alpha_1}$, $w_i = \frac{\beta_1 u_i + \varepsilon_i}{1-\beta_1\alpha_1}$, then we will have:

$$X_i = \pi_0 + \pi_1 Z_{1i} + \pi_2 Z_{2i} + w_i \text{ ……..(3.6)}$$

Remember that Z_{1i} and Z_{2i} are assumed to be exogenous. Therefore, from equation (3.6) X_i and u_i are correlated if w_i and u_i are correlated. But, we can see that w_i is linear function of $u_i$and ε_i, so it is ***correlated*** with u_i.Note that w_i and u_i will be correlated if

- ✓ $\beta_1 \neq 0$ even if u_i and ε_i are uncorrelated, or
- ✓ u_i and ε_i are correlated even if $\beta_1 = 0$

Both cases mean$X_i$is simultaneously determined with Y_i.

In general, when a variable X_i is correlated with u_i because of simultaneity, we say that OLS suffers from simultaneity bias. Obtaining the direction of the bias in the coefficients will generally get more complicated as the explanatory variables in the model increase.

Equation (3.6), expresses X_i in terms of the exogenous variables and the error terms. This is called the **reduced form equation** for X_i. A reduced-form equation is one that expresses an endogenous variable solely in terms of the exogenous variables and the stochastic

disturbances. The parameters, π_{21} and π_{22} are called **reduced form parameters**. These parameters are nonlinear functions of the **structural parameters**, which appear in the structural equation (3.3).

Since u_i and ε_i are each uncorrelated with Z_{1i} and Z_{2i}, the **reduced form error term**$(w_i)$is also uncorrelated with Z_{1i} and Z_{2i}because it is a linear function of u_i and ε_i. Therefore, we can consistently estimate π_{21} and $\pi_{22}$by OLS, something that is used for two-stage least squares (2SLS) estimation[3].

3.3. Identification and Estimation of Structural Equations in SEM

Earlier we have said that SEM, unlike linear regression model, cannot be estimated directly using OLS technique for it will give us biased and inconsistent estimates. Rather there are other estimation techniques-like Instrumental Variable (IV) estimation or two-stage least squares estimation method ((2SLS). However, a SEM has to first pass the criteria called identification which we promised to discuss right now under section 3.1. Hence, we will first discuss the issue of identification condition under section (3.3.1) and will then proceed to estimation under section (3.3.2).

3.3.1. Identification of Structural Equation (Order and rank conditions) (without proof)

By ***identification*** it is to mean whether numerical estimates of the parameters of a structural equation can be obtained from the estimated reduced-form coefficients. Thus, when we say that an equation is identified, it means we can estimate the parameters of a structural equation from the estimated reduced-form coefficients. Identification is a concern of model formulation, not estimation as the latter depends up on the empirical data and the form of the model. Let's elucidate this concept with the following example.

Consider the following SEM involving ***two-equation*** system.

$$\left.\begin{aligned} Y_1 &= \alpha_0 + \alpha_1 Y_2 + \alpha_2 Z_1 + u \\ Y_2 &= \beta_0 + \beta_1 Y_1 + \beta_2 Z_2 + \varepsilon \end{aligned}\right\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.7)$$

[3] Refer section 3.3.2 to know about 2SLS.

Where,Y_1 andY_2 are endogenous variables (because they are correlated with the error terms), and Z_1 and$Z_2$are exogenous (because they're assumed to be independent of, or uncorrelated with, the error terms).

For the first equation to be *identified*, there should be *at least* one variable which is ***excluded*** from the first equation- but included in the other equation - and is uncorrelated with the error term of the first equation. This is known as **exclusion restrictions.** Therefore**,** sinceZ_2 is excluded from the first equation, and since Z_2 is uncorrelated with the error term u-because it is assumed to be exogenous-the first equation is identified.

Similarly, for the second equation to be identified there should be *at least* one variable which is ***excluded*** from the second equation, and is uncorrelated with the error term (ε)of this equation. SinceZ_1 is excluded from the second equation, and sinceZ_2 is assumed to be exogenous (uncorrelated with the error term, ε) the second equation is identified.

Take another SEM which involves ***three-equation*** system.

$$\left.\begin{aligned} Y_1 &= \alpha_0 + \alpha_1 Y_2 + \alpha_3 Y_3 + \alpha_4 \boldsymbol{Z_1} + u \\ Y_2 &= \beta_0 + \beta_1 Y_1 + \beta_2 \boldsymbol{Z_1} + \beta_3 \boldsymbol{Z_2} + \beta_4 \boldsymbol{Z_3} + \varepsilon \\ Y_3 &= \gamma_0 + \gamma_1 Y_2 + \gamma_2 \boldsymbol{Z_1} + \gamma_3 \boldsymbol{Z_2} + \gamma_4 \boldsymbol{Z_3} + \gamma_5 \boldsymbol{Z_4} + v \end{aligned}\right\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.8)$$

Where, Y_1, Y_2, and Y_3 are endogenous variables, and $\boldsymbol{Z_1}$, $\boldsymbol{Z_2}$, $\boldsymbol{Z_3}$and$\boldsymbol{Z_4}$ are exogenous variables.α_i, β_iandγ_i are estimators of the three structural equations respectively.

It is generally difficult to show that an equation is identified in an SEM with more than two equations, but it is easy to see when certain equations are *not* identified.

The first equation is identified (is at least promising) because three exogenous variables Z_2, Z_3andZ_4 are excluded from this equation.

The second equation is identified (is at least promising). Why?

But, the third equation is not identified because no exogenous variables excluded from this equation. That, is Z_1, appears in the first equation,Z_2 and Z_3 appear in the second equation and Z_4appear in the third equation itself.

Formally speaking, there are two rules/conditions which must be fulfilled for an equation to be identified.

A. The order condition for identification

This condition is based on a counting rule of the exogenous and endogenous variables included and excluded in the SEM. This condition states that an equation in any SEM satisfies the order condition for identification if the number of *excluded* exogenous variables from the equation is at least as large as the number of right-hand side endogenous variables in the equation.

Now, identification condition for equation (3.8) can be answered based on order condition as:

Table 3.1: order condition for identification

Equation No	**No of endogenous variables**	**No of excluded exogenous variables**	**Identification condition**
1st	2 (=Y_2 and Y_3)	3 (= $\boldsymbol{Z_2, Z_3 and Z_4}$)	(3 ≥ 2) over-identified
2nd	**1**	**1**	(1 = 1)just identified
3rd	**1**	**0**	(0 ≤ 1)underidentified

Note from the above that in the first equation the number of excluded exogenous variables (=3) is greater than the number of endogenous variables (=2). Such an equation is called ***over-identified***. In the second equation, the number of excluded exogenous variables (=1) is exactly equal to the number of endogenous variables (=1). In this case, the equation is called ***just-identified***. In the third equation, the number of excluded exogenous variables (=0) is less than the number of endogenous variables (=1) and the equation is called **unidentified equation**. **Under-identified** equation means we can't estimate the parameters of a structural equation from the estimated reduced-form coefficients.

Order condition is a necessary (but not sufficient) condition for identification. For example, from equation (3.8), we have said that the second equation is identified because of the presence of an excluded variable Z_4 from this equation. But, if $\gamma_5 = 0$, it means Z_4 is not correlated with Y_1, Y_2 *or* Y_3 and will be eliminated from the model, so the second equation will remain unidentified. This again illustrates that identification of an equation depends not only on the presence of an excluded variable but also on the values of the parameters (which we can never know for sure) in the other equations. The sufficient condition for identification is called the **rank condition** which is discussed below.

B. The rank condition for identification

The rank condition states that in an SEM containing G equations any particular equation is identified if and only if it is possible to construct at least one non-zero determinant of order (G-1) from the coefficients of the variables excluded from that particular equation but contained in the other equations of the model.

Remember from your linear algebra course that, the term **rank** refers to the rank of a matrix and is given by the largest-order square matrix(contained in the given matrix) whose determinant is nonzero. Alternatively, the rank of a matrix is the largest number of linearly independent rows or columns of a matrix.

To understand the order and rank conditions, let's introduce the following notations:

Let, M = number of endogenous variables in the model

m = number of endogenous variables in a given equation

K = number of exogenous variables in the model including the intercept

k = number of exogenous variables in a given equation

- ***Order Condition***
  - ✓ In a model of M simultaneous equations in order for an equation to be identified, it must exclude *at least* $M-1$ variables (endogenous as well as exogenous) appearing in the model. If it excludes exactly $M-1$ variables, the equation is just identified. If it excludes more than M−1 variables, it is over-identified.

- ✓ In a model of M simultaneous equations, in order for an equation to be identified, the number of exogenous variables excluded from the equation must not be less than the number of endogenous variables included in that equation less 1, that is,

$$K - k \geq m - 1$$

If $K - k = m - 1$, the equation is just identified, but if $K - k > m - 1$, it is over-identified.

- ➢ ***Rank condition***
  - ✓ In a model containing M equations in M endogenous variables, an equation is identified if and only if *at least* one ***nonzero*** determinant of order $(M - 1)(M - 1)$ can be constructed from the coefficients of the variables (both endogenous and predetermined) excluded from that particular equation but included in the other equations of the model.
  - ✓ In a model containing in simultaneous equations:
    - If $K - k > m - 1$ and the rank of the **A** matrix is $M - 1$, the equation is over-identified.
    - If $K - k = m - 1$ and the rank of the matrix **A** is $M - 1$, the equation is exactly identified.
    - If $K - k \geq m - 1$ and the rank of the matrix **A** is less than $M - 1$, the equation is under-identified.
    - If $K - k < m - 1$, the structural equation is unidentified. The rank of the **A** matrix in this case is bound to be less than $M - 1$. (Why?)

Steps of checking rank condition

1. Bring all items of each equation, except the error term, to the left of the equal sign
2. Put all the endogenous and exogenous variables in a row
3. Put the corresponding coefficients of each variable beneath each variable
4. Construct a matrix from excluded variables (both exogenous and endogenous) and check for its rank
5. If we can form at more than one (M-1) by (M-1) matrix of non-zero determinant, the matric is said to be over identified. If we can form at exactly one (M-1) by (M-1) matrix of non-zero determinant, the matric is said to be just identified. If we can not

form at least (M-1) by (M-1) matrix of non-zero determinant, the matric is said to be under identified.

