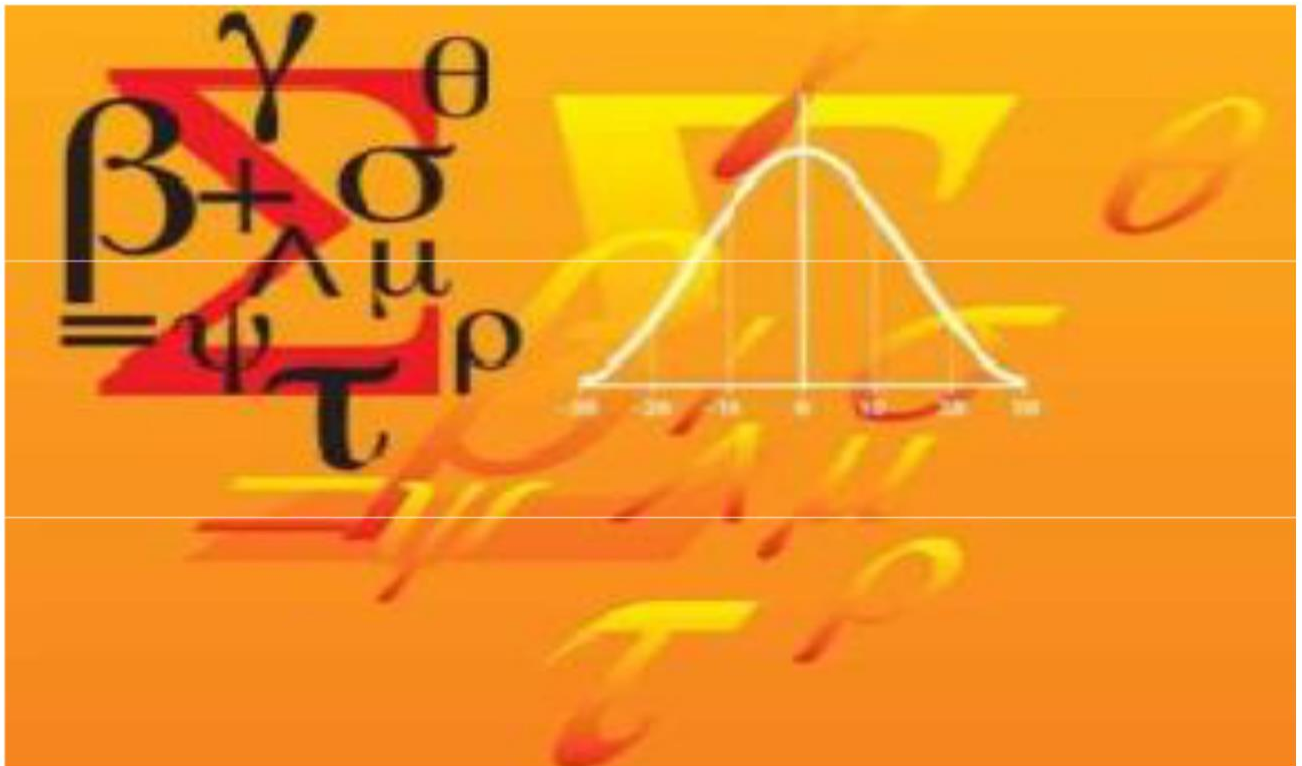




ARBAMINCH UNIVERSITY
COLLEGE OF BUSINESS & ECONOMICS
DEPARTMENT OF ECONOMICS

ECONOMETRICS II (ECON- 3022)



Compiled By: Mesfin Menza (PhD)

MAY, 2020

ARBAMINCH, ETHIOPIA

A TEACHING MATERIAL ON Econometrics II [ECON- 3022]
FOR DISTANCE STUDENTS



By: Mesfin Menza (PhD)

ARBA MINCH UNIVERSITY
COLLEGE OF CONTINUING AND DISTANCE EDUCATION
Tel: +251-046-881-49-72
+251-046-881-07-71
P.O.BOX: 21, ARBAMINCH UNIVERSITY
FAX: +251-6-81-02-79
+251-6-81-02-79

May, 2020

ARBA MINCH, ETHIOPIA

PREFACE

Literally, Econometrics is about adding some 'metrics' into Economics. In recent times, it has become common to see econometric applications in economic planning, research, and policies. Econometrics has also become an integral part of social, agricultural, health and other researches. An applied econometrics usually starts with an economic model, constructed either by observation or by other form of deduction and needs the important ingredients: Data, Methodology, knowledge & understanding of economic theory and Computer software. Particularly, the tremendous advancement in computer technology has made econometrics a handy tool of economists to explain the complex realities of the actual world. Hence, this module, which is a simple extension of econometrics I, is prepared with the objective of introducing students of economics with the basic knowledge of quantitative analysis and enabling them to develop strong background about the application of econometrics to policy analysis, decision making, and research. Concisely, the material is designed to enable students to know how to measure economic variables (both quantitative and qualitative) so as to tell the exact extent of the effect that some variables may have on some other variables.

This material assumes that readers have sufficient background of Business statistics, Calculus for Economists, Linear algebra for Economists, Research methods for Economists, Statistics for Economists and Econometrics I. As the module is mainly meant for students of economics, it requires solid knowledge of the basic principles and theories of economics as a pre-requisite.

The module is organized in such a way that the first chapter introduces students to the new era called violations of the assumptions of classical linear regression models. The second chapter deals with the regression analysis on dummy variables, while the third chapter pioneer dynamic econometric models or time series econometrics in brief. Finally, chapter four introduces simultaneous equation models.

Dear learners and other interested readers! I hope that you will find this material worth reading. And perhaps, if any editorial or technical errors are committed, they are all mine.

Mesfin Menza (PhD)

CHAPTER ONE

VIOLATIONS OF BASIC ASSUMPTIONS OF CLASSICAL LINEAR REGRESSION MODELS

Overview

Dear Distance Learners, on the chapters of Classical Linear Regression Models of Econometrics I (both simple and multiple), you are introduced with the basic concepts, **assumptions**, estimation methods, and interpretations of the classical linear regression models and their applications in economic science. The main objectives in fitting a regression equation in both cases are (1) to estimate the regression coefficients and related parameters and (2) to predict the value of the dependent variable in terms of that of the independent variable (or variables).

The ordinary least square (OLS) estimation method discussed in Econometrics I is based on some basic assumptions for its applicability. Among others, in both the simple and multiple regression models, we assumed that ' u_t ' was random variable with mean zero and constant variance ($\text{var}(u_t) = \sigma^2$), and that the errors corresponding to different observations are uncorrelated ($\text{cov}(u_i, u_j) = 0$ (for $i \neq j$)) and in multiple regression we assumed that there is no perfect correlation between the independent variables. But in many real world instances, in applied regression analysis a number of problems may occur and the assumptions of classical linear regression models may be violated. Thus, in this chapter we will try to relax some of these assumptions and see the effect of these relaxations on OLS estimates, and inferences and predictions made using the OLS estimates. Specifically, in this chapter we will try to relax three of the basic assumptions, namely the assumptions of homoscedastic variance of the error terms, non auto-correlated error terms, and no perfect multicollinearity among the explanatory variables.

Consequently, in the first part of this chapter we will try to analyze the causes, consequences, detection mechanisms and the remedial measures for heteroscedasticity of the variance of the error term. In the subsequent sections of the chapter, we will treat issues related to autocorrelation of the error term and multicollinearity among the explanatory variables.

Chapter Contents

Dear distance learners, this chapter contain the following major sub-topics:

General Overview

Heteroscedasticity

- Introduction
- Causes for Heteroscedasticity
- Consequences of Heteroscedasticity
- Detecting Heteroscedasticity
- Solutions for

Heteroscedasticity Autocorrelation

- Introduction
- Causes for Autocorrelation
- Consequences of Autocorrelation
- Detecting Autocorrelation
- Solutions for the problem of

Autocorrelation Multicollinearity

- Introduction
- Causes for Multicollinearity
- Consequences of Multicollinearity
- Detecting Multicollinearity
- Solutions for the incidence of Multicollinearity

At the end of this chapter students will be able to:

- ❖ explain the causes and consequences of the heteroscedasticity of the variance of the error term
- ❖ know how to detect whether the variance of the error term is heteroscedastic or homoscedastic
- ❖ take remedial measures to treat the consequential problem of heteroscedastic.
- ❖ describe the causes and consequences of auto-correlated error terms;
- ❖ differentiate between different types of autocorrelation schemes
- ❖ use different mechanisms to detect whether the error terms are autocorrelated
- ❖ take remedial measures to cure the problem of autocorrelation, if any
- ❖ elaborate the causes and consequences of perfect and near to perfect multicollinearity among the explanatory variables
- ❖ detect multicollinearity using different tests
- ❖ suggest possible remedial measures when multicollinearity become severe

1.1. Heteroscedasticity

1.1.1. Introduction

Dear students, as you might remember from Econometrics I, the *third assumption* classical linear regression model required for the OLS estimator to be efficient states that the variance of the error term has to be constant and the same for all observations. This is referred to as a **homoscedastic error term**. In other words, the probability distribution of the disturbance term remains same over all observations of X; i.e. the variance of each u_i is the same for all the values of the explanatory variable. Symbolically,

$\text{var}(u_i) = \sigma^2$; constant value. Which is equivalent to the variance of error term does not depend on i .

When this assumption is violated and the variance is different for different observations we refer to this as heteroskedasticity. Thus, we say that u 's are heteroscedastic when:

$\text{var}(u_i) = \sigma_i^2$ (a value that varies). That means the variance of the error term depends on i .

In words, it arises when the variances of the error terms of a regression model are not constant over different sample observations.

NB: In the case of homoskedastic errors the error terms are stochastic while in the case of heteroskedastic errors there is a systematic pattern in the errors.

Heteroskedasticity can arise *empirically, through theoretical considerations and from model misspecification*. Empirically, heteroskedasticity is often encountered when using **cross-sectional data** on a number of microeconomic units such as firms or households than in time-series data. One deals with members of population at a given point of time, such as individual consumers or their families, firms, industries. These members may be of different size such as small, medium or large firms or low, medium or high income. In time series data on the other hand, the variables tend to be of similar orders of magnitude because one generally collects data for the same entity over a period of time.

Heteroscedasticity poses a problem when we are estimating parameters in the regression model, because the least-squares estimation procedure places more weight on observations that have large errors and variances. Thus the regression line is adjusted to give a good fit for the large-variance portion of the observations but largely ignores the small-variance part of the data. The result is that the variances of the estimates do not have a minimum variance.

Therefore, this sub-section will discuss types of heteroscedasticity, some heteroscedastic functions, the causes and consequences of violating the homoskedasticity assumption, how to detect any deviations from the assumption and how to solve the problem when present in detail.

1.1.2. Types of Heteroscedasticity

The case of Heteroscedasticity is shown by increasing or decreasing dispersion of the observations from the regression line. The pattern of the observations on a scatter diagram depends on the form of Heteroscedasticity, i.e., on the form of the relationship between δ_{ui}^2 and X_i .

Generally, we can encounter three types of Heteroscedasticity:

- i. **Increasing Heteroscedasticity:** This is the case of (monotonically) increasing variance of the stochastic term, u , i.e., as X increases, so does the variance of u . This is the common form of heteroscedasticity assumed in econometric applications.
- ii. **Decreasing Heteroscedasticity:** As X assumes higher values the deviation of the observations from the regression line decreases, i.e., the variance of the random variable changes in the opposite direction to the explanatory variable.
- iii. **Cyclical Heteroscedasticity:** The variance of u decreases initially as X assumes higher values, but after a certain level of X , the variance of u increases with X .

Graphical representation of heteroscedasticity and homoscedasticity

The assumption of homoscedasticity states that the variation of each u_i around its zero mean does not depend on the value of explanatory variable. The variance of each u_i remains the same irrespective of small or large values of X ; the explanatory variable. Mathematically, σ_u^2 is not a function of X ; i.e. $\sigma_u^2 \neq f(X_i)$.

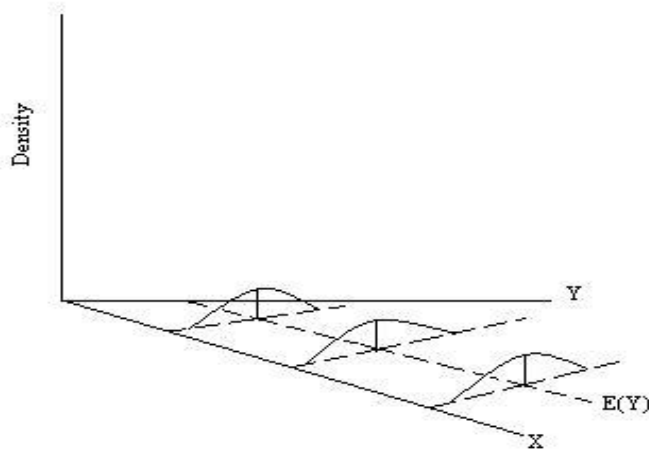


Figure (a) Homoscedastic error variance

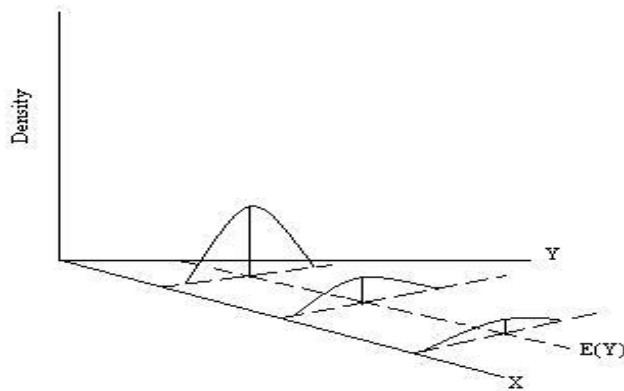
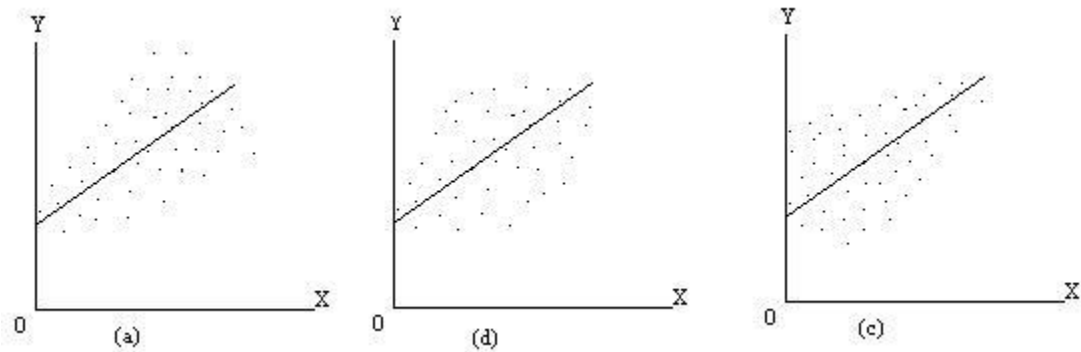


Figure (b) Heteroscedastic error variance

If the variance of U were the same at every point or for all values of X , definite restriction would be placed on the scatter of Y against X when plotted in the three dimensions, we should observe something approximating the pattern of figure (a). In contrast, consider figure (b) shows that the conditional variance of Y_i (which in fact is u_i) increases as X increases.

If σ_u^2 is not constant but its value depends on the value of X; it means that $\sigma_u^2 = f(X_i)$. Such dependency is depicted diagrammatically in the following figures. Three cases of heteroscedasticity: all shown by increasing or decreasing dispersion of the observation from the regression line.



In panel (a) σ_u^2 seems to increase with X while in panel (b) the error variance appears greater in X's middle range, tapering off toward the extremes. Finally, in panel (c), the variance of the error term is greater for low values of X, declining and leveling off rapidly as X increases.

Generally, it should be clear that the pattern of heteroscedasticity depends on the signs and values of the coefficients of the relationship between σ_{ui}^2 and X_i .

$$\Rightarrow \sigma_{ui}^2 = f(X_i).$$

However, since the u 's are not observable, we do not know the true pattern of heteroscedasticity. In applied research econometricians usually make the assumption that heteroscedasticity, if exists, will take the following forms:

- i. $\sigma_{ui}^2 = K_i^2 (X_i^2)$
- ii. $\sigma_{ui}^2 = K^2 (X_i)$
- iii. $\sigma_{ui}^2 = \frac{K}{\Lambda_i}$ etc.

Where, K_i 's are constants to be estimated from the model.

1.1.3. Plausibility of the Assumption of Homoscedasticity.

Dear students, do you think that the variance of the stochastic term is constant in real world applications? -----

In many econometric applications the assumption of constant variance of the random variable may well be expected not to hold. Why?

1. As u expresses the influence of measurement errors on the values of the dependent variable, there is a cogent reason for expecting the variance of u to vary over time, in most cases. For example as Y increases, errors of measurement tend to increase, because it becomes more difficult to collect data, and check its consistency and reliability. Furthermore, the errors of measurement tend to be cumulative over time, so that their size tends to increase. In this case the variance of u increases with increasing values of X .
2. The sampling techniques and data collection methods may continuously improve over time, and thus errors of measurement may decrease, in which case δ_{ui}^2 decreases over time.
3. Many of the variables omitted from the model tend to change in the same direction with X . Thus, they cause an increase in the variation of the observations of the dependent variable from the regression line, which implies an increase in the variance of u as the values of X increase.
4. On a priori grounds there are reasons to believe that the assumption of Heteroscedasticity may often be violated in practice.

Some Heteroscedastic functions in practice

As we have mentioned, heteroscedasticity is the major problem of cross sectional data. Let us see some of the cross sections where heteroscedasticity is the severe problem. **Size of the firm:** Heteroscedasticity would be a problem in a study of sales for a cross section of firms in an industry, because the error terms for large firms would be likely to have larger variances than those for small firms. In other words, the high volatility in sales for larger firms might pose problems for the researcher.

The magnitude of error terms is a function of firm size. For example, if we have three firms i , j , and k , it may be the case that $\sigma_i^2 > \sigma_j^2 > \sigma_k^2$.

Production Function: Suppose we are required to estimate the production function $X = f(K, L)$ of the sugar industry from a cross-section random sample of firms of the industry. Disturbance terms in the production function would stand for many factors; like entrepreneurship, technological differences, selling and purchasing procedures, differences in organizations, etc. other than inputs, labor (L) and capital (K) considered in the production function. The factors mentioned above, which are not considered explicitly in the production function show considerable variance in large firms than in small ones. This leads to breakdown of our assumption on homogeneity of variance terms.

Another commonly cited example of heteroscedasticity is the relationship between family expenditures and income. High-income families are likely to exhibit a higher variance in spending than lower-income families. Thus, the expenditures for consumers with lower income have smaller error terms.

Consumption Function: Suppose we are to study the consumption expenditure from a given cross-section sample of family budgets: $C_i = \alpha + \beta Y_i + U_i$

Where: C_i = Consumption expenditure of i^{th} household; Y_i = Disposable income of i^{th} household.

At low levels of income, the average consumption is low, and the variation below this level (autonomous consumption, α) is less possible; consumption cannot fall too far below because this might mean starvation. On the other hand, it cannot rise too far above because money income does not permit it or constrains us. Such constraints may not be found at higher income levels. Thus, consumption patterns are more regular at lower income levels than at higher levels. This implies that at high incomes the u 's will be high, while at low incomes the u 's will be small. The assumption of constant variance of u 's is therefore, does not hold when estimating the consumption function from across section of family budgets.

Saving function: in estimating the savings function from a cross-section of family budget the assumption of constant variance of the u 's is not appropriate because high income families show a much greater variability in their saving behavior than do low income families. High income families tend to stick to a certain standard of living and when their income falls they cut down their savings rather than cutting down their consumption expenditure. On the other hand, low income families save for certain purposes, and thus their saving patterns are more regular. This implies that at high incomes the u 's will be high, while at low incomes the u 's will be small.