Illustration

Given the following simultaneous equation model, check whether it is identified or not using both rank and order conditions?

$$Y_1 = \alpha_0 + \alpha_1 Y_2 + \alpha_2 Y_3 + \alpha_3 \boldsymbol{Z_1} + u$$
$$Y_2 = \beta_0 + \beta_1 Y_3 + \beta_2 \boldsymbol{Z_1} + \beta_3 \boldsymbol{Z_2} + \varepsilon$$
$$Y_3 = \gamma_0 + \gamma_1 Y_1 + \gamma_2 \boldsymbol{Z_1} + \gamma_3 \boldsymbol{Z_2} + v$$
$$Y_4 = \theta_0 + \theta_1 Y_1 + \theta_2 Y_2 + \theta_3 \boldsymbol{Z_3} + \omega$$

To check the order condition for identification, look at the following table

Table 3.2: order condition for identification

Equation No	**No of endogenous variables**	**No of excluded exogenous variables**	**Identification condition**
1st	2 (=Y_2 *and* Y_3)	2 (= $\boldsymbol{Z_3 and Z_4}$)	(3 ≥ 2) exactly-identified
2nd	**1** (=Y_3)	**1**(= $\boldsymbol{Z_3}$)	(1 = 1)just identified
3rd	**1** (=Y_1)	1 (=Z_3)	(1 = 1)exactly-identified
4th	2 (=Y_1 *and* Y_2)	2 (= $\boldsymbol{Z_1 and Z_2}$)	(3 ≥ 2) just-identified

As it can be seen from the table, all equations are just identified.

Now, Let us recheck with the rank condition.

First, let's bring all items, except error terms, to the left

$$Y_1 - \alpha_0 - \alpha_1 Y_2 - \alpha_2 Y_3 - \alpha_3 \boldsymbol{Z_1} = u$$
$$Y_2 - \beta_0 - \beta_1 Y_3 - \beta_2 \boldsymbol{Z_1} - \beta_3 \boldsymbol{Z_2} = \varepsilon$$
$$Y_3 - \gamma_0 - \gamma_1 Y_1 - \gamma_2 \boldsymbol{Z_1} - \gamma_3 \boldsymbol{Z_2} = v$$
$$Y_4 - \theta_0 - \theta_1 Y_1 - \theta_2 Y_2 - \theta_3 \boldsymbol{Z_3} = \omega$$

The second step is, put all the endogenous and exogenous variables in a row and put the corresponding coefficients beneath each variable

Table 3.3: Rank condition for identification

Eq. No	**1**	Y_1	Y_2	Y_3	Y_4	Z_1	Z_2	Z_3
1st	$-\alpha_0$	1	$-\alpha_1$	$-\alpha_2$	**0**	$-\alpha_3$	**0**	**0**
2nd	$-\beta_0$	**0**	1	$-\beta_1$	0	$-\beta_2$	$-\beta_3$	0
3rd	$-\gamma_0$	$-\gamma_1$	0	1	0	$-\gamma_2$	$-\gamma_3$	0
4th	$-\theta_0$	$-\theta_1$	$-\theta_2$	**0**	**1**	**0**	**0**	$-\theta_3$

Consider the first equation, which excludes variables **Y**4, $\mathbf{Z}_2$, and Z_3 (this is represented by zeros in the first row of table 3.3). For this equation to be identified, we must obtain at least one nonzero determinant of order 3 × 3 from the coefficients of the variables excluded from this equation but included in other equations. To obtain the determinant we first obtain the relevant matrix of coefficients of variables **Y**4, $\mathbf{Z}_2$, and $\mathbf{Z}_3$ included in the other equations. In the present case there is only one such matrix, call it **A**, defined as follows.

$$A = \begin{bmatrix} 0 & -\beta_3 & 0 \\ 0 & -\gamma_3 & 0 \\ 1 & 0 & -\theta_3 \end{bmatrix}$$

If we find the determinant of matrix A, it is equal to zero. This implies the rank of the matrix is less than 3 and it is not identified. Therefore, although the order condition shows that the SEM is identified, the rank condition shows that it is not.

As noted, the rank condition is both a necessary and sufficient condition for identification.

3.3.2. Indirect Least Squares (ILS), Instrumental Variable (IV) and Two-Stage Least Squares (2SLS) estimation of structural equations

Once we finished modeling SEM, the next task is estimation. Yet, estimation problem is rather complex because there are a variety of estimation techniques with varying statistical properties.

In an SEM, two approaches may be adopted to estimate the structural 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*.

Although the systems method-such as the **full information maximum likelihood (FIML) method-** may be good to preserve the spirit of simultaneous-equation models, in reality they are not commonly used for different reasons. Some of these include:

i. High burden of the computation: for example, estimating 20 equations require incorporating 151 coefficients for US economy in 1955.
ii. The systems methods, such as FIML, lead to solutions that are highly non-linear in the parameters and are, therefore, often difficult to determine.
iii. If there is a specification error (say, a wrong functional form or exclusion of relevant variables) in one or more equations of the system, that error is transmitted to the rest of the system. As a result, the systems methods become very sensitive to specification errors.

Due to the above problems, therefore, single-equation methods are often used in practice. These include:

A. Ordinary least squares (OLS)
B. Indirect least squares (ILS)
C. Two-stage least squares (2SLS)

A. Ordinary least squares (OLS):

OLS can be used for **recursive, triangular,** or **causal** models. Since this is out of the course outline, we will not go for it; you can explore details from books you can have access to.

B. Indirect least squares (ILS)

For a just or an *exactly identified* structural equation, the method of obtaining the estimates of the structural coefficients from the OLS estimates of the reduced-form coefficients is known as the **method Indirect Least Squares(ILS),** and the estimates obtained are known as the **indirect least squares estimates.**

Steps in ILS

Step 1: Obtain the reduced-form equations; solve for the endogenous variable in each equation in terms solely of the exogenous variables and the stochastic error term. This gives the reduced-form equations.

Step 2: Apply OLS to the reduced-form equations individually. Since the explanatory variables in these equations are predetermined/exogenous variables which are uncorrelated with the stochastic disturbances, this operation is permissible and the estimates obtained are consistent.

Step 3: Obtain estimates of the original structural coefficients from the estimated reduced-form coefficients obtained in Step 2. If an equation is exactly identified, there is a one-to-one correspondence between the structural and reduced-form coefficients; that is, one can derive unique estimates of the former from the latter.

Example 3.3:

$$\left.\begin{aligned} Q_t &= \alpha_0 + \alpha_1 P_t + \alpha_2 I_t + u_t \\ Q_t &= \beta_0 + \beta_1 P_t + \varepsilon_t \end{aligned}\right\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.9)$$

Where, Q_t and P_t are quantity and price which are endogenous,I_t is income and is exogenous in the first equation, u_t is error term of the first equation and ε_t is error term of the second equation. Then, given equation(3.9)

i. Differentiate the demand function and the supply function?
ii. Determine the identification condition of equation system?
iii. Find the estimators of the structural equation using ILS?

Solution

i. The first equation is demand function (Why?). As a result, the second equation has to be supply function.

ii. In the first equation, there is no excluded variable; hence the demand function is under-identified. But, Income,I_t, is excluded from the second equation; hence supply function is just-identified.

iii. Since the second equation is just identified, we can apply ILS and solve for Q_t and P_t in this equation as in the following.

At equilibrium Supply = Demand. Hence, substitute Q_t from the first equation in to Q_t of the second equation.

$$\alpha_0 + \alpha_1 P_t + \alpha_2 I_t + u_t = \beta_0 + \beta_1 P_t + \varepsilon_t$$

$$\alpha_1 P_t - \beta_1 P_t = \beta_0 - \alpha_0 - \alpha_2 I_t + \varepsilon_t - u_t$$

$$P_t(\alpha_1 - \beta_1) = \beta_0 - \alpha_0 - \alpha_2 I_t + \varepsilon_t - u_t \quad \text{Where, } \alpha_1 \neq \beta_1 \text{ (why?)}$$

$$P_t = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{-\alpha_2}{\alpha_1 - \beta_1} I_t + \frac{\varepsilon_t - u_t}{\alpha_1 - \beta_1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.9b)$$

$$P_t = \pi_0 + \pi_1 I_t + v_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.10)$$

Where, $\pi_0 = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.11)$

$\pi_1 = \frac{-\alpha_2}{\alpha_1 - \beta_1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.12)$

$v_t = \frac{\varepsilon_t - u_t}{\alpha_1 - \beta_1}$

To estimate Q_t, substitute the estimate of P_t from equation (3.9b) in to either the demand or supply function of equation (3.9):

$$Q_t = \beta_0 + \beta_1 P_t + \varepsilon_t$$

$$Q_t = \beta_0 + \beta_1\left(\frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{-\alpha_2}{\alpha_1 - \beta_1} I_t + \frac{\varepsilon_t - u_t}{\alpha_1 - \beta_1}\right) + \varepsilon_t$$

$$= \beta_0 + \beta_1\left(\frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1}\right) + \frac{-\beta_1\alpha_2}{\alpha_1 - \beta_1} I_t + \beta_1\left(\frac{\varepsilon_t - u_t}{\alpha_1 - \beta_1}\right) + \varepsilon_t$$

$$= \frac{\beta_0\alpha_1 - \beta_0\beta_1 + \beta_1\beta_0 - \beta_1\alpha_0}{\alpha_1 - \beta_1} + \frac{-\beta_1\alpha_2}{\alpha_1 - \beta_1} I_t + \frac{\beta_1\varepsilon_t - \beta_1 u_t + \alpha_1\varepsilon_t - \beta_1\varepsilon_t}{\alpha_1 - \beta_1}$$

Simplifying this gives:

$$Q_t = \frac{\beta_0\alpha_1 - \beta_1\alpha_0}{\alpha_1 - \beta_1} + \frac{-\beta_1\alpha_2}{\alpha_1 - \beta_1} I_t + \frac{\alpha_1\varepsilon_t - \beta_1 u_t}{\alpha_1 - \beta_1} \quad \ldots\ldots\ldots\ldots\ldots\ldots (3.13)$$

$$Q_t = \gamma_0 + \gamma_1 I_t + w_t \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.14)$$

Where, $\gamma_0 = \frac{\beta_0\alpha_1 - \beta_1\alpha_0}{\alpha_1 - \beta_1}$ $\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.15)$

$$\gamma_1 = \frac{-\beta_1\alpha_2}{\alpha_1 - \beta_1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.16)$$

$$w_t = \frac{\alpha_1\varepsilon_t - \beta_1 u_t}{\alpha_1 - \beta_1}$$

Look once again back to the structural equations (3.9). It consists of five structural coefficients/parameters; namely $\alpha_0, \alpha_1, \alpha_2, \beta_0, and\ \beta_1$. But, there are only four equations to estimate those structural coefficients, namely, the four reduced-form coefficients π_0, π_1, γ_0 and γ_1 given by equations (3.11), (3.12), (3.15) and (3.16) respectively.