1.1.4. Causes for Heteroscedasticity

There are several reasons why the variances of u_i may be variable. Some of these are:

1. **Error learning model:** it states that as people learn their error of behavior become smaller over time. In this case σ_i^2 is expected to decrease.

Example: as the number of hours of typing practice increases, the average number of typing errors and as well as their variance decreases.

2. **Poor data collection technique:** As data collection technique improves, σ_{ui}^2 is likely to decrease. Thus banks that have sophisticated data processing equipment are likely to commit fewer errors in the monthly or quarterly statements of their customers than banks without such facilities.

3. Heteroscedasticity can also arise as a result of the presence of **outliers**. An outlier is an observation that is much different (either very small or very large) in relation to the other observation in the sample. It is an observation from a different population to that generating the remaining sample observations.

4. **Specification bias (omitted variable):** This is also called functional misspecification. In such a case the residuals obtained from the regression may give the distinct impression that the error variances may not be constant.

5. **Skewness:** the distribution of some variables such as income, wealth, etc... is skewed. It is well known that the distribution of income and wealth in most societies is uneven, with the bulk of the income and wealth being owned by a few at the top.

6. **Incorrect data transformation**, etc.

1.1.5. Consequences of Heteroscedasticity

The classical assumptions made on the error terms are that they are uncorrelated,

with mean zero and constant variance

². In technical terms this means that

$$E[U_i] = 0 \quad (1.1.1)$$

$$\text{Var}[U_i] = \sigma^2 \quad (1.1.2)$$

$$\text{Cov}[U_i, U_j] = 0 \quad (1.1.3)$$

Assumptions (1.1.1) and (1.1.3) are in use to make the OLS estimators unbiased and consistent. Assumption (1.1.2) is important for the OLS estimator to be **efficient**. Hence, if (1.1.2) is ignored (become heteroscedastic), we can no longer claim that our estimator is the best estimator among linear unbiased estimators. This means that it is possible to find another linear unbiased estimator that is more efficient.

So what happens when we use ordinary least squares procedure to a model with heteroscedastic disturbance terms?

In the presence of heteroskedasticity:

1. The OLS estimators of the population parameters are still linear, consistent and unbiased.

The fact that the parameter estimators are unbiased can be seen as follows:

$$\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} = \frac{\sum_{i=1}^n x_i (x_i \beta + u_i)}{\sum_{i=1}^n x_i^2} = \beta + \frac{\sum_{i=1}^n x_i u_i}{\sum_{i=1}^n x_i^2}$$

Taking the expected value of both sides:

$$E\left[\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}\right] = \beta + \frac{\sum_{i=1}^n x_i E[u_i]}{\sum_{i=1}^n x_i^2}, \text{ given } E[u_i] = 0 \quad (1.1.4)$$

This means as we increase our observation, negative deviations will tend to be offset by positive deviations from the regression line.

Similarly; $\hat{Y} = \frac{\sum_{i=1}^n Y_i}{n} = \frac{\sum_{i=1}^n (X_i \beta + u_i)}{n} = \beta + \frac{\sum_{i=1}^n u_i}{n}$

Still taking the expected value of both sides and given $E[u_i] = 0$:

$$E(\hat{Y}) = \beta + \frac{\sum_{i=1}^n E[u_i]}{n} = \beta \quad (1.1.5)$$

2. OLS estimators are inefficient (do not have minimum variance). This means they are no longer BLUE. Thus, the usual standard errors of the estimated parameters are biased and inconsistent as the true standard error of β_1 shall be understated. As such the t-value associated with it will be overestimated that leads to the conclusion of is statistically significance, which in fact may not be true.

It is important to understand that the violation of (1.2) makes the standard errors of the OLS estimators and the covariance's among them biased and inconsistent. Therefore tests of hypothesis are no longer valid, since the standard errors are wrong.

To see this, consider the variance of the estimator for the slope coefficient of the simple regression model:

$$V(\hat{\beta}_1) = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 V[Y_i]}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 \sigma_i^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (1.1.6)$$

The expression given by (1.1.6) represents the correct variance that should be used. Unfortunately it involves the unknown population variance of the error term which is different for different observations.

Since the error term is heteroskedastic, each observation will have a different error variance. The expression will therefore deviate from the variance estimated under homoskedasticity, that is:

$$V(\hat{\beta}_1) = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 \sigma_i^2}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 E[\sigma_i^2]}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (1.1.7)$$

As can be seen in (1.1.7) the variance of the OLS estimator is different from the expected value of the sample variance of the estimator that works under the assumption of a constant variance.

3. The usual formulae of the variances of the coefficients are not appropriate to conduct tests of significance and construct confidence intervals. The tests are inapplicable.

When heteroskedasticity is present, ordinary least squares estimation places more weight on observations with large variances than those with smaller variances.

The weighting occurs because the sum of squared residuals associated with large error variances are likely to be larger the sum of squared residuals associated with low error variances.

In computing the variances of the OLS parameters, σ^2 was assumed to be constant.

$$Var \frac{\sum_{i=1}^n \frac{e_i^2}{X_i^2}}{\sum_{i=1}^n X_i^2} \quad (1.1.8)$$

In the presence of heteroskedasticity, however, the error variance is assumed to vary from one observation to another and to be related to independent variable(s).

$$Var \frac{\sum_{i=1}^n \frac{e_i^2}{X_i^2}}{\sum_{i=1}^n X_i^2} \quad (1.1.9)$$

Thus, in the presence of heteroskedasticity the usual hypothesis tests and confidence intervals based on the various test statistics do not hold and become unreliable and leading to wrong conclusions.

4. An important use of the regression equation is that of making predictions and forecasts of the future. Since the OLS estimators are unbiased and consistent, so will the forecasts. However, since the estimators are inefficient, the uncertainty of the forecasts will increase, and the confidence interval of the forecast will be biased and inconsistent; thus, our inference and prediction about the population coefficients would be incorrect.

1.1.6. Detecting Heteroscedasticity

Dear Distance Learners, from the discussion we had above we have observed that the consequences of heteroscedasticity are serious on OLS estimates. If there is heteroscedasticity in the model, estimation, statistical inferences, predictions etc will be fruitless. Hence, before proceeding with any statistical or econometric analysis, we have to figure out whether there is the problem of heteroscedasticity. Fortunately there are a number of detection mechanisms, graphical as well as statistical that one can apply in order to receive an answer to the question. The most commonly used methods to detect the possible existence of heteroscedasticity are discussed as follows. Basically there are two methods of detecting heteroscedasticity. These are:

1. **Visual inspection of the residuals methods, and**
2. **Statistical Methods**

1.1.6.1. Visual inspection methods

These methods are called informal because it does not undertake the formal testing procedures such as t-test, F-test and the like. It is a test based on the nature of the **graph** and **covariance-variance matrix notation**.

A. Graphical methods

A natural starting point in detecting possible deviations from homoskedasticity is to plot the data. Since we are interested in the behavior of the error term and its variation, we may have different scatter plots that come from a linear regression model.

In graphical method to check whether a given data exhibits heteroscedasticity or not we look on whether there is a systematic relation between residual squared e_i^2 and X_i , the

dependent variable Y or with the fitted value of $(\hat{Y})$. The easiest way to check for heteroscedasticity is to look at a plot of the residuals against the independent variables or the expected/fitted values.

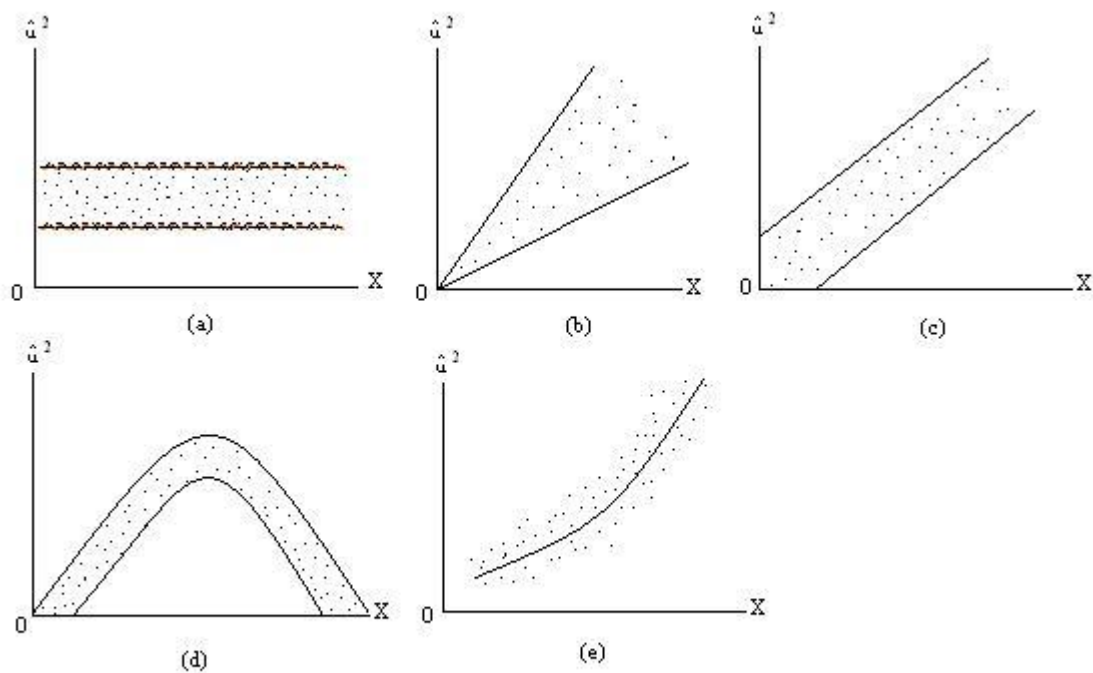
In the simple regression case with just one explanatory variable, the graphs in these two cases are always very similar. However, when using a multiple regression models the picture might be different, since the residual is a linear combination of all variables included in the model.

Since it is the partial effect of X on the residual that is of primary interest, it is advised that the residual should be plotted against all involved variables separately.

If it is possible to find a systematic pattern that give indications of differences of the variances over the observations, one should be concerned.

Thus, in the first case, the estimated residuals could also be plotted directly against X , as is done in Figures a to d.

In fig (a), we see there is no systematic pattern between the two variables, suggesting that perhaps no heteroscedasticity is present in the data. Figures b to e, however, exhibit definite patterns. . For instance, c suggests a linear relationship where as d and e indicate quadratic relationship between e_i^2 and Y_i .



Scattergram of estimated squared residuals against X

The second and the most commonly used graphical method is to plot the *residuals versus the dependent variable or fitted (predicted) values*.

To estimate the error term, we first compute the predicted dependent value by the regression model:

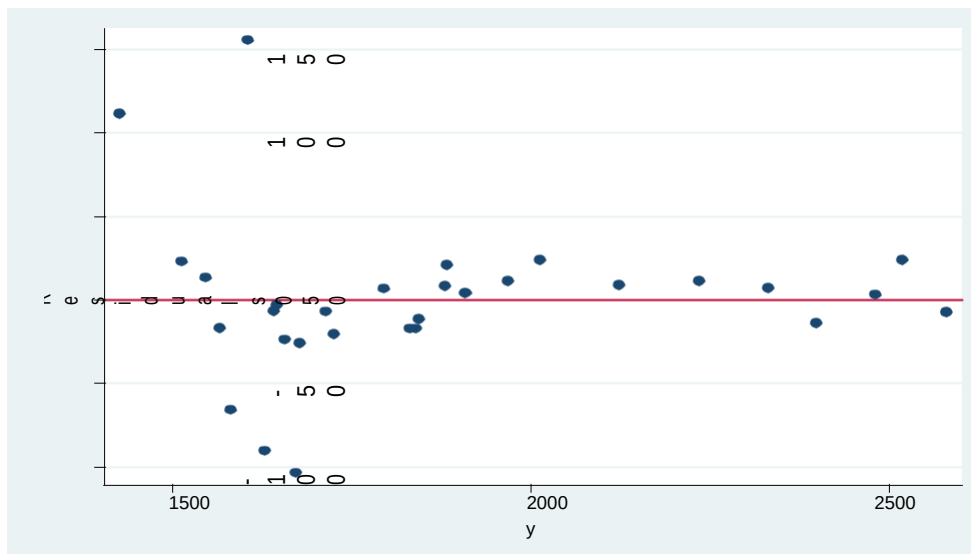
$$\hat{y}_i = a + b_1 x_{1i} + b_2 x_{2i} + \dots + b_k x_{ki} \quad (1.1.10)$$

Then we calculate the error term by taking the actual value for Y_i and subtracting the predicted value: $e_i = y_i - \hat{y}_i$.

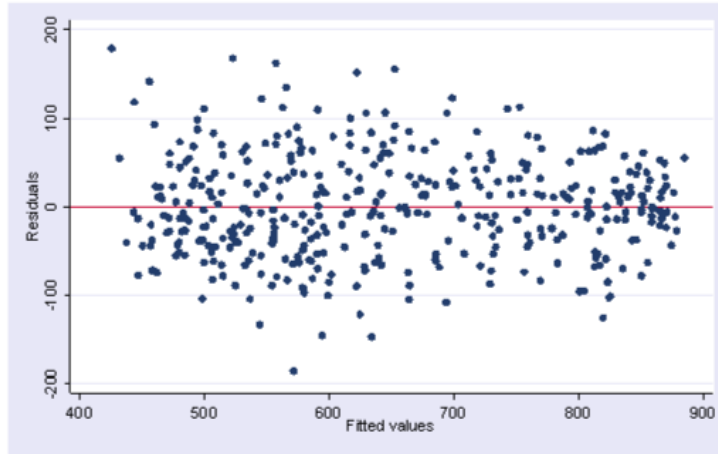
In practice, we generally plot the residuals range e , against the predicted values, $(\hat{Y})$. If the residuals appear to be random and the width of the scatter diagram seems constant throughout the data—that is, if no pattern is apparent—then no heteroscedasticity is present.

We do this by issuing the **rvpplot y** or **rvfplot** command in STATA after regression. Below we use the **rvpplot y** and **rvfplot** command with the **ylines(0)** option to put a reference line at $y=0$ and can produce the following graphs. In both cases, we see that the pattern of the data points is getting a little narrower towards the right end, which is an indication of heteroscedasticity.

rvpplot y



rvfplot, ylines(0)



Graphical methods are useful, but sometimes it is difficult to say if heteroskedasticity is present and found harmful. *It is therefore necessary to use statistical test.* The graphical method is therefore merely a first step in the analysis that can give a good picture of the nature of the heteroscedasticity we might have, which will be helpful later on when we are correcting for it.

B. Matrix Notation of Heteroscedasticity

The variance-covariance of the error terms can be represented in a matrix form with n columns and n rows.

The diagonal terms of the matrix represent the variances of the individual disturbances and the off-diagonal terms represent the covariance between them. If all the diagonal terms are the same, the disturbances are said to have uniform variance (homoskedastic) and if they are not the same they are said to be heteroskedastic. This means the disturbance term is thought of as being drawn from a different distribution for each observation.

If the disturbances are homoskedastic, their variance-covariance matrix can be written as:

$$E^2 I \quad (1.10)$$

Where, I is an identity matrix of $n \times n$ (with 1's along the diagonal and zero's on the off-diagonal sides).

The variance covariance matrix is represented as:

$$\begin{array}{ccccccc}
 (U_i^2) & (UU) & \dots & (UU) \\
 (UU)_1 & (U^2) & \dots & (U^2) \\
 \vdots & \vdots & & \vdots \\
 (UU)_n & (U^2) & \dots & (U^2)
 \end{array}
 \quad (1.1.11)$$

If (U_i^2) is a constant value, then $(UU)' = I_n$ given that $(U_i U_j) = 0$.

If (U_i^2) is a value that varies

$$(UU)' = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 2 & \dots & 0 \\ \vdots & \vdots & & \vdots \end{pmatrix} \quad (1.1.12)$$

Where (U_i^2) is a value that varies. In other words, variance covariance matrix in the present case is a diagonal matrix with unequal elements in the diagonal.

1.1.6.2. Statistical Tests

These are called the formal methods. There are several formal methods of testing heteroscedasticity which are based on the formal testing procedures. The most commonly used tests to detect the possible existence of heteroscedasticity are discussed as follows.

A. The Spearman rank-correlation test

This is the simplest test, which may be applied to either small or large samples. This test assumes that the variance of the disturbance term is either increasing or decreasing as x increases.

The steps to conduct this test are outlined as follows.

Step 1: Formulate the null hypothesis of homoscedasticity against the alternative hypothesis of heteroscedasticity as:

$$H_0 : \rho^s = 0$$

$$H_1 : \rho^s \neq 0$$

Where, ρ^s is the population rank correlation coefficient and its sample counterpart is

r^s . **Step 2:** Regress Y on X as shown below

$$Y_i = \beta_0 + \beta_1 X_i + U_i \quad (1.1.13) \text{ and we obtain the residuals, } e_i \text{ which are estimates of the } u_i \text{'s.}$$

Step 3: We order the e 's (ignoring their sign) and the X values in ascending or descending order and we compute the Spearman rank correlation coefficient, r^S using the following

formula: (1.1.14)

$$\frac{\sum D_i^2}{n(n^2 - 1)}$$

where D_i = difference between the ranks of corresponding pairs of X and e , and n = observations in the sample.

A high rank correlation coefficient is an implication of the presence of heteroscedasticity.

The student might wonder why the rank correlation coefficient is chosen and not the simple correlation coefficient r_{ex} . The answer is that $r_{ex} = 0$ always if OLS has been

used, since $\sum e_i = 0$ and $\sum X_i = 0$ can be

used as a measure of heteroscedasticity. This correlation coefficient is the basis of Glejser's test.

Step 4: Compute the value of t ($t^S_{calculated}$) from equation 1.1.15.

Assuming that the null hypothesis is true (i.e., population rank correlation coefficient, ρ^S is zero) and $n > 8$, the significance of sample rank correlation coefficient, r^S , can be tested by using t -test. The formula used to obtain the value of t for this purpose is given as:

$$t^S_{calculated} = \frac{r^S}{\sqrt{\frac{1-r^S^2}{n-2}}} \quad (1.1.15)$$

Step 5: Make decision by using the following rule

Decision Rule

Reject the null hypothesis of homoscedasticity, if the value of t obtained from equation 1.15 ($t^S_{calculated}$) is greater than the value of t obtained from t -tables for degree of freedom of $n - 2$, and conclude that there is the problem of heteroscedasticity.

NB: If we have a relationship with more explanatory variables we may compute the rank correlation coefficient between e_i and each one of the explanatory variables separately.

NB: As alternative, we can directly compare the calculated value of Spearman rank correlation coefficient with standard error, $r^S = \frac{r^S}{SE(r^S)}$ and if former is greater than the latter, we are able to reject the null hypothesis of homoscedasticity and conclude that there is heteroscedasticity problem in that particular data.

Example 1.1

The Capital Market Line (CML) Portfolio theory postulates a linear relationship between expected return (E_i) and risk (as measured by the standard deviation, σ) of a portfolio as: $E_i = \beta_0 + \beta_1\sigma_i + u_i$, where: E_i (average annual return, %) and σ_i (standard deviation of annual return). Using the data related to 10 mutual funds with different size and investment goals in table below, check heteroscedasticity using Spearman rank correlation coefficient test.

Mutual funds	E_i	σ_i
1	12.4	12.1
2	14.4	21.4
3	14.6	18.7
4	16	21.7
5	11.82	12.5
6	10.0	10.4
7	16.2	20.8
8	10.4	10.2
9	13.1	16.0
10	11.3	12.0

Solution

By applying the steps from 1-3, we can produce the following table results

Mutual funds	E_i	σ_i		$ e_i - \bar{e} $	Rank of e_i	Rank of σ_i	D_i	D_i^2
1	12.4	12.1	11.37	1.03	9	4	5	25
2	14.4	21.4	15.64	1.24	10	9	1	1
3	14.6	18.7	14.40	0.20	4	7	-3	9
4	16	21.7	15.78	0.22	5	10	-5	25
5	11.82	12.5	11.56	0.26	6	5	1	1
6	10.0	10.4	10.59	0.59	7	2	5	25
7	16.2	20.8	15.37	0.83	8	8	0	0
8	10.4	10.2	10.50	0.10	3	1	2	4
9	13.1	16.0	13.16	0.06	2	6	-4	16
10	11.3	12.0	11.33	0.03	1	3	-2	4
								$\Sigma = 110$

Thus, the Spearman rank correlation coefficient becomes:

$$r_s = \frac{\Sigma D_i^2}{n(n-1)}$$

$$= 1 - \frac{110}{10(10-1)}$$

$$= 0.33$$

Using the correlation coefficient, we can compute the t-calculated value as follows:

$$t_{s \text{ calculated}} = \frac{r_s \sqrt{n-2}}{\sqrt{1-r_s^2}}$$

$$= \frac{0.33 \sqrt{8}}{\sqrt{1-(0.33)^2}}$$

$$= 0.9998$$

For 8 degree of freedom the t- table value (at 10% = 1.39, at 5% = 1.86, and at 1% = 2.896). Since the t-calculated value is less than t-tabulated even at 10%, we accept the null hypothesis and reject the alternative hypothesis.

Conclusion is that there is no evidence of a systematic relationship between the explanatory variable (standard deviation of annual return) and the absolute values of the residuals, which might suggest that there is no heteroscedasticity problem in the data given.

B. Park test

Park formalizes the graphical method by suggesting that the variance of random disturbance σ_i^2 is some function of the explanatory variable X_i . The functional form he suggested was: $\sigma_i^2 = X_i e^{v_i}$

Or taking natural logarithm and since $\ln e = 1$,

$$\ln \sigma_i^2 = \ln X_i + v_i \quad (1.1.16)$$

where v is the stochastic disturbance term.

Since σ_i^2 is generally not known, park suggests using e_i^2 as proxy and running the following regression.

$$\ln e_i^2 = \ln X_i + v_i$$

$$\text{Since } \sigma^2 \text{ is constant, } \ln \sigma_i^2 = \ln X_i + v_i \quad (1.1.17)$$

Equation (1.1.17) indicates on how to test heteroscedasticity by testing on whether there is a significant relation between X_i and e_i^2 .

The null hypothesis: $H_0: \rho = 0$ is tested against the alternative hypothesis:

$$H_1: \rho \neq 0$$

If turns out to be statistically significant, it would suggest that heteroscedasticity is present in the data. If it turns out to be insignificant, we may accept the assumption of homoscedasticity. The park test is thus a two-stage test procedure; in the first stage, we run OLS regression disregarding the heteroscedasticity question. We obtain e_i from this regression and then in the second stage we run the regression in equation (1.1.17) above.

Example 1.2: Suppose that from a sample of size $n=100$ we estimate the relation between compensation and productivity.

$$\begin{aligned} Y &= 1992.342 + 0.2329X_i + e_i \\ SE &= (936.479) (0.0098) \\ t &= (2.1275) (2.333) \quad R^2 = 0.4375 \end{aligned}$$

The results reveal that the estimated slope coefficient is significant at 5% level of significant on the bases of one tail t-test since t-calculated is greater than t-tabulated.

The equation shows that as labour productivity increases by, say, a birr, labor compensation on the average increases by about 23 cents.

The residual obtained from the above regression were regressed on X_i as suggested by equation (1.14) giving the following result.

$$\begin{aligned} \ln e_i^2 &= 35.817 + 2.8099 \ln X_i + v_i \\ SE &(38.319) \quad (4.216) \\ t &(0.934) \quad (0.667) \quad R^2 = 0.0595 \end{aligned}$$

The above result revealed that the slope coefficient is statistically insignificant implying there is no statistically significant relationship between the two variables. Following the park test, one may conclude that there is no heteroscedasticity in the error variance. Although empirically appealing, the park test has some problems.

Gold Feld and Quandt have argued that the error term v_i entering into $\ln e_i^2 = \ln X_i + v_i$ may not satisfy the OLS assumptions and may itself be heteroscedastic. Nonetheless, as a strict explanatory method, one may use the park test.

C. Glejser test

The Glejser test is similar in spirit to the park test, but uses the absolute value of the estimated residuals as independent variable.

To detect the possible existence of heteroscedasticity using the Glejser test, we commonly use the following steps.

1. We perform the regression or estimation of Y on all the explanatory variables and we compute the residuals, e 's.

2. We regress the absolute values of e 's ($|e_i|$) on the explanatory variable with which

is thought, on *a priori* grounds, to be associated.'

The actual form of this regression is usually not known, so that one may experiment with various formulations, containing various powers of X , for example .

$$\begin{aligned} 1. |e_i| &= \alpha_0 + \alpha_1 X_i + v_i \\ 2. |e_i| &= \alpha_0 + \alpha_1 X_i^2 + v_i \\ 3. |e_i| &= \alpha_0 + \alpha_1 X_i + \alpha_2 X_i^2 + v_i \\ 4. |e_i| &= \alpha_0 + \alpha_1 X_i + \alpha_2 X_i^2 + \alpha_3 X_i^3 + v_i \end{aligned}$$

$$5. |e_i| a_0 \sqrt{a_1 X_i} v_i$$

$$6. |e_i| a_0 \sqrt{a_1 X_i^2} v_i \text{ and so on.}$$

Finally, among these regressions, we choose the form of regression which gives the best fit in the light of the correlation coefficient and the standard errors of the coefficients

and . Then, by using the coefficients (a_0 and a_1) of the chosen model, we can decide on the existence of heteroscedasticity problem as follows:

- i. if $=0$, while $\neq 0$, the situation is referred to as **strict/ strong or pure heteroscedasticity**;
- ii. if both $\neq 0$ and $\neq 0$, the case is referred to as **mixed heteroscedasticity**.

Heteroscedasticity is judged in the light of the statistical significance of and ; that is we perform any standard test of significance for these coefficients, and if they are found significantly different from zero, we accept that the 's are heteroscedastic.

The Glejser test has the advantage that it gives also information on the form of heteroscedasticity, that is, on the particular way in which is connected with X_i . This information is important, as we shall see for the 'correction' of the heteroscedastic disturbance term. Once heteroscedasticity has been established on either of the statistical tests one may experiment with various forms of the function suggested by Glejser $|e_i| = f(X_i)$ in order to decide what is the appropriate transformation of the original data required, to overcome the difficulties of heteroscedastic 's.

However, some scholars preferred other tests over the Glejser's test. For instance, J. Johnston seems to accept that the Spearman rank-correlation test and the Goldfeld and Quandt test (we will discuss it latter on) are preferable to Glejser's test for the detection of heteroscedasticity.

Goldfeld and Quandt point out that error term v_i has some problems in that its expected value is non-zero, i.e. it is serially correlated and it is heteroscedstic.

An additional difficulty with the Glejser method is that models such as $|e_i| = \sqrt{a_0 + a_1 X_i} v_i$ and $|e_i| = \sqrt{a_0 + a_1 X_i^2} v_i$ are non-linear in parameters and therefore cannot be estimated with the usual OLS procedure.

Glejser has found that for large samples the first four preceding models give generally satisfactory results in detecting heteroscedasticity. As a practical matter, therefore, the Glejser technique may be used for large samples and may be used in small samples strictly as qualitative device to learn something about heteroscedasticity.

D. Goldfield-Quandt (GQ) test

This test is mainly applicable when the sample size is large. To apply this test the number of observations on the variables included in the model must be at least twice as many as the parameters to be estimated from the model. The test assumes normality and serially independent u 's.

This popular method is applicable if one assumes that the heteroscedastic variance σ_i^2 is positively related to one of the explanatory variables in the regression model. For simplicity, consider the usual two variable models:

$$Y_i = \beta_0 + \beta_1 X_i + U_i \quad (1.1.18)$$

Suppose σ_i^2 is positively related to X_i as:

$$\sigma_i^2 = \lambda^2 X_i^2; \text{ where } \lambda^2 \text{ is constant.}$$

If the above equation is appropriate, it would mean σ_i^2 would be larger, the larger values of X_i . If that turns out to be the case, heteroscedasticity is most likely to be present in the model.

Therefore, if heteroskedasticity is supposed to be related to one of the explanatory variables and the variance of observations increases with that variable, say x , the GQ-test is an appropriate test of heteroscedasticity.

The following steps can be followed to conduct the **Goldfeld and Quandt test**.

Step 1: Formulate the hypotheses to be tested.

The customary hypotheses used in the Goldfeld and Quandt test are:

H_0 : u_i 's are homoscedastic, and is tested against the alternative hypothesis

H_1 : u_i 's are heteroscedastic (with increasing variances).

Step 2: Order or rank the observations according to the values of X_i beginning with the lowest X value

Step 3: Select arbitrarily a certain number (c) of central observations which we omit from the analysis. (It has been found from some experiments by Goldfeld and Quandt that for samples larger than $n = 30$ the optimum number of central observations to be omitted from the test is approximately a quarter of the total observations, for example 8 for $n = 30$, 16 for $n = 60$.), and divide the remaining $(n-c)$ observations into two groups each of $\frac{(n-c)}{2}$ observations, one including the small values of X and the other including the large values of X .

Step 4: Fit separate OLS regression to the first $\frac{(n-c)}{2}$ observations and the last $\frac{(n-c)}{2}$ observations, and obtain the respective residual sums of squares RSS_1 and RSS_2 ;

RSS_1 representing RSS from the regression corresponding to the smaller X_i values (the small variance group) and RSS_2 that from the larger X_i values (the large variance group).

These RSS each have

$$\frac{(n-c)}{2} - K \text{ and } \frac{(n-c-2K)}{2} \text{ df, where: } K \text{ is the number of parameters to be}$$

estimated, including the intercept term; and df is the degree of freedom.

Step 5: Compute the F-calculated value

If each of these sums is divided by the appropriate degrees of freedom, we obtain estimates of the variances of the u 's in the two sub-samples. The ratio of the two

$$F = \frac{\frac{RSS_1 / ((n-c)/2 - K)}{RSS_2 / ((n-c-2K)/2)}}{\frac{RSS_2 / ((n-c-2K)/2)}{RSS_1 / ((n-c)/2 - K)}} \quad (1.1.19)$$

has an F distribution (with $V_1 = V_2 = [(n-c)/2] - K = [(n-c-2K)/2]$ degrees of freedom, where n = total number of observations, C = central observations omitted, K = number of parameters estimated from each regression).

NB: As a rule of thumb, one should always put the larger variance in the numerator.

If the two variances are the same (that is, if the u 's are homoscedastic) the

value of F^* will tend to 1. If the variances are differ, F^* will have a large value (given that by the design of the test $\Sigma > \Sigma$).

NB: The higher the observed F^* ratio the stronger the heteroscedasticity of the u 's.

Step 6: Make decision

The observed or calculated, F^* is compared with the theoretical or tabulated value of F with $V_1 = V_2 = (n - c - 2K)/2$ degrees of freedom (at a chosen level of significance).

The theoretical (obtained from the F -tables) value of F is the value of F that defines the critical region of the test.

Decision rule

- i. If $F^* > F$, we accept that there is heteroscedasticity (that is we reject the null hypothesis of no difference between the variances of u 's in the two sub-samples).
- ii. If $F^* < F$, we accept that the u 's are homoscedastic (in other words we accept the null hypothesis).

Example 1.3:

We are interested in the rate of the return to education and estimate the coefficients of the following human capital model:

$$\ln Y = \beta_0 + \beta_1 ED + \beta_2 ED^2 + \beta_3 year + \beta_4 year^2 + U$$

We use a sample of 1483 individuals with information on hourly wages in logarithmic form ($\ln Y$), years of schooling (ED), and years of work experience ($year$).

We are going to investigate the above model to see if we can identify any heteroskedasticity using the GQ-test.

Therefore, we sort the data set in an increasing order of years of schooling and delete the 33 percent in the middle. We use the two remaining sub-samples and estimate a regression for each sample. Using the results from these regressions we calculate the corresponding variance for each regression:

$$\begin{array}{ccccccc} \Sigma & \frac{RSS_1}{n_1 - k} & \frac{59.9641}{490} & 0.1223756 & e^2 & \frac{RSS_2}{n_2 - k} & \frac{43.6254}{489} & 0.08921339 \end{array}$$

Using the estimated variances for the two sub samples we can calculate the test value:

$$F = \frac{0.1223756}{0.0892133} = 1.3717$$

Choosing a significance level of 5 percent we found a critical value of F equal to 1.16. Since our test value is larger than the critical value we conclude that our model suffers from heteroskedasticity and that year of schooling is at least partly responsible.

Most statistical tables do not offer information on critical values for the degrees of freedom we have in this example; thus, we can approximate using Excel or other packages. To approximate the critical value by using the numbers valid for infinity in both numerator and denominator would not be meaningful.

As we can see from F-table, as degree of freedom increases, the F-critical value become decreasing.

A general problem with this test is that it tends to reject the null hypothesis very often. That is, it is very sensitive to very small differences, especially when the degrees of freedom are in level with those that we have in this example, since that produce very small critical values.

E. The Breusch-Pagan test (BP)

This is also known as **Cook-Weisberg test or Lagrangian Multiplier (LM)** test and is a popular test procedure used in most applied econometric researches. The BP-test is slightly more general than the GQ-test, since it allows for more than one variable at the time to be tested and is designed to detect any linear form of heteroscedasticity.

The test does not require ordering of observations but requires the assumption of normality. It requires the specification of the relationship between the true error variance σ_i^2 and independent variable (s) Z_i .

To conduct the test, we can follow the following steps:

Assume the usual regression form, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ (1.1.20) And also suppose the error variance, σ_i^2 is a function of some variables, $\sigma_i^2 = f(\delta_0 + \delta_i Z_i)$. This implies the error variance is a linear function of an intercept and one or more independent variables, the Z_i 's, where Z_i could be some other variable or group of variables.

Step 1: Run the standard regression equation in eq-1.1. 20

$$Y_i = \beta_0 + \beta_1 X_i + e_i \quad (1.1.21)$$

Step 2: Calculate the least squares residuals ($e_i = y_i - \hat{y}_i$) from the original regression equation or obtain the residuals from this regression

Step 3: Estimate the regression variance

$$s^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} \quad (1.1.22)$$

Step 4: Run the auxiliary regression $\frac{e_i^2}{2} = Z_i + v_i$ (1.1.23)

The functional form is not expressed explicitly in (1.1.23) as stated but we are going to use a linear specification just as for the model we use.

$$H_0 : \delta_1 = \delta_2 = \dots = \delta_n = 0$$

The hypothesis of 'this test is: (1.1.24)

$$H_1 : \delta_j \neq 0 \text{ for at least one } j, j = 1, 2, \dots, P$$

Step 5: Even though it looks like we could use the classical approach of using an F-test to test the joint hypothesis, it turns out not to be possible since the dependent variable is a construction based on another model. Instead the following test statistic could be used to test the null hypothesis:

$$F = \frac{R^2 / p}{(1 - R^2) / (n - p - 1)} \quad (1.1.25)$$

where n is the number of observations used in the regression of (1.1.23) and is the coefficient of determination received from (1.1.23) It turns out that the product of those two terms is chi-squared distributed with p degrees of freedom, where p is the number of restrictions, which in this case corresponds to the number of variables included in (1.1.23). The last value should therefore be compared with a critical value received from the Chi-square table for a suitable level of significance.

In other words, after obtaining the regression sum of squares from the equation 1,1,23, we can calculate the following statistic and compare it with Chi-square table value.

$$\Theta = \frac{RSS_1}{2} \quad (1.1.26)$$

When there are (P) independent variables, the relevant test would be:

$$\frac{RSS}{2} \quad \chi^2_p$$

Step 6: Make decision

Decision rule

If the calculated χ^2 value is greater than the tabulated (critical) value at chosen significance level and the degree of freedom (number of variables in auxiliary regression excluding the constant term), we reject the null hypothesis of homoskedasticity and conclude that there is heteroscedasticity problem in the data

Example 1.4:

In this example we will use the same data set and the same model as in Example 1.3. But this time the test will involve both the variables included in the model. We choose not to include the squared terms, even though they in principle could be included. Following the basic procedure of the BP-test, we specify and estimate the variance function with standard errors given within parenthesis:

$$\hat{e}^2 = 0.063 - 0.001ED + 0.002year$$

$$(0.060) \quad (0.004) \quad (0.001)$$

$$R_e^2 = 0.0028 \quad n = 1483$$

Using this information we are able to calculate the test value:

$$LM = n R_e^2 = 1483 \times 0.0028 = 4.1524$$

Choosing significance at the 5 percent level, the Chi-square table with 2 degrees of freedom shows a critical value of 5.99. Hence, the test or calculated value is smaller than the critical value and we are unable to reject the null hypothesis. This means that we have received a conflicting result compared with the GQ-test result. Since the GQ-test is very sensitive to small differences, we believe that the result of this test is more useful. However, the BP-test is a test that requires large data sets to be valid, and is sensitive to any violation of the normality assumption. Since we have more than 1000 observations we believe that our sample is sufficiently large, but in order to be sure we will move on with yet another common test called the **White's test**.

F. White's test

This test is very similar to the BP-test but does not assume any prior knowledge of the heteroskedasticity, but instead examines whether the error variance is affected by any of the regressors, their squares or cross products. This test is designed to detect non-linear form of heteroscedasticity. It is based on an auxiliary regression with squared residuals as dependent variable and regressors given by: the regressors of the initial model, their squares and their cross products. Therefore, it is also a large sample test but it does not depend on any normality assumption. Hence, this test is more robust than the other test procedures described above, and is sometimes also called the **White's General Heteroskedasticity test (WGH)**.

The basic steps in the procedure are as follows for a model with two explanatory variables, where (1.1. 27) represents the main model and (1.1. 28) the variance function that contains all the variables of the main function and their squares and cross products:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i \quad (1.1. 27)$$

$$U_i^2 = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 X_{1i}^2 + \gamma_4 X_{2i}^2 + \gamma_5 X_{1i} X_{2i} + U_i \quad (1.1.28)$$

Step 1: Estimate the parameters of equation (1.1.27) and create and save the residual.

Step 2: Square the residual and run the auxiliary regression model given by (1.1.28).

Step 3: Using the results from the auxiliary regression you can calculate the test value using (1.1.25).

Step 4: Making decision

Decision rule

If the test value is larger than the critical value at chosen significance level and the degree of freedom (number of variables in auxiliary regression excluding the constant term), you reject the null hypothesis of homoskedasticity.

Example 1.5:

We repeat the test executed in Example 1.4 and apply the WGH-test instead. Observe that the only difference is in the specification of the variance function. Following the basic steps given above we received the following results with standard errors reported within parenthesis:

$$\begin{array}{l} \hat{e}^2 \quad 0.230 \ 0.023 \ ED \ 0.001 \ year \ 0.001 \ ED^2 \ 0.000 \ year^2 \ 0.000 \ ED \times year \ 0.196 \ 0.024 \\ \quad \quad 0.007 \ 0.001 \ 0.000 \ 0.000 \\ R_e^2 \quad 0.0039 \ n \quad \quad 1483 \end{array}$$

Observe that the coefficients and their standard errors are different from zero even though some of them appear to be zero since they are expressed with just three decimal points. Their t-values are definitely different from zero. Using these results we can calculate the test value:

$$LM \quad nR_e^2 \quad 0.0039 \times 1483 \quad 5.78$$

The critical value from the Chi-square table, with 5 degrees of freedom and a significance level of 5 percent, equals 11.07, which is larger than the test value. Hence this test confirms the conclusions from the previous test and we are unable to reject the null hypothesis of homoscedasticity.

Activity 1.1.1

The following table shows the personal savings and personal income (in millions of dollar) of a country over a 31 year period.