Since the number of coefficients in the structural equations (=5) is greater than the number of coefficients in the reduced form equations (=4), unique solution of all the structural coefficients is not possible.(Relate this from your linear algebra lessons).

Question: Why is that the number of parameters in the structural equation is greater than the number of equations in the reduced form equation in this example?

As a result, ***only*** the parameters of the supply function can be identified as:

$$\beta_0 = \gamma_0 - \beta_1\pi_0 \text{ and } \beta_1 = \frac{\gamma_1}{\pi_1} \quad (3.17)$$

Remember from *chapter of econometrics I* that estimating equation (3.10) using OLS results:

$$\hat{\pi}_1 = \frac{\sum p_t\, i_t}{i_t^2} \quad (3.18)$$

$$\hat{\pi}_0 = \bar{P} - \hat{\pi}_1\bar{I}. \quad (3.19)$$

Similarly, estimating equation (3.14) using OLS results:

$$\hat{\gamma}_1 = \frac{\sum q_t\, i_t}{i_t^2} \quad (3.20)$$

$$\hat{\gamma}_0 = \bar{Q} - \hat{\gamma}_1\bar{I} \quad (3.21)$$

Where, the lowercase letters, as usual, denote deviations from mean value, and $\bar{Q}$, $\bar{P}$ *and* $\bar{I}$are the sample mean values of *Q* and *P and I.*

Substituting estimates of the reduced-form coefficients in to equation (3.17) gives the ILS estimators of structural equation parameters for supply functions:

$$\hat{\beta}_1 = \frac{\hat{\gamma}_1}{\hat{\pi}_1} \text{ and } \hat{\beta}_0 = \hat{\gamma}_0 - \hat{\beta}_1\hat{\pi}_0 = \hat{\gamma}_0 - \frac{\hat{\pi}_0}{\hat{\pi}_1}\hat{\gamma}_1 \quad (3.22)$$

But, as far as the demand function is under-identified, there is no unique way of estimating the parameters of, and remains under-identified.

Self Test 1:

If estimation of equations (3.10) and (3.14) gives the following result: *Source: (undisclosed)*

$$\hat{P}_t = 72.3 + 0.0043I_t$$

$$\hat{Q}_t = 84.07 + 0.002I_t$$

Assuming all estimators are statistically significant, find the ILS estimates of $\hat{\beta}_0$, $\hat{\beta}_1$ for the supply function in equation (3.9)?

Answer: $\hat{\beta}_0 = 51.05 \hat{\beta}_1 = 0.456$

So, the estimated ILS regression is: $\hat{Q}_t = 51.05 + 0.456P_t$

C. Two-Stage Least Squares (2SLS)

If we have over-identified equation(s) in an SEM,

- ✓ OLS will not be appropriate because the existence of endogenous variable(s) on the right side of an equation will give biased and inconsistent estimated due to endogeniety problem.(*Refer to section 3.2*)
- ✓ ILS estimation will not be appropriate because it will give more than one estimate for a single coefficient, for example for β_1 in equation (3.23 below).

Therefore, other estimation techniques have to be used such as ***2SLS*** discussed as under.

Illustration

Consider the following SEM:

$$\left.\begin{aligned} Income &= \alpha_0 + \alpha_1 Asset + \alpha_2 Experience + \alpha_2 Education + u_i \\ Asset &= \beta_0 + \beta_1 Income + \varepsilon_i \end{aligned}\right\} \ldots\ldots\ldots\ldots\ldots (3.23)$$

Where, experience and level of education (both measured in years) are assumed exogenous.

Applying the order condition of identification, we can see that the income equation is under-identified whereas the asset equation is over-identified.

If we can find a variable, Z, which satisfies the following conditions:

i. highly correlated with the endogenous variable (income in this case)
ii. But, uncorrelated with error term (ε_i in this case)

Then, we can substitute Z in place of income in the second equation and estimate the asset model using OLS directly. This means since Z is correlated with income, it can serve as a "proxy" or "substitute" for income. Z is then known as ***instrumental variable (IV)*** and the estimation is called instrumental variable estimation.

But, sometimes we may find more than one exogenous variables (like $Experience$ and $Education$ in this example) which satisfy the above two conditions. In this case, which variable shall we use as a proxy? The answer is, as far as these variables are uncorrelated with the error term (s), any linear combination of these exogenous variables $Experience$ and $Education$ is also uncorrelated with the error term(s), and can be used as a valid instrumental variable for income. That is, to find the best instrumental variable for income, we choose the linear combination $Experience$ and $Education$ that is most highly correlated with income using the following:

$$Income = \pi_0 + \pi_1 Experience + \pi_2 Education + v \quad \ldots\ldots (3.24)$$

Where, $E(v) = 0$, Cov$(Experience, v) = 0$, Cov$(Education, v) = 0$

This is the concept of two-stage least squares (2SLS) estimation. As the name 2SLS implies estimation of SEM using 2SLS involves two stages. These are:

Stage 1: Estimate the endogenous variable using all exogenous variable of the SEM

To get rid of the likely correlation between endogenous variable(*income)* and the error term,ε_i, we find the best linear combination among *all* the exogenous variables in the ***whole system, not just that equation***. The linear combination of Experience and Education in equation (3.24), which we call$Income^*$ becomes*:*

$$Income^{*} = \pi_0 + \pi_1 Experience + \pi_2 Education \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.25)$$

But, since we don't know the exact value of$Income^{*}$,we can only estimate, using OLS, by regressing income on experience and education as:

$$\widehat{Income} = \hat{\pi}_0 + \hat{\pi}_1 Experience + \hat{\pi}_2 Education \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.26)$$

Or,

$$Income = \widehat{Income} + \hat{v} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.27)$$

Then, conduct a joint significance of variables for equation (3.26) using F-test.

If the variable are found jointly significant (not larger than 5%), then use the fitted values of income, $\widehat{Income}$, as an IV. All the above tasks involve *Stage-I*.

Stage 2: Substitute the estimated value of the endogenous variable obtained from Stage-1 and estimate the over-identified model of the SEM

Substitute equation (3.27) in the second equation of (3.23), and estimate the asset model using OLS method yields:

$$Asset = \widehat{Income} + \hat{v} + \varepsilon_i = \beta_0 + \beta_1 \widehat{Income} + wi. \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.28)$$

Where, $wi = \hat{v} + \varepsilon_i$

Equation (3.28) is very similar in appearance, with the second equation of (3.23) with the only difference being that actual value of *income* is replaced by its estimated value, $\widehat{Income}$, using exogenous variables .

***Advantage of doing so**:*

- ✓ The error term,ε_i,is correlated with income, but not $\widehat{Income}$.Why?
- ✓ As a result, OLS estimation of on equation (3.28), will give unbiased and consistent estimate unlike equation (3.23).

Features of 2SLS

i. It can be applied to an individual equation in the SEM without directly taking into account any other equation(s) in the system. For this reason the method has been used extensively in practice, for solving econometric models involving a large number of equations.
ii. ILS provides multiple estimates of a parameter in the over identified equations, but 2SLS provides only one.
iii. It is easy to apply because all one needs to know is the total number of exogenous or variables in the system, without knowing any other variables in the system.
iv. It can also be applied to exactly identified equations and gives identical estimates with ILS.
v. If the R^2 values in the reduced-form regressions (that is, Stage-1 regressions) are very high, say, in excess of 0.8, the classical OLS estimates and 2SLS estimates will be very close. Why?
vi. In reporting the ILS regression, we did not state the standard errors of the estimated coefficients. But we can do this for the 2SLS estimates because the structural coefficients are directly estimated from the second-stage (OLS) regressions, though there may be some modification (see$wi = \hat{v} + \varepsilon_i$ under equation (3.28) above).

Endogenous variable

Exogenous variable

Full information methods

Identification

Indirect least square

Instrumental variable

Observed shifters

Order condition

Rank condition

Reduced form equation

Simultaneity bias

Simultaneous equation

Structural equation

Systems method

Two stage least squares

Unobserved shifters

Chapter Four

Introduction to Panel Data Regression Models

Dear learner!, you have learned about three types of data in econometrics I. these are cross sectional data, time series data and panel data. You have learned models involving time series data and cross sectional data in the previous lessons. In this chapter will learn about models involving panel data.

In section 4.1, you will learn about types of panel data and advantages of panel data. In this section, you will look at the endogenous variable and exogenous variable. The second section of the chapter covers estimation of panel data. Here, you will learn about the two techniques of estimating panel data. These are the fixed-effects approach and the random-effects approach.

Objective of the chapter

At the end of the chapter you are expected to:

- Be familiar with the nature of panel data.
- Distinguish between panel data and pooled data.
- Understand the characteristics panel data.
- Know the advantages of using panel data over cross-sectional or time series data.
- Understand what does it mean by the fixed-effects approach of estimating panel data.
- Identify the different possibilities/assumptions of fixed-effects approach
- Differentiate between the with-in estimator, between-estimator and over-all estimator.
- Understand what does it mean by the random-effects approach of estimating panel data.
- Know how to estimate panel data using random-effects approach

What is panel data?

Dear learner, what do you know about simultaneous equations from your lessons of econometrics I and statistics courses? Write what comes to your mind on the following spaces ----------
---
---.

4.1. Introduction

So far, you have covered regression analysis using either ***cross sectional*** or ***time series*** data alone. Although these two cases arise often in applications, cross-sectional data across time-a situation where the data set has both cross sectional and time series dimensions-are being used more and more often in empirical research.

We know that, multiple regression is a powerful tool for controlling for the effect of variables on which we have data. If data is not available for some of the variables, however, they cannot be included in the regression and the OLS estimators of the regression coefficients could have omitted variables bias.

This chapter describes a method for controlling for some types of omitted variables without actually observing them. This method requires a specific type of data, called ***panel data,*** in which each observational unit, or entity, is observed at two or more time periods. By studying changes in the dependent variable over time, it is possible to eliminate the effect of omitted variables that differ across entities but are constant over time.

Basically sampling cross sectional data across time involves two kinds of data sets: Panel data and pooled data. Panel data (also known as longitudinal or cross-sectional time-series data) is a dataset in which the behavior of entities are observed across time. These entities could be states, companies, individuals, countries, etc.

The structure of panel dataset looks like the following.

country	year	Y	X1	X2	X3
1	2000	6.0	7.8	5.8	1.3
1	2001	4.6	0.6	7.9	7.8
1	2002	9.4	2.1	5.4	1.1
2	2000	9.1	1.3	6.7	4.1
2	2001	8.3	0.9	6.6	5.0
2	2002	0.6	9.8	0.4	7.2
3	2000	9.1	0.2	2.6	6.4
3	2001	4.8	5.9	3.2	6.4
3	2002	9.1	5.2	6.9	2.1

Panel data allows you to control for variables you cannot observe or measure like cultural factors or difference in business practices across companies; or variables that change over time but not across entities (i.e. national policies, federal regulations, international agreements, etc.). That is, it accounts for individual heterogeneity.