Period	Savings (S)	Income (X)	Estimated savings (by OLS)	OLS Residuals
1	264	8,777	97.85	166.15
2	105	9,210	134.46	-29.46
3	90	9,954	197.35	-107.35
4	131	10,508	244.18	-113.18
5	122	10,979	250.18	-128.18
6	107	11,912	362.87	-255.86
7	406	12,747	433.46	-27.45
8	503	13,499	497.02	5.97
9	431	14,269	562.12	-131.12
10	588	15,522	668.04	-80.04
11	898	16,730	770.15	127.84
12	950	17,663	849.03	100.97
13	779	18,575	926.12	-147.12
14	819	19,635	1015.73	-196.73
15	1,222	21,163	1144.90	77.10
16	1,702	22,880	1290.04	411.96
17	1,578	24,127	1395.45	182.54
18	1,654	25,604	1520.32	133.68
19	1,400	26,500	1596.06	-196.05
20	1,829	27,670	1694.96	134.04
21	2,200	28,300	1748.22	451.78
22	2,017	27,430	1674.67	342.33
23	2,105	29,560	1854.73	250.27
24	1,600	28,150	1735.54	-135.54
25	2,250	32,100	2069.45	180.55
26	2,420	32,500	2103.26	316.74
27	2,570	35,250	2335.73	234.27
28	1,720	33,500	2187.80	-467.80
29	1,900	36,000	2399.14	-499.14
30	2,100	36,200	2416.04	-316.04
31	2,300	38,200	2585.11	-285.11

Check for possible existence of heteroscedasticity using:

- i. Spearman's Rank correlation test and
- ii. Goldfeld and Quandt test

1.1.7. Solutions for Heteroscedastic Disturbances

Dear Distance learners, in the discussions we made so far in sub-section 1.1.6, we tried to detect the heteroscedasticity problem using informal as well as formal methods. However, the issue should not be simply detection; we have to try to solve the problem. And also we have seen that, heteroscedasticity does not destroy the unbiasedness and consistency property of the OLS estimators, but they are no longer efficient. This lack of efficiency makes the usual hypothesis testing procedure of dubious value. *Therefore, remedial measures concentrate on the variance of the error term.*

The appropriate solution is to transform the original model in such a way as to obtain a form in which the transformed disturbance term has constant variance. We, then apply the method of classical least squares to the transformed model.

Applying OLS to the transformed variables is known as the method of Weighted Least squares (WLS) or Generalized Least Squares (GLS). This method is used when the nature of the heteroscedasticity (the true error variance of each observation) is known or unknown; however, the former is less likely operational in practice since the population parameters (β 's) are unknown. Thus, we will stick to the latter.

In short WLS or GLS is OLS on the transformed variables that satisfy the standard least squares assumptions. The estimators thus obtained are known as GLS estimators, and it is these estimators that are BLUE.

The transformation of the model reduces to the adjustment of the original data. The adjustment of the model depends on the particular form of homoscedasticity, that is, on the form of the relationship between the variance and the values of the explanatory variable(s): = ()

In general, the transformation of the original model consists in dividing through the original relationship by the square root of the term which is responsible for the homoscedasticity.

To illustrate this statement assume the original model:

$$Y_i = \beta_0 + \beta_1 X_i + U_i, \quad (1.1.29)$$

Where: U_i is heteroscedastic, but satisfies all other assumption of the stochastic term.

If we apply OLS to the above then it will result in inefficient parameters since $\text{var}(u_i)$ is not constant.

The remedial measure is transforming the above model so that the transformed model satisfies all the assumptions of the classical regression model including homoscedasticity.

Case I: Suppose the heteroscedasticity is of the form $\text{var}(u_i) = k^2 X_i^2$ (where k is a finite constant to be estimated from the model), that is, the variance of U increases proportionately with X^2 . Solving for the constant factor of proportionality, k^2 , we obtain $k^2 = \frac{\text{var}(u_i)}{X_i^2}$. This suggests that the appropriate transformation of the original model consists of the division of the original relationship by X_i , which means that the appropriate transformation version is

$$\frac{Y_i}{X_i} = \frac{\beta_0}{X_i} + \beta_1 + \frac{u_i}{X_i} \quad (1.1.30)$$

The new transformed random term $\frac{u_i}{X_i}$, is homoscedastic.

Proof

The variance of $\frac{u_i}{X_i}$ can be given in the following form

$$\text{var}\left(\frac{u_i}{X_i}\right) = \frac{1}{X_i^2} \text{var}(u_i) = \frac{1}{X_i^2} k^2 X_i^2 = k^2$$

In our case it was assumed that $\text{var}(u_i) = k^2 X_i^2$

Thus, $\text{var}\left(\frac{u_i}{X_i}\right) = k^2$, which is constant.

Therefore the new random term has a finite constant variance (equal to k^2), hence we can conclude that through the above transformation, we solved the problem of heteroscedasticity. Consequently, we can apply classical least squares to the transformed version of the model.

Note that in the above transformation the position of the coefficients has changed: the parameter of the variable $\frac{1}{X_i}$ in the transformed model is the constant intercept of the original model, while the constant of term of the transformed model is the parameter of the explanatory variable X in the original model. Therefore, to get back to the original model, we shall have to multiply the estimated regression by K_i .

Case II. Suppose the heteroscedasticity is of the form: $(u_i^2) = k^2 X_i$

The appropriate transformation of the original model consists of the division of the original relationship by $\sqrt{X_i}$

The transformed model is:

$$\frac{Y_i}{\sqrt{X_i}} = \beta_0 \frac{1}{\sqrt{X_i}} + \beta_1 \frac{X_i}{\sqrt{X_i}} + \frac{U_i}{\sqrt{X_i}} \quad (1.1.31)$$

Note that the transformed random term $\frac{U_i}{\sqrt{X_i}}$ in the transformed model given in equation 1.1.31 is homoscedastic with constant variance equal to k^2 .

Proof

The variance of the stochastic term in the transformed model can be given as

$$\frac{u_i^2}{X_i} = \frac{1}{X_i} (U_i)^2 = \frac{1}{X_i} k^2 X_i = k^2 \text{ Constant variance.}$$

Hence, we can conclude that with the above transformation we solved the problem of heteroscedasticity, and hence we can estimate the transformed model with OLS.

There is no intercept term in the transformed model. Therefore, one will have to use the 'regression through the origin' model to estimate β 's. In this case, therefore, to get back to the original model, we shall have to multiply the estimated regression by $\sqrt{X_i}$.

In practice this is carried out by just transforming Y_i , X_i and creating a new constant equal to $1/\sqrt{X_i}$. Hence when running this specification in a computer you have to ask the software to run the regression through the origin, since we now have a model specific constant that moves with X_i . All computer software has that option, and once you have found how to do that, you just regress $\frac{Y_i}{\sqrt{X_i}}$ on $\frac{1}{\sqrt{X_i}}$, $\frac{X_i}{\sqrt{X_i}}$, $\frac{U_i}{\sqrt{X_i}}$.

When you transform the variables in this way, you automatically transform the error term, which now is divided by the square root of X_i .

Case III: Suppose heteroscedasticity is of the form

$$(\sigma^2) = (h(X_i))$$

Then, the appropriate transformation will be to divide the original equation by

$$\frac{Y_i}{h(X_i)} = \frac{\beta_0}{h(X_i)} + \frac{\beta_1 X_i}{h(X_i)} + \frac{u_i}{h(X_i)}, \text{ i.e.,} \quad (1.1.32)$$

Now we can show that the stochastic term of the transformed model given in equation 1.1.32, has constant variance.

Proof

$$\begin{aligned} \text{Var}\left(\frac{u_i}{h(X_i)}\right) &= \frac{1}{h(X_i)^2} \text{E}(u_i^2) \\ &= \frac{1}{h(X_i)^2} (h(X_i)) = \frac{1}{h(X_i)}, \text{ which is constant.} \end{aligned}$$

Hence, with the above transformation we solved the problem of heteroscedasticity, and hence we can estimate the transformed model with OLS.

NB: In general if the heteroscedasticity is of the form: $(\sigma^2) = (h(X_i))$

(where k is a finite constant and $f(X_i)$ is a function of the X 's), the solution is the transformation of the original function by dividing through the original model by $\frac{1}{h(X_i)}$.

In applied research, econometricians usually assume that the variance of u , changes in proportion to the square of the explanatory variable. The main justification for this assumption is that the resulting transformation is computationally easy. Thus, the first case is commonly operational.

Example 1.6

Given the data in Activity 1.1 and it is verified that there is heteroscedasticity problem in saving function (see check your progress).

Thus, following the procedures below, we can solve the problem.

We assume that the pattern of heteroscedasticity is

So that the appropriate transformation of the original model is

The transformed variables (S_t/X_t and $1/X_t$) are shown in the last two columns on table below.

Period	Savings S	Income X	Estimated savings (by OLS)	OLS Residuals	S/X (x 1,000)	1/X (x 10,000)
1	264	8,777	97.85	166.15	30.079	1.139
2	105	9,210	134.46	-29.46	11.401	1.086
3	90	9,954	197.35	-107.35	9.042	1.005
4	131	10,508	244.18	-113.18	12.467	0.952
5	122	10,979	250.18	-128.18	11.112	0.911
6	107	11,912	362.87	-255.86	8.983	0.839
7	406	12,747	433.46	-27.45	31.851	0.785
8	503	13,499	497.02	5.97	37.262	0.741
9	431	14,269	562.12	-131.12	30.205	0.701
10	588	15,522	668.04	-80.04	37.882	0.644
11	898	16,730	770.15	127.84	53.676	0.598
12	950	17,663	849.03	100.97	53.785	0.566
13	779	18,575	926.12	-147.12	41.938	0.538
14	819	19,635	1015.73	-196.73	41.711	0.509
15	1,222	21,163	1144.90	77.10	57.742	0.473
16	1,702	22,880	1290.04	411.96	74.388	0.437
17	1,578	24,127	1395.45	182.54	65.404	0.414
18	1,654	25,604	1520.32	133.68	64.599	0.391
19	1,400	26,500	1596.06	-196.05	52.830	0.377
20	1,829	27,670	1694.96	134.04	66.100	0.361
21	2,200	28,300	1748.22	451.78	77.739	0.353
22	2,017	27,430	1674.67	342.33	73.533	0.365
23	2,105	29,560	1854.73	250.27	71.211	0.338
24	1,600	28,150	1735.54	-135.54	56.838	0.355
25	2,250	32,100	2069.45	180.55	70.093	0.312
26	2,420	32,500	2103.26	316.74	74.462	0.308
27	2,570	35,250	2335.73	234.27	72.908	0.284
28	1,720	33,500	2187.80	-467.80	51.343	0.299
29	1,900	36,000	2399.14	-499.14	52.778	0.278
30	2,100	36,200	2416.04	-316.04	58.011	0.276
31	2,300	38,200	2585.11	-285.11	60.209	0.262

Applying OLS to the new variables we obtain

$$\begin{aligned} &= -718.88 + 0.088X_t \\ (71.27) & \quad (0.004) \\ R^2 &= 0.770 \end{aligned}$$

Following the same steps of Spearman rank correlation test of activity, the value of the Spearman correlation coefficient of the transformed equation is $r' = -0.22$, which is less than its standard error of ≈ 0.14 shows that the transformation has eliminated heteroscedasticity.

Thus the new savings function is

$$\begin{aligned} S_t^* &= -718.88 + 0.088X_t & R^2 &= 0.770 \\ (71.27) & (0.004) \end{aligned}$$

As compared with the old equation

$$\begin{aligned} S_t &= -644.1 + 0.085X_t & R^2 &= 0.903 \\ (117.6) & (0.005) \end{aligned}$$

Note that the R^2 of the transformed model is lower than R^2 in the original model. This is due to the way in which weights are assigned to the individual observations in the transformation process.

1.2 Autocorrelation

1.2.1. Introduction

Dear Distance Learners, as you might remember from your discussion on simple and multiple classical regression models of Econometrics I, one of the assumptions is that the successive values of error terms are uncorrelated; in other words, the correlation between error terms is equal to zero. i.e., $\text{cov}(u_i, u_j) = 0$. However, this assumption might not be always satisfied in practice. If the assumption is not satisfied, that is, if the value of U in any particular period is correlated with its own preceding value(s), we say that there is autocorrelation of the random terms.

Hence, *autocorrelation is defined as a correlation between successive error or stochastic terms which are ordered in time.*

Dear students, is there any difference between correlation and autocorrelation?

Yes, there is difference between two concepts. Autocorrelation is a special case of correlation which refers to the relationship between successive values of the same variable (disturbance terms), while correlation refers to the relationship between two or more different variables (independent variables). Autocorrelation is also sometimes called as serial correlation but strictly speaking the two terms may be distinct; thus, some economists distinguish between these two terms. According to G.Tinner, autocorrelation is the lag correlation of a given series with itself but lagged by a number of time units. The term serial correlation is defined by him as “lag correlation between two different series.” Thus, correlations between two time series such as $u_1, u_2, \dots, u_{10}$ and $u_2, u_3, \dots, u_{11}$, where the former is the latter series lagged by one time period, is autocorrelation. Whereas correlation between time series such as $u_1, u_2, \dots, u_{10}$ and $v_1, v_2, \dots, v_{11}$ where U and V are two different time series, is called serial correlation. *Although the distinction between the two terms may be useful, we shall treat these terms synonymously in our subsequent discussion.*

We are quite likely to encounter only uncorrelated errors when dealing with cross-sectional data. However, autocorrelation or serial correlation- the correlation of an error term and a lagged version of itself is likely to occur with time-series data because errors made in a particular time period are readily carried over to future time periods.

For example, an underestimate of the GNP in one year can generate more underestimates in future time periods.

This sub-section will discuss the most important issues related to autocorrelation that an applied researcher need to be aware of, such as its nature and causes, effect on the estimated parameters when ignored, how to detect it and how to solve the problem when present.

1.2.2. Nature of autocorrelation

An auto correlated error term can take a range of different specifications to manifest a correlation between pair wise observations. The most basic form of autocorrelation is referred to as the **first order autocorrelation** and is specified in the following way:

$$U_t = \rho U_{t-1} + V_t \quad (1.2.1)$$

Where, U refers to the error term of the population regression function.

As can be seen from (1.2.1) the error term at period t is a function of itself in the previous time period $t-1$ times the coefficient, ρ , which is referred to as the **first order autocorrelation coefficient** (This is the Greek letter rho, pronounced "row"). The last term V_t , is a so called **white noise error term**, and suppose to be completely random. It is often assume to be standard normal and satisfy the following assumptions:

$$(v) = 0, \quad (v^2) = \sigma_v^2 \quad \text{and} \quad (v_i v_j) = 0 \quad \text{for } i \neq j$$

This type of autocorrelation is called **auto regression** because the error term is a function of its past values. Since U is a function of itself one period back only, as oppose to several periods, we call it the **first order auto regression error scheme or first-order Markov process**, which is denoted AR(1).

On the other hand, if the values of U in a certain period depend on its value in the two previous periods, then the form of autocorrelation is called **second-order autoregressive scheme**, and so on.

This specification can be generalized to capture up to n terms. We would then refer to it as the n^{th} order of autocorrelation and it would be specified like this:

$$U_t = \rho_1 U_{t-1} + \rho_2 U_{t-2} + \dots + \rho_n U_{t-n} + V_t \quad (1.2.2)$$

Nonetheless, in most applied research it is assumed that if autocorrelation exists, it takes the form of the *simple first-order autoregressive scheme*. Consequently, in this module we will give emphasis to first-order autoregressive scheme.

The autocorrelation can be positive or negative in theory, and is related to the sign of the autocorrelation coefficient in (1.2.1).

- If $\rho > 0$, then there is positive serial correlation, meaning that unusually high values of the dependent variable tend to occur together and that unusually low values of the dependent variable tend to occur together.
- If $\rho < 0$, then there is negative serial correlation, meaning that an unusually high values is likely to be followed by an unusually low values of the dependent variable. That is, values tend to fluctuate back and forth from period to period.

However, in practice it is in most cases positive. The main reasons for this are economic growth and cyclic movements of the economy and interdependence among most macroeconomic variables. Most economic variables tend to grow in periods of growth or they tend to show cyclical pattern.

Example: If last month's unemployment rate was unusually high, then this month's unemployment rate will probably be unusually high too.

Dear students, why do you think we use ρ as a coefficient of u_{t-1} in equation

1.2.1? -----

Because this equation represents a linear relationship between u_t and u_{t-1} , where ρ is the slope coefficient and the constant intercept is equal to zero. Furthermore, equation 1.2.1 is a simple linear regression model with suppressed constant term.

Then, if we apply OLS on the model given in (1.2.1) we obtain:

$$\hat{u}_t = \frac{\sum_{t=1}^n u_t}{\sum_{t=1}^n u_t^2} \quad (1.2.3)$$

On the other hand, the autocorrelation coefficient, ρ , between u_t and u_{t-1} is given as:

$$\rho = \frac{\sum_{t=1}^n u_t u_{t-1}}{\sum_{t=1}^n u_t^2} \quad (1.2.4)$$

Equation 1.2.4 is a special form of correlation coefficients between any two variables (X and Y), which are defined as:

$$r = \frac{\sum_{t=1}^n x_t y_t}{\sqrt{\sum_{t=1}^n x_t^2 \sum_{t=1}^n y_t^2}} \quad (1.2.5)$$

Given that for large samples: $u_t^2 \approx u_{t-1}^2$, we observe that coefficient of autocorrelation represents a simple correlation coefficient r .

$$\hat{u}_t = \frac{\sum_{t=1}^n u_t}{\sum_{t=1}^n u_t^2} = \frac{\sum_{t=1}^n u_t}{\sqrt{\sum_{t=1}^n u_t^2 \sum_{t=1}^n u_{t-1}^2}} r \quad (1.2.6)$$

$$\rho = \frac{\sum_{t=1}^n u_t u_{t-1}}{\sum_{t=1}^n u_t^2} = \frac{\sum_{t=1}^n u_t u_{t-1}}{\sum_{t=1}^n u_t^2} \quad (1.2.7)$$

Now, we can see that equations 1.2.3 and 1.2.7 are similar, and this is the reason why in most text books, the simple first-order autoregressive scheme is given as: $U_t = U_{t-1} + V_t$

$1 \approx 1$ since $1 \approx r \approx 1$

This proves the statement “we can treat autocorrelation in the same way as correlation in general”. From our statistics background we know that:

- ✓ if the value of r is 1 we call it perfect positive correlation, if r is -1, perfect negative correlation and if the value of r is 0, there is no correlation.

By the same analogy if the value of $\hat{\rho}$ is 1 it is called perfect positive autocorrelation, if $\hat{\rho}$ is -1 it is called perfect negative autocorrelation and if 0, no autocorrelation.

NB: If $\hat{\rho} = 0$ in $u_t = \rho u_{t-1} + v_t$ i.e. u_t is not correlated.

Mean, Variance and Covariance of disturbance terms in autocorrelated model:

To examine the consequences of autocorrelation on ordinary least square estimators, it is required to study the properties of U . If the values of U are found to be correlated with simple Markov process, then it becomes: $U_t = \rho U_{t-1} + v_t$ with ρ between -1 and 1 and v_t fulfilling all the usual assumptions of a disturbance term. Our objective, here is to obtain value of u_t in terms of autocorrelation coefficient and random variable v_t .

Then, the complete form of the first-order Markov process (i.e. the pattern of autocorrelation for all the values of U) can be developed as follows:

We know that

$$U_t = f(U_{t-1}) + v_t \quad (1.2.8)$$

Then

$$U_t = f(U_{t-1}) + v_t \quad (1.2.9)$$

$$U_t = f(U_{t-1}) + v_t \quad (1.2.10)$$

$$U_t = f(U_{t-1}) + v_t \quad (1.2.11)$$

Thus, in order to define the error term in any particular period t for the first order Markov process, we work as follows. We start from the autocorrelation relationship in period t and we make use of above relations to perform continuous substitutions in $U_t = \rho U_{t-1} + v_t$ as follows.

Substituting equation 1.2.9 into equation 1.2.8 yields:

$$\begin{aligned}
 U_t &= U_{t-1} + v_t \\
 (U_{t-2} + v_{t-1}) + v_t &= U_{t-2} + v_{t-1} + v_t \\
 &= U_{t-2} + v_{t-1} + v_t
 \end{aligned} \tag{1.2.12}$$

Substituting equation 1.2.10 into equation 1.2.12 we get:

$$\begin{aligned}
 &= U_{t-3} + v_{t-3} + v_{t-2} + v_{t-1} + v_t \\
 &= U_{t-3} + v_{t-3} + v_{t-2} + v_{t-1} + v_t
 \end{aligned} \tag{1.2.13}$$

In this way, if we continue the substitution process for r lag periods (assuming that r is very large), we shall obtain:

$$= U_{t-r} + v_{t-r} + v_{t-r+1} + v_{t-r+2} + \dots + v_t \tag{1.2.14}$$

The correlation coefficient, $|\rho| < 1$. Then, as the power of ρ increases to infinity, the term with the lagged u_t , (i.e. $[\rho^r u_{t-r}]$, tends to zero. In other words, as $r \rightarrow \infty$, $\rho^r \rightarrow 0$.

Thus, u_t in equation 1.2.14 can be written as

$$\begin{aligned}
 U_t &= v_t + v_{t-1} + v_{t-2} + v_{t-3} + \dots + v_{t-r} \\
 &= v_t + v_{t-1} + v_{t-2} + v_{t-3} + \dots + v_{t-r}
 \end{aligned} \tag{1.2.15}$$

Thus, equation 1.2.15 gives the value of the error term when it is autocorrelated with a first-order autoregressive scheme.

Finally, we will establish the mean, variance and covariance using the value of u_t ; autocorrelated with a first-order autoregressive scheme.

1. Mean of autocorrelated u_t

The mean of u_t is its expected value. Hence, to find the mean of u_t , we take the expected value of both sides of equation 1.2.15.

$$(U_t)' = v_t + v_{t-1} + v_{t-2} + v_{t-3} + \dots + v_{t-r} \tag{1.2.16}$$

It is evident from equation 1.2.16 that the mean of an auto-correlated stochastic term, u_t , is zero if the autocorrelation is simple first-order autoregressive scheme.

2. Variance of autocorrelated u_t

By the definition of variance

$$\text{var}(U_t) = \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} \rho^r \rho^s \text{var}(v_{t-r}) \text{var}(v_{t-s});$$

$$\text{since } \text{var}(v_{t-r}) = \text{var}(v_{t-s}) = \sigma^2$$

$$\text{var}(U_t) = \sigma^2 \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} \rho^{2r+2s} = \sigma^2 \sum_{r=0}^{\infty} (1 + \rho^2 + \rho^4 + \rho^6 + \dots)$$

Note that the expression in the bracket is the sum of a *geometric progression of infinite* terms, whose first term is 1 and the common ratio is ρ^2 .

Following the summation rule of a geometric progression, the sum of the terms in the bracket converges to $1/(1 - \rho^2)$ since $\rho < 1$. Therefore,

$$\text{var}(U_t) = \frac{\sigma^2}{(1 - \rho^2)} \quad (1.2.17)$$

Thus, variance of autocorrelated u_i is $\frac{\sigma^2}{1 - \rho^2}$ which is a constant value.

From the above, the variance of U_i depends on the nature of variance of V_i . If the variance of V_i is homoscedastic, U_i is homoscedastic and if V_i is heteroscedastic, U_i is heteroscedastic.

However, equation 1.2.17 shows that the variance of an error term is constant even though it is autocorrelated, provided that it is autocorrelated with first order autoregressive scheme.

3. The covariance of auto-correlated u_t

By the definition of covariance: $\text{Cov} = E(U_t U_{t-1})$

$$\text{Since } u_t = v_t + \rho v_{t-1} + \rho^2 v_{t-2} + \dots$$

$$u_{t-1} = v_{t-1} + \rho v_{t-2} + \rho^2 v_{t-3} + \dots$$

Substituting the above two equations in the definition, we obtain

$$\begin{aligned}
\text{cov}(U_t, U_{t-1}) &= (v_t - v_{t-1} - \rho^2 v_{t-2} - \dots)(v_{t-1} - v_{t-2} - \rho^2 v_{t-3} - \dots) \\
&= \{v_t - (v_{t-1} - v_{t-2} - \dots)\}(v_{t-1} - v_{t-2} - \rho^2 v_{t-3} - \dots) \\
&= [v_t(v_{t-1} - v_{t-2} - \dots) - (v_{t-1} - v_{t-2} - \dots)^2]; \text{ since } E(v_t v_{t-r}) = 0 \\
&= 0 - (v_{t-1} - v_{t-2} - \dots)^2 \\
&= -(v_{t-1} - v_{t-2} - \dots)^2 \\
&= (v_{t-1}^2 - 2v_{t-1}v_{t-2} + \dots \text{ 2 times cross products}) \\
&= (v_{t-1}^2 - 2v_{t-2}^2 + \dots - 0) \\
&= (v_{t-1}^2(1 - \rho^2 - \rho^4 - \dots)) \\
&= \frac{v_{t-1}^2}{1 - \rho^2}
\end{aligned} \tag{1.2.18}$$

Since $|\rho| < 1$

$$\text{cov}(U_t, U_{t-1}) = \frac{\sigma_u^2}{1 - \rho^2} \tag{1.2.19}$$

Similarly $\text{cov}(u_t, u_{t-2}) = \rho^2 \sigma_u^2$

$\text{cov}(U_t, U_{t-3}) = \rho^3 \sigma_u^2$ and generalizing $\text{cov}(U_t, U_{t-r}) = \rho^r \sigma_u^2$ (for $r \neq t$)

From this pattern we can conclude that as the lag, say r , increases (i.e., as the lag gets far apart from the current period) the (spill over) effect of U_{t-r} on U_t declines. This is due to the fact that $\rho < 1$, and hence ρ^r declines and approaches zero as r approaches infinite

In summarizing the above cases, we can conclude that when there is autocorrelation of first-order autoregressive scheme, the auto-correlated disturbance term has the following characteristics:

$u_t \sim N\left(0, \frac{\sigma_u^2}{1 - \rho^2}\right) \text{ and } \text{Cov}(U_t, U_{t-r}) = E(U_t U_{t-r}) = 0$
--

1.2.3. Sources of autocorrelation

In practical exercise, autocorrelation between the values of U may arise from different sources. The possible sources of autocorrelation for the values of the error term are discussed as follows.

a. *Cyclical fluctuations or inertia*

Time series such as GNP, price index, production, employment and unemployment exhibit business cycle. Starting at the bottom of recession, when economic recovery starts, most of these series move upward. In this upswing, the value of a series at one point in time is greater than its previous value. Thus, there is a momentum built in to them, and it continues until something happens (e.g. increase in interest rate or tax) to slowdown them. Thus successive time series data are likely to be interdependent. Therefore, in regression involving time series data, successive error terms are likely to be correlated.

b. *Specification Bias*, which can be manifested by: i.

Excluded or omitted explanatory variable(s)

It is customary that most economic variables tend to be autocorrelated. If an autocorrelated variable has been excluded from the set of explanatory variables, obviously its influence will be reflected in the random variable U . This case may be called “**quasi – autocorrelation**”, since it is due to the auto-correlated pattern of omitted explanatory variables; and not due to the behavioral pattern of the values of the true U . For example, suppose we have the following demand model:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \quad (1.2.20)$$

Where Y = quantity of banana demanded, X_1 = price of banana, X_2 = consumer income, X_3 = price of mango, and t = time. However, for some reason we run the following regression:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + v_t \quad (1.2.21)$$

Now if (1.2.20) is the “correct” model or the “truth” or true relation, running (1.2.21) is practically the same as to letting $v_t = \beta_3 X_{3t} + u_t$. And to the extent the price of mango affects the consumption of banana, the error or disturbance term v will reflect a systematic pattern, thus creating quasi-autocorrelation.

From the discussion of the preceding paragraph it should be obvious that the assumption of temporal independence of the values of u can be easily violated in practice. Thus taking into account that in most applied econometric research only the most important (three or four) explanatory variables are included explicitly in the function; it is natural to expect that omitted variables are a frequent cause of 'quasi-autocorrelation'. In particular, if we use time series it is almost certain that some at least of these omitted variables will be serially correlated.

However, if several correlated X 's are omitted from the model, U may not be autocorrelated, since the autocorrelation patterns of the omitted regressors may be such as to offset each other.

ii. Incorrect functional form of the model (functional misspecification)

If we have adopted a mathematical form that differs from the true form of the relationship, the values of U may show serial correlations. For example if we have chosen a linear function while the true relationship between Y and the X 's is of a non-linear form like quadratic, the values of U will be temporally dependent.

Suppose the "true" or correct model in a cost-output study is as follows:

$$\text{Marginal cost}_i = \beta_0 + \beta_1 \text{output}_i + \beta_2 \text{output}_i^2 + u_i \quad (1.2.22)$$

But we fit the following model:

$\text{Marginal cost}_i = \beta_0 + \beta_1 \text{output}_i + v_i$ (1.2.23) The linear marginal cost curve will consistently underestimate or overestimate the true marginal cost. This is because the disturbance term v_i is, in fact, equal to $\text{output}_i^2 + u_i$, and hence will catch the systematic effect of the output_i^2 term on marginal cost. In this case, v_i will reflect autocorrelation because of the use of an incorrect functional form.

c. Manipulating data: e.g. taking averages, interpolation, extrapolation

Most of the published time series data involve some interpolation and "smoothing" processes. In this process, the average of the values disturbance term over successive time periods will be taken. Consequently, the successive values of U are interrelated and exhibit autocorrelation pattern.

d. Misspecification of the true random term, U

It may well be expected in many cases that the successive values of the true U are corrected. Thus even the purely random factors (such as wars, drought, storms, strikes etc) impose influences that are spread over more than one period of time. For example a strike will have disruptive effects on the production process which will persist through several future periods. An exceptionally low cropping period in the agricultural sector, caused by abnormal weather conditions, will influence the performance of almost all other economic variables in several time periods; and so on. These factors result in serially dependent values of the disturbance term, U , so that if we assume $E(u_i u_j) = 0$, we will really misspecify the true pattern of the values of U . This case of autocorrelation may be called “**true-autocorrelation**” because its root lies in the U term itself.

e. Cobweb Phenomenon

As you have discussed in *Agricultural Economics*, the supply of many agricultural commodities reflects the so-called cobweb phenomenon, where supply reacts to price with a lag of one time period because supply decisions take time to implement (the gestation period). Thus, at the beginning of this year’s planting of crops, farmers are influenced by the price prevailing last year, so that their supply function is

$$\text{Supply}_t = \beta_0 + \beta_1 P_{t-1} + u_t \quad (1.2.24)$$

Suppose at the end of period t , price P_t turns out to be lower than P_{t-1} . Therefore, in period $t + 1$ farmers may very well decide to produce less than they did in period t . Obviously, in this situation the disturbances u_t are not expected to be random because if the farmers overproduce in year t , they are likely to reduce their production in $t + 1$, and so on, leading to a **Cobweb pattern**.

f. Neglecting lags from the regression model

For various reasons the behavior of some economic variables does not change readily. Example: Consumers do not change their consumption habits readily for psychological, technological or institutional reasons.

$$= + + + \quad (1.2.25)$$

Where:

- is consumption at time t which is a function of its lag value and current income, - is income at time t , - is consumption at time $t-1$ and - is error term at time t .

If the lagged term of the dependent variable is neglected, the error term will reflect a systematic pattern which indicates autocorrelation in the model.

g. Non-stationarity

When dealing with time series data, we should check whether the given time series is stationary. A time series is stationary if and only if its characteristics (mean, variance and covariance) are time variant, that is, they do not change over time. If that is not the case, we have a non-stationary time series which causes autocorrelation.

h. Data transformation

It is common in economic analysis that, the data may be converted in to natural logarithms some specified goal or to some fractions and common denominators; hence autocorrelation problem may be encountered.

It should be noted that the source of autocorrelation has a strong bearing on the solution which must be adopted for the 'correction' of the incidence of serial correlation. In other words the type of corrective action in each particular econometric application depends on the cause or source of autocorrelation. We will discuss this topic in a subsequent section.

1.2.4. Consequences of Autocorrelation

The consequences of having an autocorrelated error term are very similar to those that appear with a heteroskedastic error term. In short we have that:

1. Autocorrelation does not affect the linearity, consistency and unbiasedness of OLS estimators. That is, the estimated slope coefficients are linear, unbiased (in the sense that their expected value is equal to the true parameters) and consistent.

In previous section we established that

$$= + \Sigma \epsilon$$

Taking expected values of both sides, we obtain:

$$= + \frac{\Sigma \epsilon}{n} \quad (1.2.26)$$

The statistical bias is the difference between the expected value of the estimate and the true parameter, we have

$$\text{Bias in } \hat{\beta} = E(\hat{\beta}) - \beta = \frac{\sum (X_i u_i)}{\sum X_i^2} \quad (1.2.27)$$

But we found that the expected value of the random term u , even when auto correlated, is zero: $E(u_i) = 0$; thus, $\text{bias in } \hat{\beta} = 0$. That is, the property of unbiasedness and consistency does not require uncorrelated error terms.

Thus, irrespective of whether the residuals are serially independent or not, the estimates of the parameters have no statistical bias, so long as the u 's and X 's are uncorrelated.

2. However, in the presence of autocorrelation, OLS estimators are inefficient (do not have minimum variance) since the OLS variances of the parameter estimates are likely to be larger than those of other econometric methods.

a. *With positive autocorrelation the standard errors are biased and too small.*

b. *With negative autocorrelation the standard errors are biased and too large.*

We will prove that the OLS variance of the estimate $\hat{\beta}$ is larger when the u 's are positively autocorrelated, and the values of X are also positively auto correlated. These conditions are often met in the **real economic world**.

It can be shown that the variance of the least squares estimate $\hat{\beta}$ is given by the formula

$$\text{Var}(\hat{\beta}) = \frac{\sum u_i^2}{\sum X_i^2} + 2 \frac{\sum X_i u_i \sum X_j u_j}{\sum X_i^2 \sum X_j^2} \quad (1.2.28)$$

When the u 's are serially independent $E(u_i u_j) = 0$ and the last term disappears, so that the variance of becomes:

$$\text{Var}(\hat{\beta}) = \frac{\sum u_i^2}{\sum X_i^2} \quad (1.2.29)$$

However, with u 's related with a first-order autoregressive scheme we have: $(\hat{\beta}) =$

And $(\hat{\beta}) =$

Therefore $\text{Var}(\hat{\beta}) = \frac{\sum u_i^2}{\sum X_i^2} + 2 \frac{\sum X_i u_i \sum X_j u_j}{\sum X_i^2 \sum X_j^2} \quad (1.2.30)$

Expanding the second term we obtain the following result for the variance of when there is autocorrelation in the function

$$= \frac{\sum u_i^2}{\sum X_i^2} + 2 \frac{\sum X_i u_i \sum X_j u_j}{\sum X_i^2 \sum X_j^2} + \dots$$

$$= \frac{1}{n} \sum_{i=1}^n (1 + 2\rho + 2\rho^2 + \dots + 2\rho^{n-1}) \quad (1.2.31)$$

If we compare the two variances of (for non-autocorrelation and for serially dependent u 's) we see that the simple least squares variance does not take into account the terms in parentheses.

If ρ is positive and if the successive values of X (in consecutive time periods) are positively autocorrelated, the expression in parentheses is almost certainly greater than unity, and hence the true variance of β_1 will be larger than the simple formula

$\text{var}(\beta_1) = \sigma^2 / \sum x_i^2$ implies. In other words we are in a situation where

autocorrelation causes the *actual (true)* variance of β_1 to increase, while the *estimate* of this variance (from the application of the above simple formula) will be smaller: the simple OLS formula will underestimate the true variance of β_1 .

Furthermore, as we will see in the next paragraph, the OLS residuals will underestimate the variance of the error term (i.e., σ^2 will be underestimated from the

OLS expression $\sum (e_i)^2 / (n-2)$ and this will be an additional source of underestimation of the true variance of the parameter estimate β_1 .

3. The variance of the random term may be seriously underestimated if the u 's are autocorrelated. This can be another source of the underestimation of the true variance of the parameter estimate, β_1 . In particular, the underestimation of the

variance of u will be more serious in the case of positive autocorrelation of the error term (u_t) and of positively autocorrelated values of X (in successive time periods).

To see how the underestimation of $\text{var}(u)$ occurs, let us establish that if there is no autocorrelation

$$\text{Cov}(u_t, u_s) = 0 \quad \text{or} \quad \sum_{t \neq s} u_t u_s = 0 \quad (1.2.32)$$

If we assume a first-order autoregressive scheme, it can be shown that

$$\text{Cov}(u_t, u_s) = -2\rho \quad \text{if } |t-s|=1 \quad \text{and} \quad \text{Cov}(u_t, u_s) = 0 \quad \text{if } |t-s| > 1 \quad (1.2.33)$$

most likely underestimate because it would ignore the expression in parentheses which is almost certainly positive.

Note that if X is not auto correlated, but is approximately random, then, even if u is autocorrelated ($\rho \neq 0$), the bias in $\text{var}(u)$ and in var is not likely to be serious.

In the real world, however, the successive values of most explanatory variables (X 's) are positively autocorrelated, and this would tend to underestimation (sometimes substantial) of the variance of u and of the variance of the estimates of the β 's if OLS were applied.

4. Invalid statistical inferences. The usual t and F statistics are not reliable.

5. The formula to compute the error variance $\frac{ESS}{d.f.}$, is biased estimator of the true error variance (might underestimate it in most cases). As a consequence the estimated R^2 could be unreliable.

6. Finally, if the values of u are auto correlated, the predictions based on ordinary least squares estimates will be inefficient, in the sense that they will have a larger variance as compared with predictions based on estimates obtained from other econometric techniques. Since these variances are not minimal as compared with other techniques (such as Generalized Least Squares), the standard error of the forecast (from OLS) will not have the least value, due to auto correlated u 's.

1.2.5. Detecting Autocorrelation

From the previous discussion we understand that autocorrelation is bad which emphasize the importance of learning how to detecting it. Like that of hetroscedasticity, autocorrelation can be detected using ***visual inspection methods and formal tests.***

1.2.5.1. Visual Inspection Methods

A. Graphical method

Graphically, some clues about the possible existence of autocorrelation among the values of the error term can be obtained. Thus, this method gives some rough idea of the existence and the pattern of autocorrelation which may be gained by plotting the regression residuals either against their own lagged value(s), or against .time.

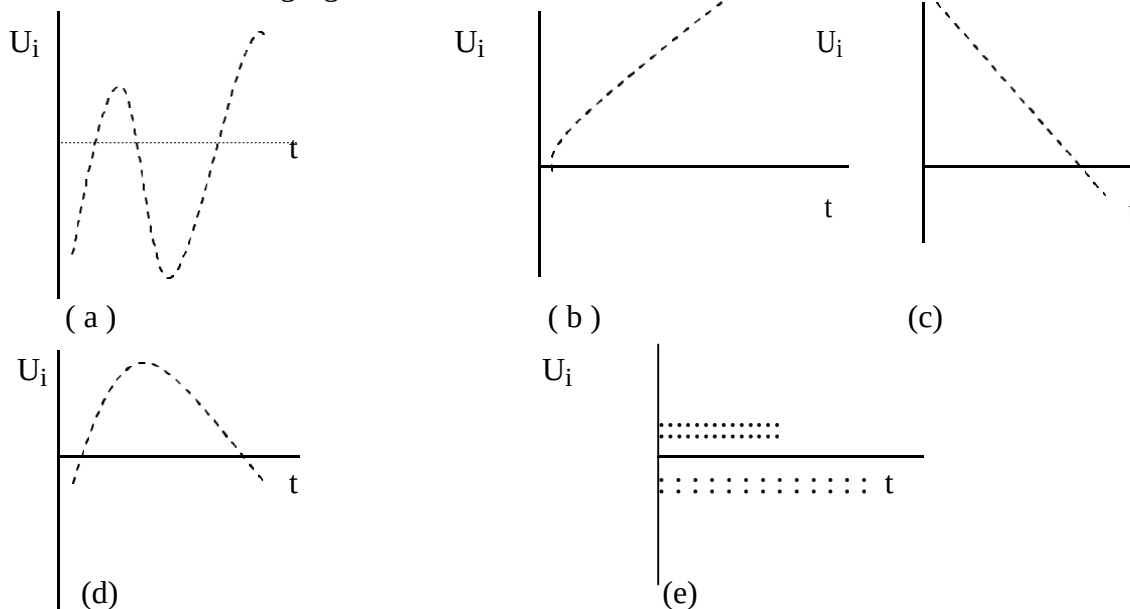
The graphs used for this purpose can be:

I. The graphs obtained by plotting the regression residuals against time:

Since autocorrelation is correlation between members of series of observations ordered in time, we will see graphically the trend of the random variable by plotting time horizontally and the random variable (U_i) vertically.