4.1.1. Pooled data

This involves sampling randomly from a large population at different points in time. Samples drawn in different times may ***not*** be the ***same***. The advantage here is samples consist of *independently* sampled observations. This was also a key aspect in our analysis of cross-sectional data: among other things, it rules out correlation in the error terms across different observations. Pooling is helpful only if the relationship between the dependent variable and at least some of the independent variables remain constant over time.

4.1.2. Panel data/Longitudinal data

In panel data the ***same*** cross-sectional units are surveyed over time. The problem here is if we lose any observation for whatever reason (e.g. because of death), we can no longer use panel data.

Despite the existence of some variations, both pooled data and panel data essentially connote movement over time of cross-sectional units. We will, therefore, use the term panel data in a generic sense to include one or more of these terms.

Panel Data Model Examples

- **Labor economics**: effect of education on income, with data across time and individuals.
- **Economics**: effects of income on savings, with data across years and countries.

Panel data characteristics

1. Panel data provide information on individual behavior, both across individuals and over time – they have both cross-sectional and time-series dimensions.
2. Panel data include *N* individuals observed at *T* regular time periods.
3. Panel data can be balanced when all individuals are observed in all time periods ($T_i = T$for all i) or unbalanced when individuals are not observed in all time periods ($T_i \neq T$).
4. We assume correlation (clustering) over time for a given individual, with independence over individuals.
    - Example: the income for the same individual is correlated over time but it is independent across individuals.

Panel data types

- Short panel: many individuals and few time periods (we use this case in class)
- Long panel: many time periods and few individuals
- Both: many time periods and many individuals

Regressors

- Varying regressors x_{it}
  - annual income for a person, annual consumption of a product
- Time-invariant regressors $x_{it} = x_i$ for all t.
  - gender, race, education
- Individual-invariant regressors $x_{it} = x_t$ for all i.
  - time trend, economy trends such as unemployment rate

Variation for the dependent variable and regressors

- Overall variation: variation over time and individuals.
- Between variation: variation between individuals.
- Within variation: variation within individuals (over time).

Id	Time	Variable	Individual mean	Overall mean	Overall deviation	Between deviation	Within deviation
i	t	x_{it}	$\bar{x}_i$	$\bar{x}$	$x_{it} - \bar{x}$	$\bar{x}_i - \bar{x}$	$x_{it} - \bar{x}_i$
1	1	9	10	20	-11	-10	-1
1	2	10	10	20	-10	-10	0
1	3	11	10	20	-9	-10	1
2	1	20	20	20	0	0	0
2	2	20	20	20	0	0	0
2	3	20	20	20	0	0	0
3	1	25	30	20	5	10	-5
3	2	30	30	20	10	10	0
3	3	35	30	20	15	10	5

Individual mean $\bar{x}_i = \frac{1}{T}\sum_t x_{it}$

Overall mean $\bar{x} = \frac{1}{NT}\sum_i \sum_t x_{it}$

- Time-invariant regressors (race, gender, education) have zero within variation.
- Individual-invariant regressors (time, economy trends) have zero between variation.

- We need to check the data to see if the between or within variation is larger for each variable.

4.1.3. Advantages of using panel data

1. The techniques of panel data estimation can take into account heterogeneity relating to firms, states, countries, etc., over time, explicitly by allowing for individual-specific variables.

2. Increases precision estimators with more power of test statistics: By combining time series of cross-section observations, panel data give "more informative data, more variability, less collinearity among variables, more degrees of freedom and more efficiency."

3. By studying the repeated cross section of observations, panel data are better suited to study the *dynamics of change.* Example: Spells of unemployment, job turnover, and labor mobility.

4. Panel data can better detect and measure effects that simply cannot be observed in pure cross-section, or pure time series data. *Ex*: the effects of minimum wage laws on employment and earnings.

5. Panel data enables us to study more complicated behavioral models. For example, phenomena such as economies of scale and technological change.

6. By making data available for several thousand units, panel data can minimize the bias that might result if we aggregate individuals or firms into broad aggregates.

Panel data models have the following general form:

$$Y_{it} = \alpha_0 + \alpha_1 X_{1it} + \alpha_2 X_{2it} + \cdots + \alpha_n X_{nit} + u_{it} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.1)$$

Where,i stands for the i[th] cross-sectional unit in the time period t. $X's$ are regressors,Yas usual refers to the dependent variable. $u_{it}$is the error term assumed to follow the classical assumption of zero mean and constant variance. As a matter of convention, we usually let *i* denote the cross-section identifier and *t* the time identifier.

According to the above denotation, Y_{it}shows the value of the dependent variable for the ith cross-sectional unit in the time period t. X_{1it} shows the value of the first independent variable for the ith cross-sectional unit in the time period t, etc.

Illustration

Suppose you want to estimate, "The determinant of real investment (I) in Ethiopian manufacturing industries from 2001 to 2015". For this, say, you selected four firms: Muger Cement Factory (MCF), Abay Steel and Plastic Factory (APF), Kombolcha Textile Factory (KTF) and Dashen Brewery (DB). So, you will collect a 15 years data for each industry. Suppose, again, the explanatory variables are the real value of the firm (X_1) and real capital stock (X_2).

- ✓ If you want to run ***time series*** regressions only, you can have ***four*** regressions, one for each company.
- ✓ If you want to run ***cross sectional*** regression only, you can have ***fifteen*** regressions, one for each year.
- ✓ But, if you want to run ***panel data*** regression, you can have ***one*** regression having 15*4=60 observations.

Individual-specific effects model

- ➢ We assume that there is unobserved heterogeneity across individuals captured by a_i.

  Example: unobserved ability of an individual that affects wages.

- ➢ The main question is whether the individual-specific effects a_i are correlated with the regressors. If they are correlated, we have the fixed effects model. If they are not correlated, we have the random effects model.

4.2. Estimation of Panel Data Regression Models

4.2.1. The Fixed Effects Approach

- The fixed effects model allows the individual-specific effects to be correlated with the regressors x.
- We include $a_i$as intercepts.
- Each individual has a different intercept term and the same slope parameters.

$$y_{it} = a_i + x_{it}\beta + u_{it}$$

- We can recover the individual specific effects after estimation as:

$$\hat{a_i} = \bar{y_i} - \bar{x_i}\hat{\beta}$$

  In other words, the individual-specific effects are the leftover variation in the dependent variable that cannot be explained by the regressors.
- Time dummies can be included in the regressors x.

Estimation of (4.1) depends on the ***assumptions*** we make about the *intercept, the slope coefficients, and the error term*. There are several possibilities/assumptions

Case 1: The intercept and slope coefficients are constant across time and space and the error term captures differences over time and individuals

Case 2: The slope coefficients are constant but the intercept varies over individuals.

Case 3: The slope coefficients are constant but the intercept varies over individuals and time.

Case 4: Both the intercept as well as slope coefficients vary over individuals.

Case 5: Both the intercept as well as slope coefficients vary over individuals and time.

Note that complexity and reality increase as we move from case 1 to case 5.

Case 1: All Coefficients Constant across Time and Individuals

This is the simplest approach which requires OLS regression by disregarding the space and time dimensions of the pooled data.

If you have a model under illustration 1, you will get a result which looks like the following:

$$I = -63 + 0.11X_1 + 0.3X_2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.2)$$

Where, I is real investment of an industry, X_1 is real value of the firm, and X_2 is real capital stock.

Suppose the estimators are statistically significant and the slope coefficients for real value of the firm (X_1) and real capital stock (X_2) are expected to be positive and they are! This estimation is pretty easy. Nevertheless, this result assumes that the estimates of the intercept value as well as slope coefficients are the same for every industry: MCF, APF, KTF, DB.

For example, if the estimator of the coefficient of capital stock (X_2) is statistically significant, a unit increase in real capital stock increases real investment, on average, by 0.3 units across all firms, MCF, APF, KTF, DB.

This means, the model does not take in to account the specific nature of each company. This means, for example, a unit change in real capital stock (X_2) has the same effect in each industry, which is unlikely. Is it?

Therefore, what we need to do is find some way to take into account the specific nature of the four companies; case 2 can better do this.

Case 2: The slope coefficients are constant but the intercept varies across individuals: The Fixed Effects or Least-Squares Dummy Variable (LSDV) Regression Model

This assumption tries to take into account two possibilities:

A. **Cross section effect**: the "individuality" of each company or each cross-sectional unit

B. **Time Effect:** such as changes in technology, government policy, war (let it not be, indeed), etc.

All in all, these will be incorporated by letting the intercept vary for each company but still assumes that the slope coefficients are constant across firms. This is done by putting the subscript *i* on the intercept term (α_0) as:

$$Y_{it} = \alpha_{0i} + \alpha_1 X_{1it} + \alpha_2 X_{2it} + \cdots + \alpha_n X_{nit} + u_{it} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.3)$$

In equation (4.3), α_{0i} shows the intercept varies across firms but not time. Equation (4.3) is known as the **fixed effects** regression model (**FEM).** But, the equation assumes that slope coefficients remain fixed.

A. Incorporating "individuality" of each cross-sectional unit

Question: How do we actually allow for the intercept to vary between companies?

Answer: Introduce dummy variables called ***differential intercept dummies*** in the equation. This is shown on equation (4.4):

$$Y_{it} = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + \alpha_1 X_{1it} + \alpha_2 X_{2it} + u_{it} \ldots\ldots\ldots\ldots\ldots\ldots (4.4)$$

Where, $D_{1i} = 1$ if the observation belongs to Muger Cement Factory (MCF)

=0, otherwise

$D_{2i} = 1$ if the observation belongs to Abay Steel and Plastic Factory (APF),

=0, otherwise

$D_{3i} = 1$ if the observation belongs to Kombolcka Textile Factory (KTF)

= 0, otherwise

Since we are using dummies to estimate the fixed effects, equation (4.4) is known as the **least-squares dummy variable (LSDV) model** or **covariance model.** So, the terms fixed effects and LSDV can be used interchangeably.

Note that, since equation (4.4) has a constant term, we have to omit one category: Dashen Brewery (DB) in this case (Why?). Hence, the *differential intercept* coefficients β_1, β_2, and

β_3 tell us by how much the intercepts of MCF, APF, and KTF differ from the intercept of the omitted category (DB) respectively.

Usually, if a company owns unique features, intercepts in equation (4.4) are more likely to be statistically significant.