If the values of the regression residuals in successive periods show a regular pattern, we conclude that the error term is autocorrelated. Specifically, if the successive values of the regression residuals change sign frequently, then there will be negative autocorrelation. In contrary, the regression residuals do not change their sign frequently so that several positive values of them are followed by several negative values, then it can be a signal for positive autocorrelation. If they show no pattern, then it can be concluded that there is no autocorrelation.

Consider the following figures



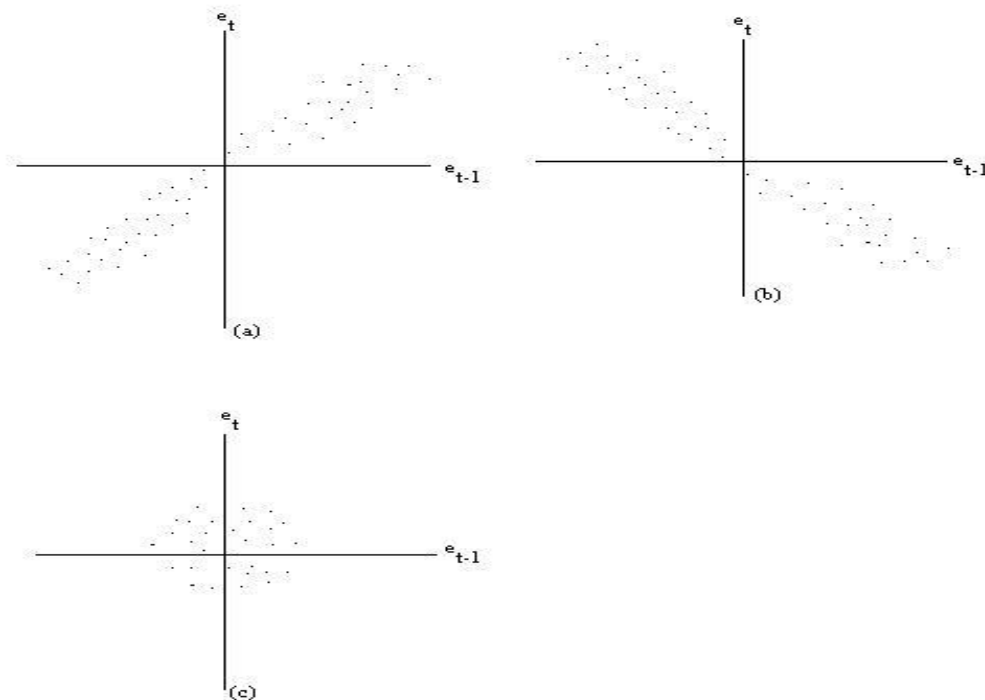
The figures (a) –(d) above, show a cyclical pattern among the U 's indicating autocorrelation i.e. figures (b) and (c) suggest an upward and downward linear trend and (d) indicates quadratic trend in the disturbance terms. Figure (e) indicates no systematic pattern supporting non-autocorrelation assumption of the classical linear regression model.

II. The graphs obtained by plotting the regression residuals against their own lagged values:

We can also show autocorrelation graphically by plotting successive values of the random disturbance term vertically (e_t) and horizontally (e_{t-1}).

In these plotted diagrams, if most of the points fall in quadrants I and III, then it shows that there is positive autocorrelation. On the other hand, if most of the points fall in quadrants II and IV, then it gives the clue that there is negative autocorrelation among the values of the error term. If there is no noticeable pattern, then we can expect that there is no autocorrelation.

In general, if the disturbance terms follow systematic pattern as in **a** (positive autocorrelation) and **b** (negative autocorrelation), there is autocorrelation or serial correlation and if there is no systematic pattern as (c), there is no autocorrelation.



B. Matrix approach for autocorrelation

The variance-covariance matrix of the error terms developed so far is also can be used as rough diagnosis of autocorrelation as of hetroscedasticity detection and is given as:

$$(UU') = \begin{pmatrix} (u_1^2) & (u_1 u_2) & \dots & (u_1 u_n) \\ (u_2 u_1) & (u_2^2) & \dots & (u_2 u_n) \\ \vdots & \vdots & \ddots & \vdots \\ (u_n u_1) & (u_n u_2) & \dots & (u_n^2) \end{pmatrix}$$

We remember that if all main diagonal elements are equal, error terms are homoscedastic. However, if all the off-diagonal elements of the variance-covariance are zero, we can judge that there is no serial correlation among the successive error terms.

$$(UU') = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} = I_n \quad (1.2.34)$$

Where I_n is an identity matrix of $n \times n$ dimensions.

1.2.5.2. Formal Tests of Autocorrelation

We said that some rough idea of the existence and the pattern of autocorrelation may be gained by plotting the regression residuals either against their own lagged value(s), or against time and also by using Variance-covariance matrix approach.

Generally, these methods give rough idea about the existence and pattern of autocorrelation, which makes the use of other methods compulsory.

This method is called formal because the tests are based on the formal testing procedure you have discussed in your statistics for economists or Econometrics I courses. It is based on either the z-test, t-test, F-test or X^2 test.

Thus, in this module we deal with some of the more accurate tests for the incidence of autocorrelation.

The traditionally applied tests are the von Neumann ratio and the Durbin-Watson test.

A. The Von Neumann Ratio

This is the ratio of the variance of the first differences of any variable X over the variance of X . Mathematically this ratio is given as:

$$VNR = \frac{\sum (x_t - x_{t-1})^2}{\sum x_t^2} \quad (1.2.35)$$

The Von Neumann ratio is applicable for directly observed series and for variables which are random, that is, variables whose successive values are not auto correlated.

In the case of the u 's, their values are not directly observable but are estimated from the OLS residuals (e 's) and for the large samples ($n > 30$) the Von Neumann ratio which is given in the following form could be applied approximately (with ≈ 0 by definition), as:

$$VNR = \frac{\sum (e_t - e_{t-1})^2}{\sum e_t^2} \quad (1.2.36)$$

However, this is not possible because the values of the OLS residuals (e 's) are not independently distributed, even if the population u_t 's are independently distributed (not serially correlated). Hence this test is not applicable for testing the autocorrelation of the u 's, especially if the sample is small ($n < 30$).

Thus, this method is less applicable in practice; thus, there is no need of investing much time here, rather we have to search for other reasonable tests. Among others, the most celebrated test for detecting serial correlation is the one that is developed by statisticians Durbin and Watson.

B. The Durbin-Watson d test:

Durbin and Watson have suggested the most commonly used method that generally used to detect first-order autocorrelation or first-order autoregressive scheme ($U_t = \rho U_{t-1} + V_t$). Though the test is appropriate for first order autoregressive scheme, it is applicable to small as well as large samples. Since first order autocorrelation is most likely to appear in time series data, the test is very relevant, and all statistical software has the option of calculating it automatically for you.

It is important to note the assumptions underlying the d-statistics

1. The regression model includes an intercept term. If such term is not present, as in the case of the regression through the origin, it is essential to rerun the regression including the intercept term to obtain the Residual Sum Squares (RSS).
2. The explanatory variables (X 's), are non-stochastic, or fixed in repeated sampling.
3. The disturbances U_t are generated by the first order autoregressive scheme:

$$U_t = \rho U_{t-1} + V_t$$

4. The regression model does not include lagged value of Y the dependent variable as one of the explanatory variables. Thus, the test is inapplicable to models of the type

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 X_{2t} + \alpha_3 X_{3t} + \dots + \alpha_k X_{kt} + \epsilon_t$$

Where y_{t-1} the one period lagged value of y is such models are known as *autoregressive models*. If d-test is applied mistakenly, the value of d in such cases will often be around 2, which leads to wrong conclusion of the absence of first order autocorrelation. Durbin developed the so-called *h-statistic* to test serial correlation in such autoregressive.

5. There are no missing observations in the data.

In using the Durbin –Watson test, it is, therefore, to note that it cannot be applied in violation of any of the above **five** assumptions.

The test involves the calculation of a test statistic based on the residuals from the OLS regression procedure.

Formally, the test may be outlined as follows:

The null hypothesis is: $H_0: \rho = 0$ (the U 's are not autocorrelated with a first-order scheme)

This hypothesis is tested against the alternative hypothesis:

$H_1: \rho \neq 0$ (the U 's are autocorrelated with a first-order scheme).

To test the null hypothesis we use the Durbin-Watson (DW) d statistic which is defined as:

$$d = \frac{\sum_{t=2}^T (e_t - e_{t-1})^2}{\sum_{t=1}^T e_t^2} \quad (1.2.37)$$

with e being the estimated residual from a sample regression model.

This statistic is calculated by summing the difference in the error terms separated by one period and dividing by the squared error term.

Note that, in the numerator of d statistic the number of observations is $n - 1$ because one observation is lost in taking successive differences.

The relationship between the Durbin-Watson d statistic and the Von Neumann ratio

The Durbin-Watson d statistic is related to the Von Neumann ratio by the relationship

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2} \quad (1.2.38)$$

Because from the Von Neumann ratio we obtain

$$\frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2} = d \quad (1.2.39)$$

For large *sample size* we have $n \approx n - 1$, and thus d tends to the Von Neumann ratio asymptotically.

To see that this test statistic is related to the first order autocorrelation case we may rewrite (1.2.37) in the form in (1.2.38). This equation is obtained by taking the following facts in to considerations: After Expansion of equation 1.2.37 we supposed that for large

$$\sum_{t=2}^n e_t^2 \approx \sum_{t=1}^n e_t^2, \text{ because in the second two cases only one observation}$$

is lost or they differ in only one observation, thus, they are approximately equal and also

we proved that $\frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2} = d$. Thus equation 1.2.37 yields us:

$$d = \frac{\sum_{t=2}^T e_t^2}{\sum_{t=1}^T e_t^2} - \frac{\sum_{t=2}^T e_t e_{t-1}}{\sum_{t=1}^T e_t^2} \frac{\sum_{t=2}^T e_{t-1}^2}{\sum_{t=1}^T e_{t-1}^2} \quad (1.2.40)$$

Where, $\hat{\rho}$ on the right hand side is the autocorrelation coefficient from the first order auto regression scheme. However, it is only an approximation since the expressions in the numerator sum from 2 to T instead of 1 to T as is the case in the denominator. The larger the value of T the better is the approximation.

From (1.2.38) it is possible to see that the DW test statistic (d) only takes values between 0 and 4 since the autocorrelation coefficient only takes values between -1 and 1. Hence we have different cases:

1. If $\hat{\rho} = 0$ and $d = 2$, there is no autocorrelation. Thus if from the sample data we find

$d \approx 2$, we accept that there is no first-order autocorrelation, either positive or

negative, in the function.

2. If $\hat{\rho} = +1$, $d = 0$ and we have perfect positive autocorrelation. Therefore, in the range of $0 < d < 2$, there is some degree of positive autocorrelation, which is stronger the closer d is to zero.
3. If $\hat{\rho} = -1$, $d = 4$ and we have perfect negative autocorrelation. Therefore if $2 < d < 4$, there is some degree of negative autocorrelation, which is stronger the higher the value of d .

General speaking, whenever, therefore, the calculated value of d turns out to be sufficiently close to 2, we accept null hypothesis, and if it is close to zero or four, we reject the null hypothesis that there is no autocorrelation.

NB: It should be clear that in the Durbin-Watson test the null hypothesis of zero autocorrelation ($\rho = 0$) is carried out indirectly, by testing the equivalent hypothesis $d = 2$. Thus, the test can also be stated in the form:

The null hypothesis is: $H_0: \rho = 0$ ($d = 2$) $\rightarrow$ the U 's are not autocorrelated with a first-order scheme

This hypothesis is tested against the alternative hypothesis:

$H_1: \rho \neq 0$ ($d \neq 2$) (the U 's are autocorrelated with a first-order scheme).

The next step is to use the sample residuals (e_t 's) and compute the empirical value of the Durbin-Watson statistic, d^* . Finally, the empirical d^* must be compared with the theoretical values of d ; that is, the values of d which define the critical region of the test. The problem with this test is that the exact distribution of d is not known: However, Durbin and Watson have established upper (d_u) and lower (d_L), limits for the significance levels of d which are appropriate to test the hypothesis of zero first-order autocorrelation against the alternative hypothesis of *positive* first-order autocorrelation; as we mentioned previously, positive autocorrelation is the severe problem of macroeconomic variables.

Durbin and Watson have tabulated these upper and lower values at the 5 per cent and the 1 per cent level of significance (**see appendices pages**). The tables assume that the u 's are normal, homoscedastic, and not autocorrelated, and that the X 's are truly exogenous. As usual n is the sample size, while k' is the number of exogenous explanatory variables in the equation being estimated.

The test itself compares the empirical d^* value, calculated from the regression residuals, with the d_L and d_u in the Durbin-Watson tables, and with their trans- forms $(4 - d_L)$ and $(4 - d_u)$.

The comparison using d_L and d_U investigates the possibility of *positive* autocorrelation, which comparison with $(4 - d_L)$ and $(4 - d_u)$ investigates the possibility of *negative* autocorrelation:

1. If $d^* < d_L$, we reject the null hypothesis of no autocorrelation and accept that there is positive autocorrelation of the first order.
2. If $d^* > (4 - d_L)$, we reject the null hypothesis of no autocorrelation and accept that there is *negative* autocorrelation of the first order.
3. If $d_u < d^* < (4 - d_u)$, we accept the null hypothesis of no autocorrelation.
4. If $d_L < d^* < d_u$ or if $(4 - d_u) < d^* < (4 - d_L)$ the test is inconclusive. It means in these regions, one cannot make any decision.

The mechanisms of the D.W test are as follows, assuming that the assumptions underlying the tests are fulfilled.

Step 1: Run the OLS regression and obtain the residuals

Step 2: Obtain the computed value of d using the formula given in equation 1.2.37

Step 3: For the given sample size and given number of explanatory variables, find out critical d_L and d_U values.

Step 4: Now follow the decision rules given below.

Null hypothesis	Decision	If
No positive autocorrelation	Reject	$0 < d < d_L$
No positive autocorrelation	No decision	$d_L < d < d_U$
No negative autocorrelation	Reject	$4 - d_L < d < 4$
No negative autocorrelation	No decision	$4 - d_U < d < 4 - d_L$
No autocorrelation, positive or negative	Accept	$d_U < d < 4 - d_U$

Example 1.2.1

Suppose for a hypothetical model: $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + U_t$ and assume that you have a time series with 150 observations, and two explanatory variables that will be used to explain the dependent variable. Running the regression you received a DW d statistic value equal to 1.63. In the table with critical values for the Durbin Watson test at 5% significance level, you found that $L=1.71$ and $U=1.76$.

Test the null hypothesis of no autocorrelation.

Since the test value is outside the inconclusive interval and below the lower value we have to draw the conclusion that our model suffer from positive autocorrelation.

Example 1.2.2

Given the same set up as in the previous example but with different DW-test value equal to 2.35, what can you say about autocorrelation?

In the table we found the value for $d_L = 1.71$ and the value for $d_U = 1.76$. Using these values we can calculate the inconclusive interval related to DW-values larger than 2. $(4 - d_U) = (4 - 1.76) = 2.24$ and $(4 - d_L) = (4 - 1.71) = 2.29$.

Since the test value of 2.35 is outside the interval and larger than the upper value of 2.29 we must conclude that our model suffer from negative autocorrelation.

Activity 1.1.2

Suppose for a hypothetical model: $Y_t = \beta_0 + \beta_1 X_t + U_t$, if we found

$$d = 0.1380; \quad d_L = 1.37 \text{ and } d_U = 1.50$$

Based on the above values test for autocorrelation

Short comings of the Durbin- Watson test

The Durbin-Watson test has several shortcomings.

Firstly, the d statistic is not an appropriate measure of autocorrelation if among the explanatory variables there are lagged values of the endogenous variable.

Secondly, the range of values of d over which the Durbin-Watson test is inconclusive ($d_L < d^* < d_U$ and $(4 - d_U) < d^* < (4 - d_L)$) has been a drawback to its application.

Thirdly, the Durbin-Watson test is inappropriate for testing for higher order serial correlation or for other forms of autocorrelation (for example for nonlinear forms of serial dependence of the values of U_t).

C. Alternative tests for autocorrelation

Various writers including Durbin himself have suggested alternative tests for serial correlation, which are more accurate and more power-full than the Durbin-Watson test. However, these tests are invariably more complicated and costly in computations. Furthermore, some of them are based on stronger assumptions. Given the shortcomings of the alternative tests several econometricians have followed the practice of applying the Durbin-Watson test in the following amended form:

Reject the null hypothesis ($H_0: \rho = 0$) if $d^* < du$ or $d^* > (4 - du)$

or Accept the null hypothesis if $du < d^* < (4 - du)$.

In the amended test the rejection (critical) region includes not only the values of $d < d_L$ and $d > (4 - d_U)$ but also the inconclusive values in the original Durbin-Watson test.

The above amendment of the original Durbin-Watson test is inaccurate, because the levels of significance of the original test are certainly affected by extending the rejection region over the range of the inconclusive d values. The amendment, however, may be justified on the grounds of the seriousness of the effects of autocorrelation.

Since the consequences of incorrectly accepting the null is so much more serious than those of incorrectly rejecting it, one might have a preference for rejecting the null hypothesis in cases of doubt.

I. The Durbin h test statistic

As been described above, the DW-test is made for the purpose of resting for first order autocorrelation. Furthermore, it assumes that none of the explanatory variables tire lagged dependent variables which would be the case when estimating a dynamic model. When that is the Case the *DW-test* has a tendency to be close to 2 even though the error terms are serially correlated; hence, increasing the risk of a type II error (wrongly accepting the hypothesis to be rejected). Hence, the DW test should not be used with the following kind of model:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 X_t + U_t \quad (1.2.41)$$

Fortunately there is an easy alternative to the *DW-test* that could be seen as a modified version of it and for that reason is called the *Durbin's h statistic*. However, this test is based on Z-score (standard normal table value). It is defined in the following way:

$$h = \frac{d}{2} \sqrt{\frac{n}{1 - \text{Var}(\beta_1)}} \quad (1.2.41)$$

Alternatively,

$$h = \frac{d}{2} \sqrt{\frac{n}{1 - \text{Var}(\beta_1)}} \quad (1.2.42)$$

Where: $\hat{\rho}$ is an estimate of ρ in the first autoregressive process ($AR(1)$), d is DW statistic, n is the number of observations in the regression and $\text{Var}(\beta_1) = S^2_{\beta_1 Y_{t-1}}$ is the square of the standard error of the estimated parameter for the lagged dependent variable. The test statistic has been shown to be standard normally distributed under the null hypothesis of no autocorrelation, which means that the test value should be compared with a critical value from the standard normal table.

The presence of autocorrelation in models that include lagged dependent variables is even more affected than the standard model. When the error term is serially correlated in a dynamic model, the estimated parameters are *biased and inconsistent*. It is therefore very important to correct for the problem before using the estimates for anything.

Example 1.2.3

Assume that we have estimated the parameters of a dynamic model and obtained the following results with standard errors within parenthesis:

$$Y_t = 1.324 + 0.789Y_{t-1} + 0.821X_t \quad d=1.529$$

(0.321) (0.043) (0.032)

We use quarterly data over a 30 years period, which means that $n=120$.