Example: suppose estimation of equation (4.4) gives the following result, where values in brackets are standard errors.

$$Y_{it} = \underset{(0.862)}{4.8} + \underset{(0.244)}{0.6}\, D_{1i} - \underset{(0.112)}{0.35}\, D_{2i} + \underset{(0.32)}{0.4}\, D_{3i} + \underset{(0.027)}{0.5}\, X_{1it} + \underset{(0.061)}{0.23}\, X_{2it}$$

Find the intercept of each factory

Solution

- The constant term is the intercept of the omitted category, DB. Hence, the intercept for DB is equal to 4.8.
- D1 is the dummy for, MCF and its coefficient is statistically significant. Hence, the intercept for MCF is equal to 4.8+0.6=5.4
- D2 is the dummy for, APF and its coefficient is statistically significant. Hence, the intercept for ABF is equal to 4.8-0.35=4.45
- D3 is the dummy for, KTF and its coefficient is not statistically significant. Hence, there is no statistical difference between the intercept of DB and KTF.

If both equations (4.2) and (4.4) are significant we may choose the better model based on significance of t-values, Durbin-Watson d-statistic, the value of R^2 etc. But, R^2 will be higher in (4.4) (why?). Nonetheless, the formal test is to use restricted F-test we discussed in chapter three of econometrics I. To remind you, use:

$$F = \frac{RSS_R - RSS_{UR}}{Number\ of\ restrictions} \Big/ \frac{RSS_{UR}}{n-k} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(4.5)$$

Where, RSS_R is residual sum of squares for restricted model (equation 4.2), RSS_{UR} denotes residual sum of squares for unrestricted model (equation 4.4), number of restrictions equal number of dummies, n is sample size, k number of parameters.

B. Time effect

This can be accounted by introducing time dummies for each year. In the above example since we are analyzing a 15 years-data (from 2001 to 2015), we introduce 14 time dummies. Why? The model becomes:

$$Y_{it} = \gamma_0 + \gamma_1 D_{01} + \gamma_2 D_{02} + \cdots + \gamma_{14} D_{14} + \alpha_1 X_{1it} + \alpha_2 X_{2it} + u_{it} \ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.6)$$

Where, D_{01}, D_{02} , …, D_{14} represent dummies form 2001, 2002, …, 2014 respectively

D_{01} = 1 for observations of MCF, APF, KTF, DB in year 2001,

= 0 for the rest of years

D_{02} = 1 for observations of MCF, APF, KTF, DB in year 2002, and

=0 for the rest of years

Etc.

We can, then, estimated equation (4.6) and test statistical significance of coefficients, $\gamma_i s$.

Case 3: The slope coefficients are constant but the intercept varies over individuals and time.

This implies we have different intercepts in each year for each company. This can be accounted by combining equations (4.4) and 4.6)

$$Y_{it} = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + \gamma_0 + \gamma_1 D_{01} + \gamma_2 D_{02} + \cdots + \gamma_{14} D_{14} + \alpha_1 X_{1it} + \alpha_1 X_{1it} + \alpha_2 X_{2it} + u_{it} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.7)$$

The omitted categories are DB and the year 2015

By first estimating equation (4.7), we can test for individual company effect($\beta_i s$) as well as time effect ($\gamma_i s$).

From equation (4.7), for example, If all estimators are statistically significant,

The intercept for DB in 2015 equals β_0+ γ_0

The intercept for DB in 2002 equals $\beta_0 + \gamma_0 + \gamma_2$

The intercept for MCF in 2015 equals $\beta_0 + \beta_1 + \gamma_0$

The intercept for MCF in 2002 equals $\beta_0 + \beta_1 + \gamma_0 + \gamma_2$

The intercept for KTF in 2001 equals $\beta_0 + \beta_3 + \gamma_0 + \gamma_2 + \gamma_1$

Etc.

Using this method, we may allow intercepts to vary among cross sectional units and over time.

Case 4: Both the intercept as well as slope coefficients vary over individuals.

This assumes, for the above example, that all industries MCF, APF, KTF, and DB have different investment functions. Put in other words, the effect of each regressor, the real value of the firm (X_1) and real capital stock (X_2) in this case, is different in each factory. To account this, we multiply each of the company dummy (D_{1i},D_{2i}andD_{3i}) by each regressor ($X_{1it}$and X_{2it}) which gives an additional of 6 variables given as:

$$Y_{it} = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + \alpha_1 X_{1it} + \alpha_2 X_{2it} + \gamma_1 (D_{1i} X_{1it}) + \gamma_2 (D_{1i} X_{2it}) + \gamma_3 (D_{2i} X_{1it}) + \gamma_4 (D_{2i} X_{2it}) + \gamma_5 (D_{3i} X_{1it}) + \gamma_6 (D_{3i} X_{2it}) + u_{it} \ldots . (4.8)$$

Where, β_i are differential intercept coefficients, and γ_i are differential slope coefficients, DB is the omitted cross sectional unit and 2015 is omitted year

The interpretation of intercepts is the same as that of equation 4.7. Thus, let's see only the slopes. If all estimators are statistically significant,

The slope coefficient of the real capital stock (X_2) for DB (the omitted category) equals α_2.

The slope coefficient of the real capital stock (X_2) for APF equals $\alpha_2 + \gamma_4$

The slope coefficient of the real value of the firm (X_1) for KTF equals $\alpha_1 + \gamma_5$

Etc.

Using this we can allow variation of slope coefficients among cross sectional units.

Problems of Fixed Effects Approach, or LSDV model

i. Incorporating too many dummy variables will erode the degree of freedom down. This reduces the degree of precision.
ii. Existence of too many variables will more likely cause multicollinearity problem.
iii. Error term may exhibit different behavior for different units. For example, it may be autocorrelated for KTF, where as it may not in MCI, etc.
iv. Problems with the existence of time-invariant variables

4.2.2. The Random Effects Approach

- The random-effects model assumes that the individual-specific effects a_i are distributed independently of the regressors.
- We include a_i in the error term.
- Each individual has the same slope parameters and a composite error term $\varepsilon_{it} = a_i + e_{it}$

$$y_{it} = x_{it}\beta + (a_i + e_{it})$$

As it is said earlier fixed effect (covariance model) is straightforward to apply, but require loss of large degree of freedom in the presence of large cross-sectional units. More specifically, if the dummy variables do in fact represent a lack of knowledge about the (true) model, we can express this ignorance through the disturbance term and this is what the **error components model (ECM) or random effects model (REM)** suggest**.** Let's start with one of the previous models:

$$Y_{it} = \alpha_{0i} + \alpha_1 X_{1it} + \alpha_2 X_{2it} + u_{it} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.9)$$

The random effect assumes that instead of treating α_{0i}as fixed, we assume that it is a random variable with a mean value of α_0 (no subscript *i* here). And the intercept value for an individual company can be expressed as:

$$\alpha_{0i} = \alpha_0 + \varepsilon_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.10)$$

Where, εi is a random error term with a mean value of zero and variance of σ_{ε}^{2}.

Substitute equation (4.10) in to (4.9),

$$Y_{it} = \alpha_0 + \varepsilon_i + \alpha_1 X_{1it} + \alpha_2 X_{2it} + u_{it}$$

$$Y_{it} = \alpha_0 + \alpha_1 X_{1it} + \alpha_2 X_{2it} + v_{it} \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots .. (4.11)$$

Where, $v_{it} = \varepsilon_i + u_{it}$, ε_i accounts cross-section, or individual-specific, error component, and , u_{it} accounts the combined time series and cross-section error components. Bothε_i and u_{it} are assumed to fulfill the basic assumptions of classical linear regression model (CLRM). In addition, the correlation between them should be zero.

In FEM each cross-sectional unit has its own (fixed) intercept value, in all *N* such values for *N* cross-sectional units. In REM (ECM), on the other hand, the intercept α_0 represents the mean value of all the (cross-sectional) intercepts and the error component εi represents the (random) deviation of individual intercept from this mean value. Note that εi is not directly observable, and is known as an **unobservable,** or **latent, variable.**

Let's go back to the previous example. This model says that the four firms included in our sample are drawn from a much larger population of such companies and that they have a common mean value for the intercept (= α_0) and the individual differences in the intercept values of each company are reflected in the error term ε_i.

Note also that $Var\ (v_{it}) = \sigma_u^2 + \sigma_\varepsilon^2$ (remember this from previous semester). Hence, if $\sigma_\varepsilon^2 = 0$, there is no difference between models (4.1) and (4.9). But, though v_{it} is homoscedastic, it is autocorrelated. Unless we account for this problem, estimates though are unbiased, will be inefficient. To this end, the generalized least square (GLS) is mostly used.

Example: suppose the random effect estimation of the four firms is given as under following result:

Tab1e: 4.1: ECM estimation result

Variable	***Coef.***	***Std. error***	***t-statistic***	***p* value**
Intercept	*4.84*	*2.11*	*2.294*	*0.032*
X1	*0.16*			
X2	*0.24*			
Random Effect				
MCF	*0.612*			
APF	*-0.542*			
KTF	*0.383*			
DB	*0.453*			
		R^2*=0.9124 (GLS)*		

Note from table (4.1) that:

- ✓ The sum of the random effect values given for the four companies will be zero (why?).
- ✓ The mean value of the random error component, εi, is the common intercept value of −4.84.
- ✓ The random effect value of MCF of 0.612 tells us by how much the random error component of MCF differs from the common intercept value. Similar interpretation applies to the other three values of the random effects

4.2.3: Comparison of FEM Vs ECM

☞ **Which is better?**

To decide between fixed or random-effects you can run a Hausman test where the null hypothesis is that the preferred model is random effects vs. the alternative the fixed effects. It basically tests whether the unique errors (ε_i) are correlated with the regressors, the null hypothesis is they are not.

Run a fixed effects model and save the estimates, then run a random model and save the estimates, then perform the test. If the p-value is significant (for example <0.05) then use fixed effects, if not use random effects.

- ✓ If it is assumed that εi and the X's are *uncorrelated*, ECM may be appropriate, whereas if εi and the X's are *correlated*, FEM may be appropriate.

☞ ***Which one has to be chosen?***

- ✓ If the number of time series data is large and the number of cross-sectional units is small, there is likely to be little difference in the values of the parameters estimated by FEM and ECM. Hence the choice here is based on computational convenience. On this score, FEM may be preferable.
- ✓ If the number of time series data is small and the number of cross-sectional units is large, there is significant difference and FEM is appropriate. Note that α_{0i}depends on number of cross sectional units.
- ✓ If the number of time series data is small and the number of cross-sectional units is large, and if the assumptions underlying ECM hold, ECM estimators are more efficient than FEM estimators.
- ✓ If εi and one or more X's are *correlated*, ECM will be biased. So what?

Tips for further reading

Policy analysis using Panel data model

Panel (pooled) data models are often used to evaluate policy measures. The most common one is the difference in difference estimator.

Difference-in-difference estimator (DID): For example, if your instructor wants to estimate the effect of tutorial on the result of second year economics students, he may give a test for all students before tutorial. Then, he may divide the students in to two groups: (1) those who are going to be given the tutorial land (2) those who will not take the tutorial. Then, after the tutorial is given, he will give another test for both groups. Finally, he will compare the ***average*** score for both groups before and after the tutorial is given, by controlling the effect of time itself. Since this action measures the effect of the "treatment" or policy on the average outcome, it is **also called average treatment effect (ATE)[4].**

[4] You can read from other books such as Wooldridge (2004) for further.

Check list of important points to remember from the chapter.

Between estimator

Cross section effect

Difference-in-difference

Fixed effects

Longitudinal data

Over all estimator

Panel data

Pooled data

Random effects

Time effect

Within estimator

References

Gujarati, D. N. (2004). *Basic Econometrics*, 4th edition, McGraw-Hill

Koutsoyiannis, A. (2001). *Theory of Econometrics*, Palgrave: New York.

Maddala, G. S. (1992). *Introduction to Econometrics*, 2nd edition, Macmillan.

Maddala, G. S. (1993). *Limited-dependent and qualitative variables in econometrics*, 2nd edition, Cambridge University press.

Salvatore D., Reagle. D. (2002). Statistics and econometrics, 2nd edition, McGRAW-HILL

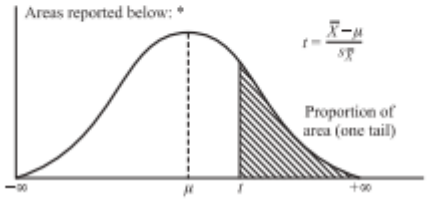
Wolters. J. (2007). *Introduction to Modern Time Series Analysis*, *publisher unknown*

Wooldridge, J. (2004). *Introductory Econometrics: A Modern Approach*, 2nd edition.

Appendices

Appendix 1:

Student's *t* Distribution



Proportions of Area for the *t* Distributions

df	*0.10*	*0.05*	*0.025*	*0.01*	*0.005*
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

*Example: For the shaded area to represent 0.05 of the total area of 1.0, value of *t* with 10 degrees of freedom is 1.812

Source: From Table III of Fisher and Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, 6th ed., 1974, published by Longman Group Ltd., London (previously published by Oliver & Boyd, Edinburgh), by permission of the authors and publishers.

Appendix 2: *The F-Distribution*

Values of *F* Exceeded with Probabilities of 5 and 1 Percent

df (denominator)	df (numerator)																							
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248	249	250	251	252	253	253	254	254	254
	4,052	**4,999**	**5,403**	**5,625**	**5,764**	**5,859**	**5,928**	**5,981**	**6,022**	**6,056**	**6,082**	**6,106**	**6,142**	**6,169**	**6,208**	**6,234**	**6,261**	**6,286**	**6,302**	**6,323**	**6,334**	**6,352**	**6,361**	**6,366**
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41	19.42	19.43	19.44	19.45	19.46	19.47	19.47	19.48	19.49	19.49	19.50	19.50
	98.49	**99.00**	**99.17**	**99.25**	**99.30**	**99.33**	**99.36**	**99.37**	**99.39**	**99.40**	**99.41**	**99.42**	**99.43**	**99.44**	**99.45**	**99.46**	**99.47**	**99.48**	**99.48**	**99.49**	**99.49**	**99.49**	**99.50**	**99.50**
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74	8.71	8.69	8.66	8.64	8.62	8.60	8.58	8.57	8.56	8.54	8.54	8.53
	34.12	**30.82**	**29.46**	**28.71**	**28.24**	**27.91**	**27.67**	**27.49**	**27.34**	**27.23**	**27.13**	**27.05**	**26.92**	**26.83**	**26.69**	**26.60**	**26.50**	**26.41**	**26.35**	**26.27**	**26.23**	**26.18**	**26.14**	**26.12**
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91	5.87	5.84	5.80	5.77	5.74	5.71	5.70	5.68	5.66	5.65	5.64	5.63
	21.20	**18.00**	**16.69**	**15.98**	**15.52**	**15.21**	**14.98**	**14.80**	**14.66**	**14.54**	**14.45**	**14.37**	**14.24**	**14.15**	**14.02**	**13.93**	**13.83**	**13.74**	**13.69**	**13.61**	**13.57**	**13.52**	**13.48**	**13.46**
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68	4.64	4.60	4.56	4.53	4.50	4.46	4.44	4.42	4.40	4.38	4.37	4.36
	16.26	**13.27**	**12.06**	**11.39**	**10.97**	**10.67**	**10.45**	**10.29**	**10.15**	**10.05**	**9.96**	**9.89**	**9.77**	**9.68**	**9.55**	**9.47**	**9.38**	**9.29**	**9.24**	**9.17**	**9.13**	**9.07**	**9.04**	**9.02**
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.96	3.92	3.87	3.84	3.81	3.77	3.75	3.72	3.71	3.69	3.68	3.67
	13.74	**10.92**	**9.78**	**9.15**	**8.75**	**8.47**	**8.26**	**8.10**	**7.98**	**7.87**	**7.79**	**7.72**	**7.60**	**7.52**	**7.39**	**7.31**	**7.23**	**7.14**	**7.09**	**7.02**	**6.99**	**6.94**	**6.90**	**6.88**
7	5.59	4.74	4.34	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57	3.52	3.49	3.44	3.41	3.38	3.34	3.32	3.29	3.28	3.25	3.24	3.23
	12.25	**9.55**	**8.45**	**7.85**	**7.46**	**7.19**	**7.00**	**6.84**	**6.71**	**6.62**	**6.54**	**6.47**	**6.35**	**6.27**	**6.15**	**6.07**	**5.98**	**5.90**	**5.85**	**5.78**	**5.75**	**5.70**	**5.67**	**5.65**
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28	3.23	3.20	3.15	3.12	3.08	3.05	3.03	3.00	2.98	2.96	2.94	2.93
	11.26	**8.65**	**7.59**	**7.01**	**6.63**	**6.37**	**6.19**	**6.03**	**5.91**	**5.82**	**5.74**	**5.67**	**5.56**	**5.48**	**5.36**	**5.28**	**5.20**	**5.11**	**5.06**	**5.00**	**4.96**	**4.91**	**4.88**	**4.86**
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07	3.02	2.98	2.93	2.90	2.86	2.82	2.80	2.77	2.76	2.73	2.72	2.71
	10.56	**8.02**	**6.99**	**6.42**	**6.06**	**5.80**	**5.62**	**5.47**	**5.35**	**5.26**	**5.18**	**5.11**	**5.00**	**4.92**	**4.80**	**4.73**	**4.64**	**4.56**	**4.51**	**4.45**	**4.41**	**4.36**	**4.33**	**4.31**
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91	2.86	2.82	2.77	2.74	2.70	2.67	2.64	2.61	2.59	2.56	2.55	2.54
	10.04	**7.56**	**6.55**	**5.99**	**5.64**	**5.39**	**5.21**	**5.06**	**4.95**	**4.85**	**4.78**	**4.71**	**4.60**	**4.52**	**4.41**	**4.33**	**4.25**	**4.17**	**4.12**	**4.05**	**4.01**	**3.96**	**3.93**	**3.91**
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79	2.74	2.70	2.65	2.61	2.57	2.53	2.50	2.47	2.45	2.42	2.41	2.40
	9.65	**7.20**	**6.22**	**5.67**	**5.32**	**5.07**	**4.88**	**4.74**	**4.63**	**4.54**	**4.46**	**4.40**	**4.29**	**4.21**	**4.10**	**4.02**	**3.94**	**3.86**	**3.80**	**3.74**	**3.70**	**3.66**	**3.62**	**3.60**
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69	2.64	2.60	2.54	2.50	2.46	2.42	2.40	2.36	2.35	2.32	2.31	2.30
	9.33	**6.93**	**5.95**	**5.41**	**5.06**	**4.82**	**4.65**	**4.50**	**4.39**	**4.30**	**4.22**	**4.16**	**4.05**	**3.98**	**3.86**	**3.78**	**3.70**	**3.61**	**3.56**	**3.49**	**3.46**	**3.41**	**3.38**	**3.36**
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60	2.55	2.51	2.46	2.42	2.38	2.34	2.32	2.28	2.26	2.24	2.22	2.21
	9.07	**6.70**	**5.74**	**5.20**	**4.86**	**4.62**	**4.44**	**4.30**	**4.19**	**4.10**	**4.02**	**3.96**	**3.85**	**3.78**	**3.67**	**3.59**	**3.51**	**3.42**	**3.37**	**3.30**	**3.27**	**3.21**	**3.18**	**3.16**
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53	2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.21	2.19	2.16	2.14	2.13
	8.86	**6.51**	**5.56**	**5.03**	**4.69**	**4.46**	**4.28**	**4.14**	**4.03**	**3.94**	**3.86**	**3.80**	**3.70**	**3.62**	**3.51**	**3.43**	**3.34**	**3.26**	**3.21**	**3.14**	**3.11**	**3.06**	**3.02**	**3.00**
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48	2.43	2.39	2.33	2.29	2.25	2.21	2.18	2.15	2.12	2.10	2.08	2.07
	8.68	**6.36**	**5.42**	**4.89**	**4.56**	**4.32**	**4.14**	**4.00**	**3.89**	**3.80**	**3.73**	**3.67**	**3.56**	**3.48**	**3.36**	**3.29**	**3.20**	**3.12**	**3.07**	**3.00**	**2.97**	**2.92**	**2.89**	**2.87**
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42	2.37	2.33	2.28	2.24	2.20	2.16	2.13	2.09	2.07	2.04	2.02	2.01
	8.53	**6.23**	**5.29**	**4.77**	**4.44**	**4.20**	**4.03**	**3.89**	**3.78**	**3.69**	**3.61**	**3.55**	**3.45**	**3.37**	**3.25**	**3.18**	**3.10**	**3.01**	**2.96**	**2.98**	**2.86**	**2.80**	**2.77**	**2.75**
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38	2.33	2.29	2.23	2.19	2.15	2.11	2.08	2.04	2.02	1.99	1.97	1.96
	8.40	**6.11**	**5.18**	**4.67**	**4.34**	**4.10**	**3.93**	**3.79**	**3.68**	**3.59**	**3.52**	**3.45**	**3.35**	**3.27**	**3.16**	**3.08**	**3.00**	**2.92**	**2.86**	**2.79**	**2.76**	**2.70**	**2.67**	**2.65**
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.29	2.25	2.19	2.15	2.11	2.07	2.04	2.00	1.98	1.95	1.93	1.92
	8.28	**6.01**	**5.09**	**4.58**	**4.25**	**4.01**	**3.85**	**3.71**	**3.60**	**3.51**	**3.44**	**3.37**	**3.27**	**3.19**	**3.07**	**3.00**	**2.91**	**2.83**	**2.78**	**2.71**	**2.68**	**2.62**	**2.59**	**2.57**
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34	2.31	2.26	2.21	2.15	2.11	2.07	2.02	2.00	1.96	1.94	1.91	1.90	1.88
	8.18	**5.93**	**5.01**	**4.50**	**4.17**	**3.94**	**3.77**	**3.63**	**3.52**	**3.43**	**3.36**	**3.30**	**3.19**	**3.12**	**3.00**	**2.92**	**2.84**	**2.76**	**2.70**	**2.63**	**2.60**	**2.54**	**2.51**	**2.49**
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28	2.23	2.18	2.12	2.08	2.04	1.99	1.96	1.92	1.90	1.87	1.85	1.84
	8.10	**5.85**	**4.94**	**4.43**	**4.10**	**3.87**	**3.71**	**3.56**	**3.45**	**3.37**	**3.30**	**3.23**	**3.13**	**3.05**	**2.94**	**2.86**	**2.77**	**2.69**	**2.63**	**2.56**	**2.53**	**2.47**	**2.44**	**2.42**
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.20	2.15	2.09	2.05	2.00	1.96	1.93	1.89	1.87	1.84	1.82	1.81
	8.02	**5.78**	**4.87**	**4.37**	**4.04**	**3.81**	**3.65**	**3.51**	**3.40**	**3.31**	**3.24**	**3.17**	**3.07**	**2.99**	**2.88**	**2.80**	**2.72**	**2.63**	**2.58**	**2.51**	**2.47**	**2.42**	**2.38**	**2.36**
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.26	2.23	2.18	2.13	2.07	2.03	1.98	1.93	1.91	1.87	1.84	1.81	1.80	1.78
	7.94	**5.72**	**4.82**	**4.31**	**3.99**	**3.76**	**3.59**	**3.45**	**3.35**	**3.26**	**3.18**	**3.12**	**3.02**	**2.94**	**2.83**	**2.75**	**2.67**	**2.58**	**2.53**	**2.46**	**2.42**	**2.37**	**2.33**	**2.31**
23	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	2.32	2.28	2.24	2.20	2.14	2.10	2.04	2.00	1.96	1.91	1.88	1.84	1.82	1.79	1.77	1.76
	7.88	**5.66**	**4.76**	**4.26**	**3.94**	**3.71**	**3.54**	**3.41**	**3.30**	**3.21**	**3.14**	**3.07**	**2.97**	**2.89**	**2.78**	**2.70**	**2.62**	**2.53**	**2.48**	**2.41**	**2.37**	**2.32**	**2.28**	**2.26**
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18	2.13	2.09	2.02	1.98	1.94	1.89	1.86	1.82	1.80	1.76	1.74	1.73
	7.82	**5.61**	**4.72**	**4.22**	**3.90**	**3.67**	**3.50**	**3.36**	**3.25**	**3.17**	**3.09**	**3.03**	**2.93**	**2.85**	**2.74**	**2.66**	**2.58**	**2.49**	**2.44**	**2.36**	**2.33**	**2.27**	**2.23**	**2.21**
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24	2.20	2.16	2.11	2.06	2.00	1.96	1.92	1.87	1.84	1.80	1.77	1.74	1.72	1.71
	7.77	**5.57**	**4.68**	**4.18**	**3.86**	**3.63**	**3.46**	**3.32**	**3.21**	**3.13**	**3.05**	**2.99**	**2.89**	**2.81**	**2.70**	**2.62**	**2.54**	**2.45**	**2.40**	**2.32**	**2.29**	**2.23**	**2.19**	**2.17**
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.10	2.05	1.99	1.95	1.90	1.85	1.82	1.78	1.76	1.72	1.70	1.69
	7.72	**5.53**	**4.64**	**4.14**	**3.82**	**3.59**	**3.42**	**3.29**	**3.17**	**3.09**	**3.02**	**2.96**	**2.86**	**2.77**	**2.66**	**2.58**	**2.50**	**2.41**	**2.36**	**2.28**	**2.25**	**2.19**	**2.15**	**2.13**
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	2.20	2.16	2.13	2.08	2.03	1.97	1.93	1.88	1.84	1.80	1.76	1.74	1.71	1.68	1.67
	7.68	**5.49**	**4.60**	**4.11**	**3.79**	**3.56**	**3.39**	**3.26**	**3.14**	**3.06**	**2.98**	**2.93**	**2.83**	**2.74**	**2.63**	**2.55**	**2.47**	**2.38**	**2.33**	**2.25**	**2.21**	**2.16**	**2.12**	**2.10**
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15	2.12	2.06	2.02	1.96	1.91	1.87	1.81	1.78	1.75	1.72	1.69	1.67	1.65
	7.64	**5.45**	**4.57**	**4.07**	**3.76**	**3.53**	**3.36**	**3.23**	**3.11**	**3.03**	**2.95**	**2.90**	**2.80**	**2.71**	**2.60**	**2.52**	**2.44**	**2.35**	**2.30**	**2.22**	**2.18**	**2.13**	**2.09**	**2.06**
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.05	2.00	1.94	1.90	1.85	1.80	1.77	1.73	1.71	1.68	1.65	1.64
	7.60	**5.42**	**4.54**	**4.04**	**3.73**	**3.50**	**3.33**	**3.20**	**3.08**	**3.00**	**2.92**	**2.87**	**2.77**	**2.68**	**2.57**	**2.49**	**2.41**	**2.32**	**2.27**	**2.19**	**2.15**	**2.10**	**2.06**	**2.03**
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16	2.12	2.09	2.04	1.99	1.93	1.89	1.84	1.79	1.76	1.72	1.69	1.66	1.64	1.62
	7.56	**5.39**	**4.51**	**4.02**	**3.70**	**3.47**	**3.30**	**3.17**	**3.06**	**2.98**	**2.90**	**2.84**	**2.74**	**2.66**	**2.55**	**2.47**	**2.38**	**2.29**	**2.24**	**2.16**	**2.13**	**2.07**	**2.03**	**2.01**
32	4.15	3.30	2.90	2.67	2.51	2.40	2.32	2.25	2.19	2.14	2.10	2.07	2.02	1.97	1.91	1.86	1.82	1.76	1.74	1.69	1.67	1.64	1.61	1.59
	7.50	**5.34**	**4.46**	**3.97**	**3.66**	**3.42**	**3.25**	**3.12**	**3.01**	**2.94**	**2.86**	**2.80**	**2.70**	**2.62**	**2.51**	**2.42**	**2.34**	**2.25**	**2.20**	**2.12**	**2.08**	**2.02**	**1.98**	**1.96**
34	4.13	3.28	2.88	2.65	2.49	2.38	2.30	2.23	2.17	2.12	2.08	2.05	2.00	1.95	1.89	1.84	1.80	1.74	1.71	1.67	1.64	1.61	1.59	1.57
	7.44	**5.29**	**4.42**	**3.93**	**3.61**	**3.38**	**3.21**	**3.08**	**2.97**	**2.89**	**2.82**	**2.76**	**2.66**	**2.58**	**2.47**	**2.38**	**2.30**	**2.21**	**2.15**	**2.08**	**2.04**	**1.98**	**1.94**	**1.91**
36	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.10	2.06	2.03	1.98	1.93	1.87	1.82	1.78	1.72	1.69	1.65	1.62	1.59	1.56	1.55
	7.39	**5.25**	**4.38**	**3.89**	**3.58**	**3.35**	**3.18**	**3.04**	**2.94**	**2.86**	**2.78**	**2.72**	**2.62**	**2.54**	**2.43**	**2.35**	**2.26**	**2.17**	**2.12**	**2.04**	**2.00**	**1.94**	**1.90**	**1.87**