Test for serial correlation of error terms.

Since our model includes a lagged variable we lose one observation in the estimation.

Using the information from the regression, we may form the Durbins h statistic:

$$h_1 = \frac{1.529}{2} \sqrt{\frac{119}{1 - 119 \times 0.043^2}} = 2.925$$

Using a one sided test at the 5 percent significance level we obtain a critical value of 1.645. Since the test value is much larger than the critical value, we reject the null hypothesis and conclude that our error terms are serially correlated.

NB: If the sample size is large, we can use DW statistic as: $h = \sqrt{(1 - 1/2 d)} \sim N(0, 1)$. A common problem with this test is that the h-statistic cannot be computed if $S^2_{\beta_1 Y_{t-1}}$ is greater than 1, which can happen if the sample size is not very large in the first formula d- assumes the regressors (independent variables) are non-stochastic, if this is not the case, then d-test is not valid; hence, h-statistic too. And also this test cannot be applied for higher order autoregressive schemes like that of original DW test. In this case Durbin suggests the following test.

II. Durbin's Alternative Test for Autocorrelation

This test has the advantage that it is applicable to any form of autocorrelation and it provides estimates of the coefficients of the autocorrelation relationship; which are required for the remedial transformations of the original observations. To apply this test we can follow the following steps:

Let's assume that the original regression takes the form in equation (1.2.41):

Step 1: We first apply OLS to the sample observations or regress Y 's on X 's and obtain the values of the regression residuals, e_t 's.

Step 2: Regress e_t on e_{t-1} and X_t

Since we cannot be *a priori* certain about the existence of autocorrelation or about its pattern, we may experiment with various forms of autoregressive structures, for example:

Or $e_t = \alpha_1 e_{t-1} + \alpha_2 e_{t-2} + \dots + \alpha_k e_{t-k} + \epsilon_t$

Or $e_t = \alpha_1 e_{t-1} + \alpha_2 e_{t-2} + \dots + \alpha_k e_{t-k} + \alpha_{k+1} X_t + \epsilon_t$

Or $e_t = \alpha_1 e_{t-1} + \alpha_2 e_{t-2} + \dots + \alpha_k e_{t-k} + \alpha_{k+1} X_t + \alpha_{k+2} X_{t-1} + \dots + \alpha_{k+m} X_{t-m} + \epsilon_t$ and so on.

Step 3: Making Decision

Autocorrelation is judged in the light of the statistical significance of the ρ and the overall fit of the above regression. That is, we may carry out anyone of the standard tests of statistical significance for the estimates ρ of the autocorrelation relationship (e.g. $t = \rho / \text{se}(\rho)$) as well as an F test for the overall significance of the regression.

Decision rule:

If the ρ 's are found statistically significant we accept that the u 's are autocorrelated, provided that the overall correlation coefficient (for the autoregressive relationship) is statistically significant and reasonably large. This test does not have the power properties of the Durbin-Watson test in finite samples, but is asymptotically equivalent.

Thus, the Durbin's second test is powerful and can be generalized to higher order autoregressions. However, it is not often used because it involves more complications.

D. The LM-test: A General Test for Higher Order Serial Correlation

The Lagrangian Multiplier (LM) test, which is derived from a general principle called the Lagrangian Multiplier (LM) principle is a more general test of autocorrelation compared to the two previous tests. This test in addition allows us for a test of autocorrelation of higher order than one, and can be used even though lagged dependent variables are included in the model. In principle, the test is similar to Durbin's second test; however, the LM test is a large sample test which means that it should be treated as an approximation when using small samples, compared to the DW-test that could be seen as an exact test.

The LM test is executed with the following steps:

Step 1: Estimate the parameters of your main model:

$$Y_t = \beta_0 + \beta_1 X_t + U_t \quad (1.2.43)$$

Step 2: Create the residual term using the estimated parameters and lag it.

Step 3: Extend your original model by including the estimated lagged residual in the specification:

$$Y_t = \beta_0 + \beta_1 X_t + \rho U_{t-1} + U_t \quad (1.2.44)$$

Where, U_t is white noise error.

Step 4: Test the null hypothesis $H_0: \rho = 0$ using a simple t-test or standard error test. If you are able to reject the null hypothesis (if $t_{cal} > t_{tab}$ or if $Se(\rho) < \alpha/2$), you can conclude that you have autocorrelation. If the estimated lag residuals are more than one, you can use joint F-test to test simultaneously. If at least one ρ is statistically significant (different from zero), you can reject null hypothesis of no serial correlation. Thus, the equation given by (1.2.44) can be extended to include more lags of the residual terms in order to test for higher order of autocorrelation.

Example 1.2.4

Assume that we have a time series model with two explanatory variables, and we suspect that the error term might be serially correlated of the second order. $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + U_t$ (*)

Since the error term is unobserved we need to estimate the residuals for the model using the estimated parameters of the model:

$$e_t = Y_t - \hat{\beta}_0 - \hat{\beta}_1 X_{1t} - \hat{\beta}_2 X_{2t} \quad (**)$$

Lag the estimated residuals from (**), re-specify equation (*) and obtain:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \rho_1 e_{t-1} + \rho_2 e_{t-2} + V_t \quad (***)$$

We estimated the parameters of the extended version of the model given by (***) and obtained the following estimates with standard errors within parenthesis:

$$\hat{Y}_t = 3.241 + 0.378X_{1t} + 0.562X_{2t} + 0.324e_{t-1} + 0.124e_{t-2} \\ (1.301)(0.091) \quad (0.192) \quad (0.029) \quad (0.101) \quad (****)$$

By investigating the significance of the coefficients of the two residual terms, we see that the first one is significantly different from zero, while the other is not. We therefore conclude that the residual term is serially correlated of the first order only.

1.2.6. Solutions for the problems of Autocorrelation

Once we have found that our error term is serially correlated we need to correct for it before we can make any statistical inference on the population. Since in the presence of serial correlation the OLS estimators are inefficient, it is essential to seek remedial measures. The remedy however depends on what knowledge one has about the nature of interdependence among the disturbances. This means the remedy depends on whether the coefficient of autocorrelation is known or not known.

As for the case of heteroskedasticity, we need to transform the involved variables and therefore use generalized least square. The solution to be adopted in each particular case depends on the source of autocorrelation. Thus if the source is omitted variables, the appropriate procedure is to include these variables in the set of explanatory variables. We may illustrate the occurrence of quasi-autocorrelation with the 'ratchet effect' in a consumption function. Consumption in period t depends not only on current income (Y_t), but also on the level of income in previous periods. In its simplest form this type of consumption function may be written as follows:

$$C_t = \beta_0 + \beta_1 Y_t + \beta_2 Y_{t-1} + U_t \quad (1.2.45)$$

If lagged income is omitted, its influence will be reflected in the random term U_t , (and probably in the coefficient of current income, which will be *biased*). Since the successive values of income are usually positively autocorrelated, omission of Y_{t-1} will result in a pattern of quasi-autocorrelated U 's.

Autocorrelation would be eliminated in this case by introducing into the function explicitly lagged income as an explanatory variable of current consumption.

The simplest way to detect whether autocorrelation is due to omitted variables is to regress the residuals, e 's, against variables which on *a priori* grounds might be relevant explanatory variables of the phenomenon being studied. Similarly, if the source of autocorrelation is the misspecification of the mathematical form of the relationship, the relevant approach is to change the initial (linear) form. This can be investigated by regressing the residuals against higher powers of the explanatory variable(s), or by computing a 'log linear' form and re-examining the resulting new residuals, etc.

Only when the above sources of autocorrelation have been ruled out should we accept that the true u 's are temporally dependent.

In the case of *true autocorrelation*, the appropriate procedure is the transformation of the original data so as to produce a model whose random variable satisfies the assumptions of classical least squares, and consequently the parameters can be optimally estimated with this method. We will illustrate the procedure starting from the simple first-order autoregressive scheme.

Once autocorrelation is detected by applying any relevant test, the appropriate corrective procedure is to obtain an estimate of the ρ 's and apply OLS to a set of transformed data. The transformation of the original data depends on the pattern of the autoregressive structure.

A. First-order autoregressive scheme

If autocorrelation is of the first-order scheme

$$U_t = \rho U_{t-1} + V_t$$

the appropriate transformation is to subtract from the original observations of each period the product of *times* the value of the variables in the previous period. That is we apply OLS to the model:

$$Y_t^* = \beta_0 + \beta_1 X_{1t}^* + \dots + \beta_k X_{kt}^* + V_t \quad (1.2.46)$$

Where

$$Y_t^* = Y_t - \rho Y_{t-1}$$

$$X_{jt}^* = X_{jt} - \rho X_{j(t-1)} \quad (j=1, 2, \dots, k)$$

$$V_t = U_t - \rho U_{t-1}$$

Rationalization of the transformation procedure

The above transformation is obtained as follows:

The original model is:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \dots + \beta_k X_{kt} + U_t \quad (1.2.47)$$

Where, $U_t = \rho U_{t-1} + V_t$ with V_t satisfying all the assumptions of a random variable.

In the same manner, the relationship for the period $t-1$ is

$$Y_{t-1} = \beta_0 + \beta_1 X_{1(t-1)} + \beta_2 X_{2(t-1)} + \dots + \beta_k X_{k(t-1)} + U_{t-1}$$

Pre-multiplying this equation by ρ we get

$$\rho Y_{t-1} = \rho \beta_0 + \rho \beta_1 X_{1(t-1)} + \rho \beta_2 X_{2(t-1)} + \dots + \rho \beta_k X_{k(t-1)} + \rho U_{t-1}$$

Subtracting the above equation from the original relationship we obtain:

$$(Y_t - \rho Y_{t-1}) = \beta_0(1 - \rho) + \beta_1[X_{1(t)} - \rho X_{1(t-1)}] + \dots + \beta_k[X_{k(t)} - \rho X_{k(t-1)}] + (U_t - \rho U_{t-1}). \quad (1.2.48)$$

The new error term $U_t - \rho U_{t-1}$ is equal to V_t and uncorrelated random variable.

Consequently, if we know ρ we can apply ordinary least squares to the transformed relationship in equation (1.2.46).

Note that with the above transformation we are able to retain in our analysis only $n - 1$ observations since we lose one observation in the process. To avoid this loss, K. R. Kadiyala has suggested the following transformation of the first observation

$$Y_1^* = Y_1 \quad \text{---} \quad 1 - \rho$$

$$X_{j1}^* = X_{j1} \quad \text{---} \quad 1 - \rho \quad (j = 1, 2, \dots, k)$$

B. Higher order autoregressive schemes

The Durbin and Watson test is not intended to be used for testing for higher order lag structures of the disturbance term. However, such structures can be tested by conducting the traditional procedures (t, standard error, F) or tests of significance for the ρ 's of the higher-order autoregressive model(s).

$$e_t = \rho_1 e_{t-1} + \rho_2 e_{t-2} + \rho_3 e_{t-3} + \dots \quad (1.2.49)$$

Once a higher order autoregressive structure has been found to exist in the function, we proceed in the transformation of the original data and the original relationship in exactly the same way as before. For example assume that a second-order autoregressive scheme has been detected in the function. The original model is in equation 1.2.47:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \dots + \beta_k X_{kt} + U_t$$

Where: $U_t = \rho_1 U_{t-1} + \rho_2 U_{t-2} + w_t$, with w_t meeting all the assumptions of a random variable (homoscedastic, non- autocorrelated, etc.).

The relationships for the periods $(t - 1)$ and $(t - 2)$ are

$$Y_{t-1} = \beta_0 + \beta_1 X_{1(t-1)} + \dots + \beta_k X_{k(t-1)} + U_{t-1}$$

$$Y_{t-2} = \beta_0 + \beta_1 X_{1(t-2)} + \dots + \beta_k X_{k(t-2)} + U_{t-2}$$

Pre-multiplying these equations with ρ_1 and ρ_2 respectively we have

$$\rho_1 Y_{t-1} = \rho_1 \beta_0 + \beta_1 \rho_1 X_{1(t-1)} + \beta_2 \rho_1 X_{2(t-1)} + \dots + \beta_k \rho_1 X_{k(t-1)} + \rho_1 U_{t-1}$$

$$\rho_2 Y_{t-2} = \rho_2 \beta_0 + \beta_1 \rho_2 X_{1(t-2)} + \beta_2 \rho_2 X_{2(t-2)} + \dots + \beta_k \rho_2 X_{k(t-2)} + \rho_2 U_{t-2}$$

Subtracting these equations from the original model we find

$$\begin{aligned} [Y_t - \rho_1 Y_{t-1} - \rho_2 Y_{t-2}] &= \beta_0 (1 - \rho_1 - \rho_2) + \beta_1 [X_{1t} - \rho_1 X_{1(t-1)} - \rho_2 X_{1(t-2)}] + \dots \\ &+ [U_t - \rho_1 U_{t-1} - \rho_2 U_{t-2}] \end{aligned} \quad (1.2.50)$$

The new disturbance term $(U_t - \rho_1 U_{t-1} - \rho_2 U_{t-2})$ is equal to w_t and hence is by assumption serially independent. Therefore if we know ρ_1 and ρ_2 (the coefficients of the second-order scheme), we can apply ordinary least squares to the transformed relationship.

NB: The extension of this procedure to higher-order lag structures of the U term is straightforward.

We see that if we know the values of ρ 's we can proceed with the above transformation of the original data and then apply ordinary least squares to the transformed relationship. However, ρ is **unknown** and we must estimate its value in some way or another. In the next section we will examine various methods for obtaining an estimate of the autocorrelation coefficient(s).

1.2.7. Methods for Estimating the Autocorrelation Parameters

A. Method I: *A priori* information on ρ (or ρ 's)

In most applied econometric research, when autocorrelation is suspected or established on the basis of a formal test, the investigator makes some 'reasonable' guess about the value of the autoregressive coefficient(s) ρ , using his knowledge and intuition about the relationship being studied.

A usual case is to assume that $\rho = 1$. Under this assumption the appropriate transformation is to take the first differences of the original data and apply ordinary least squares to the transformed model

$$(Y_t - Y_{t-1}) = \beta_1(X_t - X_{t-1}) + V_t \quad (1.2.51)$$

Where, $V_t = U_t - U_{t-1}$.

The original relationship is

$$Y_t = \beta_0 + \beta_1 X_t + U_t \quad (1.2.52)$$

Where, $U_t = \rho U_{t-1} + V_t$ with V_t satisfying all the usual assumptions of a random variable.

Assuming $\rho = 1$

$$U_t = U_{t-1} + V_t$$

or

$$U_t - U_{t-1} = V_t$$

Now, lagging the original equation by one period and pre-multiplying by ρ we obtain

$$\rho Y_{t-1} = \rho \beta_0 + \beta_1 \rho X_{t-1} + \rho U_{t-1}$$

$$Y_{t-1} = \beta_0 + \beta_1 X_{t-1} + U_{t-1} \text{ (given that } \rho = 1 \text{)}$$

Subtracting from the original equation we have

$$Y_t - Y_{t-1} = \beta_1 (X_t - X_{t-1}) + (U_t - U_{t-1}) \quad (1.2.53)$$

Where, $U_t - U_{t-1} = V_t$, serially independent by assumption.

Note that for the computation of this function we must suppress (exclude) the constant intercept. Otherwise the constant intercept that will appear in the function would imply that time is implicitly introduced in the function as an explanatory variable. Of course there is nothing *a priori* wrong in incorporating time in the function; a time factor has the meaning that there is an autonomous growth of the dependent variable over time.

If this hypothesis is justified in the particular phenomenon being studied, working with first differences is the appropriate procedure to adopt. On the other hand, if the relationship being studied is incompatible with the hypothesis of an autonomous growth, the first-differences transformation is not appropriate.

It is obvious that when one assumes $\rho = 1$ and then takes the first differences of the variables as the appropriate transformation, the computational procedure is greatly simplified. This is the reason why first differences are popular in applied research whenever autocorrelation is present in the original model.

However, the assumption $\rho = 1$ is the extreme opposite of Assumption of $\rho = 0$ the linear regression model, and the true ρ may actually lie somewhere between these two extreme values ($0 < \rho < 1$).

B. Method II: *Estimation of ρ from the d statistic*

Another crude method for the estimation of the coefficient of the auto-regressive scheme is to solve for ρ the expression.

$$d \cong 2(1 - \rho)$$

From the application of the Durbin-Watson test we obtain d^* which we may substitute in the above expression and get

$$\rho = 1 - \frac{d^*}{2} \quad (1.2.54)$$

If the sample is small the estimate ρ will not be accurate, since the relationship $d \cong 2(1 + \rho)$ holds asymptotically (for large samples).

C. Method III: *The Cochrane-Orcutt Iterative Method*

This method involves a series of iterations, each of which produces a better estimate of than does the previous one. The estimated is used the generalized differencing transformation process.

It involves a gradual approximation (convergence) to the estimate of the autocorrelation coefficient, ρ . It may be outlined as follows:

Step 1: Apply OLS to the original data and obtain estimates of the coefficients

$$= + \quad (1.2.55)$$

Step 2: Compute the 'first-round' residuals

$$e_t = Y_t - \hat{Y}_t \quad (t=1, 2, \dots, n) \quad (1.2.56)$$

Step 3: From these estimate the 'first-round' estimate of ρ can be calculated using the earlier developed formula

$$\hat{\rho} = \frac{\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=2}^n e_{t-1}^2} \quad (t=2, 3, \dots, n) \quad (1.2.57)$$

Step 4: Use $\hat{\rho}$ to transform the original data and apply OLS to the model

$$[Y_t - \hat{\rho} Y_{t-1}] = \beta_0(1 - \hat{\rho}) + \beta_1[X_t - \hat{\rho} X_{t-1}] + u_t^* \quad (1.2.58)$$

Denote the estimates of this 'second-round' by $\hat{\beta}_0$ and $\hat{\beta}_1$ (where $\hat{\beta}_0$ is an estimate of the intercept $\beta_0(1 - \hat{\rho})$).

Using $\hat{\rho}$ compute the 'second-round' residuals $\hat{e}_t$

$$\hat{e}_t = Y_t - \hat{\rho} Y_{t-1} - \hat{\beta}_0 - \hat{\beta}_1(X_t - \hat{\rho} X_{t-1}) \quad (t=1, 2, 3, \dots, n)$$

and from these obtain the 'second-round' estimate of ρ (1.2.59)

$$\hat{\rho} = \frac{\sum_{t=2}^n \hat{e}_t \hat{e}_{t-1}}{\sum_{t=2}^n \hat{e}_{t-1}^2} \quad (t=2, 3, \dots, n) \quad (1.2.60)$$

Step 3: Use $\hat{\rho}$ to transform the original variables and apply OLS to the model

$$[Y_t - \hat{\rho} Y_{t-1}] = \beta_0(1 - \hat{\rho}) + \beta_1[X_t - \hat{\rho} X_{t-1}] + u_t^{**} \quad (1.2.61)$$

This iterative procedure is repeated until the value of the estimate of ρ converges.

Sometimes researchers stop at the 'second-round' estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$. This is then called a **two-stage Cochrane-Orcutt method**.

An alternative approach is to use at each step of the iteration (for the first-order autoregressive scheme) the Durbin-Watson d statistic to test the residuals for autocorrelation. If they pass the test of zero autocorrelation, the iterations stop. If not, the iterations proceed until the hypothesis of zero autocorrelation is accepted. Tests of significance of the ρ 's are conducted only at the final iteration.

The estimate of the ρ 's is more burdensome if the autoregressive scheme is more complicated (for example if we have a second-order or third-order scheme); however, the procedure is the same as the one described above for the simple first-order autoregressive structure.