df (denominator)	df (numerator)																						
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
38	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02	1.96	1.92	1.85	1.80	1.76	1.71	1.67	1.63	1.60	1.57	1.54	1.53
	7.35	**5.21**	**4.34**	**3.86**	**3.54**	**3.32**	**3.15**	**3.02**	**2.91**	**2.82**	**2.75**	**2.69**	**2.59**	**2.51**	**2.40**	**2.32**	**2.22**	**2.14**	**2.08**	**2.00**	**1.97**	**1.90**	**1.86**	**1.84**
40	4.07	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.04	2.00	1.95	1.90	1.84	1.79	1.74	1.69	1.66	1.61	1.59	1.55	1.53	1.51
	7.31	**5.18**	**4.31**	**3.83**	**3.51**	**3.29**	**3.12**	**2.99**	**2.88**	**2.80**	**2.73**	**2.66**	**2.56**	**2.49**	**2.37**	**2.29**	**2.20**	**2.11**	**2.05**	**1.97**	**1.94**	**1.88**	**1.84**	**1.81**
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.02	1.99	1.94	1.89	1.82	1.78	1.73	1.68	1.64	1.60	1.57	1.54	1.51	1.49
	7.27	**5.15**	**4.29**	**3.80**	**3.49**	**3.26**	**3.10**	**2.96**	**2.86**	**2.77**	**2.70**	**2.64**	**2.54**	**2.46**	**2.35**	**2.26**	**2.17**	**2.08**	**2.02**	**1.94**	**1.91**	**1.85**	**1.80**	**1.78**
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98	1.92	1.88	1.81	1.76	1.72	1.66	1.63	1.58	1.56	1.52	1.50	1.48
	7.24	**5.12**	**4.26**	**3.78**	**3.46**	**3.24**	**3.07**	**2.94**	**2.84**	**2.75**	**2.68**	**2.62**	**2.52**	**2.44**	**2.32**	**2.24**	**2.15**	**2.06**	**2.00**	**1.92**	**1.88**	**1.82**	**1.78**	**1.75**
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.14	2.09	2.04	2.00	1.97	1.91	1.87	1.80	1.75	1.71	1.65	1.62	1.57	1.54	1.51	1.48	1.46
	7.21	**5.10**	**4.24**	**3.76**	**3.44**	**3.22**	**3.05**	**2.92**	**2.82**	**2.73**	**2.66**	**2.60**	**2.50**	**2.42**	**2.30**	**2.22**	**2.13**	**2.04**	**1.98**	**1.90**	**1.86**	**1.80**	**1.76**	**1.72**
48	4.04	3.19	2.80	2.56	2.41	2.30	2.21	2.14	2.08	2.03	1.99	1.96	1.90	1.86	1.79	1.74	1.70	1.64	1.61	1.56	1.53	1.50	1.47	1.45
	7.19	**5.08**	**4.22**	**3.74**	**3.42**	**3.20**	**3.04**	**2.90**	**2.80**	**2.71**	**2.64**	**2.58**	**2.48**	**2.40**	**2.28**	**2.20**	**2.11**	**2.02**	**1.96**	**1.88**	**1.84**	**1.78**	**1.73**	**1.70**
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.98	1.95	1.90	1.85	1.78	1.74	1.69	1.63	1.60	1.55	1.52	1.48	1.46	1.44
	7.17	**5.06**	**4.20**	**3.72**	**3.41**	**3.18**	**3.02**	**2.88**	**2.78**	**2.70**	**2.62**	**2.56**	**2.46**	**2.39**	**2.26**	**2.18**	**2.10**	**2.00**	**1.94**	**1.86**	**1.82**	**1.76**	**1.71**	**1.68**
60	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.86	1.81	1.75	1.70	1.65	1.59	1.56	1.50	1.48	1.44	1.41	1.39
	7.08	**4.98**	**4.13**	**3.65**	**3.34**	**3.12**	**2.95**	**2.82**	**2.72**	**2.63**	**2.56**	**2.50**	**2.40**	**2.32**	**2.20**	**2.12**	**2.03**	**1.93**	**1.87**	**1.79**	**1.74**	**1.68**	**1.63**	**1.60**
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.93	1.89	1.84	1.79	1.72	1.67	1.62	1.56	1.53	1.47	1.45	1.40	1.37	1.35
	7.01	**4.92**	**4.08**	**3.60**	**3.29**	**3.07**	**2.91**	**2.77**	**2.67**	**2.59**	**2.51**	**2.45**	**2.35**	**2.28**	**2.15**	**2.07**	**1.98**	**1.88**	**1.82**	**1.74**	**1.69**	**1.62**	**1.56**	**1.53**
80	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95	1.91	1.88	1.82	1.77	1.70	1.65	1.60	1.54	1.51	1.45	1.42	1.38	1.35	1.32
	6.96	**4.88**	**4.04**	**3.56**	**3.25**	**3.04**	**2.87**	**2.74**	**2.64**	**2.55**	**2.48**	**2.41**	**2.32**	**2.24**	**2.11**	**2.03**	**1.94**	**1.84**	**1.78**	**1.70**	**1.65**	**1.57**	**1.52**	**1.49**
100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.88	1.85	1.79	1.75	1.68	1.63	1.57	1.51	1.48	1.42	1.39	1.34	1.30	1.28
	6.90	**4.82**	**3.98**	**3.51**	**3.20**	**2.99**	**2.82**	**2.69**	**2.59**	**2.51**	**2.43**	**2.36**	**2.26**	**2.19**	**2.06**	**1.98**	**1.89**	**1.79**	**1.73**	**1.64**	**1.59**	**1.51**	**1.46**	**1.43**
125	3.92	3.07	2.68	2.44	2.29	2.17	2.08	2.01	1.95	1.90	1.86	1.83	1.77	1.72	1.65	1.60	1.55	1.49	1.45	1.39	1.36	1.31	1.27	1.25
	6.84	**4.78**	**3.94**	**3.47**	**3.17**	**2.95**	**2.79**	**2.65**	**2.56**	**2.47**	**2.40**	**2.33**	**2.23**	**2.15**	**2.03**	**1.94**	**1.85**	**1.75**	**1.68**	**1.59**	**1.54**	**1.46**	**1.40**	**1.37**
150	3.91	3.06	2.67	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.82	1.76	1.71	1.64	1.59	1.54	1.47	1.44	1.37	1.34	1.29	1.25	1.22
	6.81	**4.75**	**3.91**	**3.44**	**3.14**	**2.92**	**2.76**	**2.62**	**2.53**	**2.44**	**2.37**	**2.30**	**2.20**	**2.12**	**2.00**	**1.91**	**1.83**	**1.72**	**1.66**	**1.56**	**1.51**	**1.43**	**1.37**	**1.33**
200	3.89	3.04	2.65	2.41	2.26	2.14	2.05	1.98	1.92	1.87	1.83	1.80	1.74	1.69	1.62	1.57	1.52	1.45	1.42	1.35	1.32	1.26	1.22	1.19
	6.76	**4.71**	**3.88**	**3.41**	**3.11**	**2.90**	**2.73**	**2.60**	**2.50**	**2.41**	**2.34**	**2.28**	**2.17**	**2.09**	**1.97**	**1.88**	**1.79**	**1.69**	**1.62**	**1.53**	**1.48**	**1.39**	**1.33**	**1.28**
400	3.86	3.02	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.91	1.78	1.72	1.67	1.60	1.54	1.49	1.42	1.38	1.32	1.28	1.22	1.16	1.13
	6.70	**4.66**	**3.83**	**3.36**	**3.06**	**2.85**	**2.69**	**2.55**	**2.46**	**2.37**	**2.29**	**2.23**	**2.12**	**2.04**	**1.92**	**1.84**	**1.74**	**1.64**	**1.57**	**1.47**	**1.42**	**1.32**	**1.24**	**1.19**
1000	3.85	3.00	2.61	2.38	2.22	2.10	2.02	1.95	1.89	1.84	1.80	1.76	1.70	1.65	1.58	1.53	1.47	1.41	1.36	1.30	1.26	1.19	1.13	1.08
	6.66	**4.62**	**3.80**	**3.34**	**3.04**	**2.82**	**2.66**	**2.53**	**2.43**	**2.34**	**2.26**	**2.20**	**2.09**	**2.01**	**1.89**	**1.81**	**1.71**	**1.61**	**1.54**	**1.44**	**1.38**	**1.28**	**1.19**	**1.11**
∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79	1.75	1.69	1.64	1.57	1.52	1.46	1.40	1.35	1.28	1.24	1.17	1.11	1.00
	6.64	**4.60**	**3.78**	**3.32**	**3.02**	**2.80**	**2.64**	**2.51**	**2.41**	**2.32**	**2.24**	**2.18**	**2.07**	**1.99**	**1.87**	**1.79**	**1.69**	**1.59**	**1.52**	**1.41**	**1.36**	**1.25**	**1.15**	**1.00**

Appendix 3

ADF Critical Values

Augmented Dickey-Fuller (ADF) Test Left-Hand Critical Values (*t* test) and Right-Hand Critical Values (*F* Test): 5% Level of Significance

n	No Intercept, No Trend	Intercept, No Trend	Intercept, Trend	*F* Statistic
25	−2.26	−3.33	−3.95	7.24
50	−2.25	−3.22	−3.80	6.73
100	−2.24	−3.17	−3.73	6.49
250	−2.23	−3.14	−3.69	6.34
500	−2.23	−3.13	−3.68	6.30
∞	−2.23	−3.12	−3.66	6.25

Source: W. A. Fuller, *Introduction to Statistical Time Series*, Wiley, New York, 1976; D. A. Dickey and W. A. Fuller, "Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root," *Econometrica* **49** (1981), pp. 1057–1072.

Debre Market University

College of Business and Economics

Department of Economics

Econometrics II (Econ 2062) Assignment (30%)

1. Discuss in detail with clear examples what stochastic process (stationary, non-stationary) is?
2. What is the difference between Panel data and Pooled data?
3. Suppose a friend of you, who is a junior economist, wants to estimate "***Married Women's Annual Labor Supply in Debre Markos Town*****",** where the dependent variable is measured ***in hours worked,*** as a function of education level, experience, spouse's earning, and number of siblings. To do so, he asked you for an advice about the type of model he shall specify. As a student of econometrics class, which model would you recommend? Why?
4. What does it mean by identification of a simultaneous equation model?
5. What are Problems of Fixed Effects Approach, or LSDV model?
6. Suppose If $X_t \sim I(3)$ and $Y_t \sim I(3)$, then what do you comment about the possible linear combination between X_t and Y_t given by $Z_t = (aX_t + bY_t) = I(d^*)$?
7. Make distinction between ordered Logit and multinomial Logit?
8. Distinguish between
    i. Structural equations and reduced for equations?
    ii. Structural equation parameters and reduced for equation parameters?
    iii. Endogenous variables and exogenous variables?
    iv. Observed shifters and unobserved shifters?
9. Suppose $X_t = \pi_1 + \pi_2 X_{t-1} + e_t$

    where, X_t is a stochastic process, t is time, π_1 and π_2 are parameters and different from zero and e is the disturbance term

    A. Show X_t is non-stationary stochastic process.

    B. Show X_t would be stationary stochastic process if π_1 and π_2 were zero

10. The Dickey-Fuller unit-root test result of a stochastic process Z_t is given as follows.

$\Delta Z_t = -0.423 Z_{t-1}$

tau.....(−1.176)

$\Delta Z_t = -3.15 + 0.423 Z_{t-1}$

tau.....(−3.1)....(0.786)

$\Delta Z_t = 8.15 + 0.67t - 0.914 Z_{t-1}$

tau.....(3.1)...(0.786)...(−1.1)

where ΔZ_t is the estimated change in Z_t.

Assuming that the sample size is large and 5% significance level,

A. Based on the above result, is Z_t stationary or non-stationary?

B. Why?

11. Suppose estimation of a consumption model gives the result

$$\widehat{Y}_i = \underbrace{1200}_{(60.0)} + \underbrace{0.25}_{(0.024)} X_{1i} + \underbrace{100}_{(18.40)} D_{1i} - \underbrace{50}_{(7.42)} D_{2i}$$

Where, Y_i is consumption, X_{1i} is monthly disposable income, $D_{1i} = 1$if the person is Ethiopian and 0 otherwise, and $D_{2i} = 1$ if sex is male and 0 otherwise, $E(X_{1i}) = 800$.

i. What is the amount of average consumption of Ethiopian people?

ii. Determine the mean consumption of female people who are not Ethiopians?

12. Use both rank and order conditions to show whether the following SEM is identified?

$$\left.\begin{aligned} Y_1 &= \alpha_0 + \alpha_1 Y_2 + \alpha_3 Y_3 + \alpha_4 \mathbf{Z_1} + u \\ Y_2 &= \beta_0 + \beta_1 Y_1 + \beta_2 \mathbf{Z_1} + \beta_3 \mathbf{Z_2} + \beta_4 \mathbf{Z_3} + \beta_5 \mathbf{Z_4} + \varepsilon \\ Y_3 &= \gamma_0 + \gamma_1 Y_1 + \gamma_2 \mathbf{Z_1} + \gamma_3 \mathbf{Z_2} + \gamma_4 \mathbf{Z_3} + v \end{aligned}\right\}$$

Check whether the above SEM is identified using both order and rank conditions?