For example assume a second-order autoregressive scheme

$$U_t = \rho_1 U_{t-1} + \rho_2 U_{t-2} + V_t \quad (1.2.62)$$

From the OLS residuals of the original observations we obtain the 'first-round'

estimates of the ρ 's by using least squares on the function $e_t = e_{t-1} + e_{t-2}$

The estimates are used to obtain the transformations

$$Y_t^* = (Y_t - \rho_1 Y_{t-1} - \rho_2 Y_{t-2})$$

$$X_{jt}^* = (X_{jt} - \rho_1 X_{j(t-1)} - \rho_2 X_{j(t-2)}) \quad (j = 1, 2, \dots, k)$$

We apply OLS to the function

$$Y_t^* = \beta_0^* + \beta_1^* X_{1t}^* + \dots + \beta_k^* X_{kt}^* + u_t^* \quad (1.2.63)$$

We use the new residuals u_t^* to obtain the 'second-round' estimates of the ρ 's, by applying OLS to the function $u_t^* = \rho_1 u_{t-1}^* + \rho_2 u_{t-2}^* + v_t^*$ and so on.

The iterations cease when the estimates of the ρ 's converge.

D. Method IV: Durbin's "two-step" method of estimation of ρ

Durbin has suggested the following two-step method for estimating ρ which is applicable for any order of autoregressive scheme.

Assume the original function is

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \dots + \beta_k X_{kt} + U_t \quad (1.2.64)$$

where, $U_t = [(U_{t-1}, U_{t-2}, \dots)]$. For simplicity let $U_t = \rho U_{t-1} + V_t$.

Stage 1: We start from the transformed model

$$(Y_t - \rho Y_{t-1}) = \beta_0(1 - \rho) + \beta_1(X_{1t} - \rho X_{1(t-1)}) + \dots + \beta_k(X_{kt} - \rho X_{k(t-1)}) + (U_t - \rho U_{t-1}) \quad (1.2.65)$$

We re-write this equation as follows

$$Y_t = \beta_0(1 - \rho) + \rho Y_{t-1} + \beta_1 X_{1t} - \beta_1 \rho X_{1(t-1)} + \dots + \beta_k X_{kt} - \beta_k \rho X_{k(t-1)} + V_t \quad (1.2.66)$$

Setting: $\beta_0(1 - \rho) = \alpha_0$, $\beta_1 = \alpha_1$, $\beta_1 \rho = \alpha_2$, etc.

Thus, we may write the equation in the following form

$$Y_t = \alpha_0 + \rho Y_{t-1} + \alpha_1 X_{1t} + \alpha_2 X_{1(t-1)} + \dots + V_t \quad (1.2.67)$$

Applying least squares to this equation (1.2.67) we obtain an estimate of ρ , which is the coefficient of the lagged variable Y_{t-1} .

Stage 2: We use the estimate to obtain the transformed variables

$$\begin{aligned}(Y_t - Y_{t-1}) &= Y^* \\ (X_{1t} - X_{1(t-1)}) &= X_1^* \\ &\vdots \\ (X_{kt} - X_{k(t-1)}) &= X_k^*\end{aligned}$$

which we use in order to estimate the parameters of the original relationship

$$Y_t^* = \beta_0 + \beta_1 X_1^* + \dots + \beta_k X_k^* + V_t \quad (1.2.68)$$

Durbin's method provides estimates which have optimal asymptotic properties and are more efficient for samples of all sizes.

Example 1.2.5

Given table below that includes data on imports and the gross national product of the U.K (Cited in Koutsoyiannis):

Imports and gross national product of the U.K					
Year t	Imports Z_t	GNP X_t	<i>Estimated</i>		
			<i>imports</i> Z_t	<i>Error</i> $et = Z_t - Z_{t-1}$	$et - et-1$
1950	3,748	21,777	3,626	122	
1951	4,010	22,418	3,805	205	83
1952	3,711	22,308	3,774	-63	-268
1953	4,004	23,319	4,057	-53	10
1954	4,151	24,180	4,298	-147	-94
1955	4,569	24,893	4,497	72	219
1956	4,582	25,310	4,613	-31	-103
1957	4,697	25,799	4,750	-53	-22
1958	4,753	25,886	4,774	-21	32
1959	5,062	26,868	5,049	13	34
1960	5,669	28,134	5,403	266	253
1961	5,628	29,091	5,670	-42	-308
1962	5,736	29,450	5,771	-35	7
1963	5,946	30,705	6,121	-175	-140
1964	6,501	32,372	6,587	-86	89
1965	6,549	33,152	6,805	-256	-170
1966	6,705	34,764	6,976	-271	15
1967	7,104	34,411	7,157	-53	218
1968	7,609	35,429	7,442	167	220
1969	8,100	36,200	7,657	443	276

Applying OLS to these observations we obtain the following imports function

$$\begin{aligned} &= -2,461 + 0.28X_t \quad r^2 = 0.983 \\ &\quad (250) \quad (0.01) \end{aligned}$$

It is found that

$$\Sigma = 573,069$$

$$(\quad - \quad)2 = 537,192$$

The Durbin-Watson d statistic from the above sample is

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2} = \frac{537,192}{573,069} = 0.937$$

From the Durbin-Watson table, with 5 per cent level of significance, $n = 20$ observations, and $k' = 1$ independent variable, the significance points of d_L and d_U are:

$$d_L = 1.20 \quad d_U = 1.41$$

Since $d^* < d_L$, we conclude there is positive autocorrelation in the imports function.

We will use two of the previously developed corrective procedures, namely the Cochrane-Orcutt Iterative Method, and Durbin's two-step method.

A. We regress e_t on e_{t-1} , and e_{t-2} , and obtain the following results

$$= 0.53 e_{t-1}$$

$$(0.26)$$

$$= 0.55 e_{t-1} - 0.21 e_{t-2}$$

$$(0.29) (0.29)$$

One could experiment with other forms of autocorrelation, but we limit our example to the first-order and second-order autoregressive schemes.

The above regressions indicate a first-order autoregression, since e_{t-1} is just significant but e_{t-2} not significant at the 5 per cent-level.

Using the estimate $\rho = 0.53$ we obtain the transformed variables

$$Y_t^* = Y_t - 0.53Y_{t-1}$$

$$X_t^* = X_t - 0.53X_{t-1}$$

Applying OLS to the above transformed data we obtain

$$\begin{matrix} (235.35) & (0.02) & R^2 = 0.949 \\ \text{***} = -1394.36 + 0.296X_t^* \end{matrix}$$

B. Using Durbin's two-step method we find $d = 1.87$

$$Y_t = \alpha_0 + \rho Y_{t-1} + \alpha_1 X_t + \alpha_2 X_{t-1} + V_t$$

$$= -1107.67 + 0.6475 Y_{t-1} + 0.3403 X_t + 0.2345 X_{t-1}$$

$$(692.8) \quad (0.29) \quad (0.10) \quad (0.13)$$

From $p^* = 0.6475$ we obtain the transformed data

$$\begin{aligned} Y_t^* &= (Y_t - p^* Y_{t-1}) = Y_t - 0.6475 Y_{t-1} \\ X_t^* &= (X_t - p^* X_{t-1}) = X_t - 0.6475 X_{t-1} \end{aligned}$$

Applying OLS to the above transformed data we find

$$Y_t^* = -372.18 + 0.228 X_t^* \quad R^2 = 0.872$$

$$(223.0) \quad (0.02) \quad d = 2.11$$

We observe that with both the transformations the value of d comes close to the 'crucial' value of 2 (which corresponds to zero autocorrelation).

1.3 Multicollinearity

1.3.1. Introduction

Dear students, so far in the previous two sub-sections we were entirely concerned with the issues of the stochastic term, U_i . However, in this particular sub-topic we will divert our discussion agenda from error terms to explanatory or independent variables.

As you remember from Econometrics I, multicollinearity is the concern of ***multivariate regression***, where there are two or more explanatory variables are introduced in to the model. It means that in univariate regression, we do not bother about multicollinearity.

The assumption of no perfect multicollinearity among the explanatory variables is a crucial condition for the application of the least squares method to estimate the parameters of a model in multivariate analysis.

The term multicollinearity is used to denote the presence of linear relationship among the explanatory variables. That means the term does not rule out non-linear relationships among regressors.

For example,

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + U_i \quad (1.3.1)$$

The introduction of X_i^2 does not violate the assumption of no-multicollinearity.

Multicollinearity can be of three types:

- i) Perfect or exact multicollinearity - when the correlation coefficient between two explanatory variables is exactly equals to 1, i.e., $r_{x_i, x_j} \neq 1$)
- ii) Imperfect multicollinearity - refers to a situation with a high correlation among the explanatory variables within a multiple regression model, but not perfect.
- iii) Zero multicollinearity – here there is no correlation between explanatory variables ($r_{x_i, x_j} = 0$).

In the first case, that is, if the correlation coefficient for these variables is equal to unity, it is impossible to obtain numerical values for each parameter separately and the method of least squares breaks down. Thus, the OLS approach cannot be used to estimate the regression coefficients because the normal equations are not solvable.

If independent variables move together, it is impossible to distinguish the **separate effects** of these variables on dependent variable, Y . To illustrate the problem let us take the following example of a relationship including three explanatory variables.

Suppose that the true consumption function for a certain country is:

$C = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$, where, C = total consumption, X_1 = income of rural areas, X_2 = income of urban areas and X_3 = tax on income.

On *a priori* grounds one should expect $\beta_1 < \beta_2$, since the marginal propensity to consume (MPC) in urban areas is in general higher than the MPC in rural areas. Assume that over the sample period X_1 and X_2 happened to be equal (that is, income is equally distributed between urban and rural areas so that $X_1 = X_2$). Under these circumstances it would be impossible to obtain a separate estimate of β_1 or β_2 , because X_1 and X_2 change together.

At the other extreme if the explanatory variables are not inter-correlated at all (that is if the correlation coefficient for these variables is equal to zero), the variables are called **orthogonal** and there are no problems concerning the estimates of the coefficients, at least so far as multicollinearity is concerned.

In practice neither of the above cases is often met. Observations are more likely to be *highly* correlated than perfectly correlated. In such cases, least-squares estimates can be obtained but are difficult to interpret.

In most cases there is some degree of intercorrelation among the explanatory variables due to the interdependence of many economic magnitudes over time. This type of correlation is called imperfect multicollinearity. In this event the simple correlation coefficient for each pair of explanatory variables will have a value between zero and unity. However, as correlation coefficient approaches unity, multicollinearity gets stronger and it impairs the accuracy and stability of the parameter estimates.

It should be noted that although multicollinearity usually tends to be more common and more serious a problem in time series, it is quite frequent in cross-section data as well.

Finally, note that multicollinearity is not a condition that either exists or does not exist in economic functions, but rather it is a phenomenon inherent in most relationships due to the nature of economic magnitudes.

NB: Dear readers, multicollinearity is like bacteria; we know that bacteria can live with all human beings all the time; however, it becomes a problem when it is **active**.

Likewise, multicollinearity is the fact of all data whether time series or cross sections, but it is a severe problem when it becomes **stronger**; thus, requires remedial measures.

1.3.2 Reasons for Multicollinearity

Multicollinearity may arise for various reasons. Among others:

1. *The tendency of economic variables to move together over time.*

Economic magnitudes are influenced by the same factors and in consequence once these determining factors become operative the economic variables show the same broad pattern of behavior over time. For example in periods of booms or rapid economic growth the basic economic magnitudes grow although some tend to lag behind others. Thus, income, consumption, savings, investment, prices, employment, etc tend to rise in periods of economic expansion and decrease in periods of recession. Growth and trend factors in time series are the most serious cause of multicollinearity.

2. *The use of lagged values of some explanatory variables as separate independent factors in the relationship.*

Models with distributed lags have given satisfactory results in many fields of applied econometrics, and their use is expanding fast. For example in consumption functions it has become customary to include among the explanatory variables past as well as the present levels of income. Naturally the successive values of a certain variable are intercorrelated, for example income in the current period is partly determined by its own value in the previous period, and so on. Thus, multicollinearity is almost certain to exist in distributed lag models (that will be discussed in chapter 3).

3. *The data collection method employed.* For instance, sampling over a limited range

4. *Model specification.*

For instance adding polynomial terms so as to capture non-linearity nature of the variable.

5. *Over-determined model.*

This happens when the model has more explanatory variables than the number of observations.

1.3.3 Consequences of Multicollinearity

The assumption of no perfect multicollinearity implies that the explanatory variables should not be perfectly linearly correlated, i.e., $r_{x_i, x_j} \neq 1$. However, this assumption may be violated in real situations; hence, the method of least squares breaks down.

Thus, the consequence of multicollinearity can be observed under two categories of multicollinearity: *Perfect multicollinearity* and *less than perfect multicollinearity*.

1. If the correlation coefficient for these variables is equal to unity (*perfect multicollinearity*), the parameter estimates are not biased (), but become:
 - a) *Indeterminate*
 - b) *Their standard errors infinitely large.*

Proof (a): Suppose that the relation to be estimated is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U_i \quad (1.3.2)$$

and that X_1 and X_2 are related with the exact relation $X_2 = kX_1$, where k is any arbitrary constant number.

The formulae for the estimation of the coefficients are

$$\begin{aligned} \beta_1 &= \frac{\sum (Y - \beta_0 - \beta_2 X_2) X_1}{\sum (X_1 - \beta_0 - \beta_2 X_2) X_1} \\ &= \frac{(\sum Y) \sum X_1 - (\sum X_1) (\sum X_2)}{(\sum X_1) \sum X_1 - (\sum X_1) (\sum X_2)} \end{aligned}$$

Substituting kX_1 for X_2 we obtain

$$\beta_1 = \frac{(\sum Y) \sum X_1 - (\sum X_1) \sum Y}{(\sum X_1) \sum X_1 - (\sum X_1) \sum Y} = - \quad (1.3.3)$$

$$\frac{(\sum Y) \sum X_1 - (\sum X_1) \sum Y}{(\sum X_1) \sum X_1 - (\sum X_1) \sum Y} = - \quad (1.3.4)$$

Therefore the parameters are indeterminate: there is no way of finding separate values of each coefficient.

Proof (b): If $r_{X_1X_2} = 1$ the standard errors of the estimates become infinitely large. In the two-variable model in equation 1.3.2; if X_1 and X_2 are perfectly correlated ($X_2 = kX_1$) the variances of

$$= \frac{\sum_{i=1}^n \frac{y_i^2}{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}}}{\sum_{i=1}^n \frac{x_i^2}{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}}} = \infty$$

Substituting kX_1 , for X_2 , we obtain

$$= \frac{\sum_{i=1}^n \frac{y_i^2}{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}}}{\sum_{i=1}^n \frac{x_i^2}{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}}} = \frac{\sum_{i=1}^n y_i^2}{\sum_{i=1}^n x_i^2} = \infty \quad (1.3.5)$$

Thus the variances of the estimates become infinite unless

$r_{X_1X_2} = 0$. However, there is no

a priori reason why $r_{X_1X_2}$ should tend to zero when inter-correlation of the explanatory variables increases.

Similarly, we can develop $\sigma_{\hat{\beta}_2}^2$, which is also infinite.

2. If multicollinearity is less than perfect (i.e. near or high multicollinearity):

I. The regression coefficients are determinate

Proof: Consider the two explanatory variables model above in deviation form.

If we assume $X_2 = kX_1$ it indicates us perfect correlation between X_1 and X_2 because the change in X_2 is completely because of the change in X_1 . Instead of exact multicollinearity, we may have:

$$X_{2i} = kX_{1i} + V_i \quad (1.3.6)$$

Where $k \neq 0$ and V_i is stochastic error term such that $\sum_{i=1}^n V_i = 0$. In this case X_2 is not only determined by X_1 , but also affected by some other variables given by V_i (stochastic error term).

Substitute $X_{2i} = kX_{1i} + V_i$ in the formula of $\hat{\beta}_1$ above

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n \frac{x_{1i} y_i - \frac{(\sum_{i=1}^n x_{1i})(\sum_{i=1}^n y_i)}{n}}{\sum_{i=1}^n x_{1i}^2 - \frac{(\sum_{i=1}^n x_{1i})^2}{n}}}{\sum_{i=1}^n \frac{x_{1i}^2}{\sum_{i=1}^n x_{1i}^2 - \frac{(\sum_{i=1}^n x_{1i})^2}{n}}} = \frac{\sum_{i=1}^n x_{1i} y_i - \frac{(\sum_{i=1}^n x_{1i})(\sum_{i=1}^n y_i)}{n}}{\sum_{i=1}^n x_{1i}^2 - \frac{(\sum_{i=1}^n x_{1i})^2}{n}} \quad (1.3.7)$$

This proves that if we have less than perfect multicollinearity the OLS coefficients are determinate.

The implication of indetermination of regression coefficients in the case of perfect multicollinearity is that it is not possible to observe the separate influence of x_1 and x_2 . But such extreme case is not very frequent in practical applications. Most data exhibit less than perfect multicollinearity.

II. OLS estimates will remain BLUE

III. Although BLUE, but the OLS estimators possess large variances; hence, large standard errors and covariances, which means the coefficients cannot be estimated with great precision.

Proof:

$$\text{var}(\hat{\beta}_1) = \frac{\sum x_2^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2}$$

Multiply the numerator and the denominator by $\frac{1}{\sum x_2^2}$

$$\text{var}(\hat{\beta}_1) = \frac{\sum x_2^2 \cdot \frac{1}{\sum x_2^2}}{\sum x_1^2 \sum x_2^2 \cdot \frac{1}{\sum x_2^2} - \frac{(\sum x_1 x_2)^2}{\sum x_2^2}} = \frac{1}{\sum x_1^2 - \frac{(\sum x_1 x_2)^2}{\sum x_2^2}} \quad (1.3.8)$$

Where r_{12}^2 is the square of correlation coefficient between x_1 and x_2 ,

If x_2 x_{1i} v_i , what happens to the variance of $\hat{\beta}_1$ as r_{12}^2 increases?

As r_{12} tends to 1 or as collinearity increases, the variance of the estimator increases and in the limit when $r_{12} = 1$ variance of $\hat{\beta}_1$ becomes infinite. Similarly, from the covariance formula we developed so far:

$$\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = \frac{r_{12}}{(1 - r_{12}^2)^{3/2}} \sqrt{\sum x_2^2 \sum x_1^2} \quad (1.3.9)$$

As r_{12} increases toward one, the covariance of the two estimators increases in absolute value.

The speed with which variances and covariance increase can be seen with the **Variance-Inflating Factor (VIF)** which is defined as:

$$VIF = \frac{1}{1 - r_{12}^2} \quad (1.3.10)$$

VIF shows how the variance of an estimator is inflated by the presence of multicollinearity. As r_{12}^2 approaches 1, the VIF approaches infinity. That is, as the extent of collinearity increase, the variance of an estimator increases and in the limit the variance becomes infinite. As can be seen, if there is no multicollinearity between x_1 and x_2 , VIF will be 1.

Using this definition we can express $\text{var}(\hat{\beta}_1)$ and $\text{var}(\hat{\beta}_2)$ in terms of VIF

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum x_1^2} VIF_1 \quad \text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_2^2} VIF_2 \quad (1.3.11)$$

Which shows that variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ are directly proportional to the VIF.

- IV. *Because of consequence III, the confidence intervals tend to be much wider, leading to the acceptance of the “zero null hypothesis” more rapidly.*
- V. *Also because of consequence III, the t -ratio of one or more coefficients tend to be statistically insignificant.*
- VI. *Although the t-ratio of one or more coefficients is statistically insignificant, the overall measure of goodness of fit, R^2 can be very high.*
- VII. *The OLS estimators and their standard errors can be sensitive to small changes in the data.*
- VIII. *Explanatory power will be unaffected, even if their effects can't be separated.*
- IX. *Wrong signs of the regression coefficients.*

Activity 1.1.3

- a. Assume that consumption expenditure is linearly related to income and wealth, after regression we obtain the following regression:

$$y_i = 24.7747 + 0.9415X_{2i} - 0.0424X_{3i}$$

$$(6.7525) \quad (0.8229) \quad (0.0807)$$

$$t = (3.6690) \quad (1.1442) \quad (-0.5261)$$

$$R^2 = 0.9635 \quad F^2 = 0.9531 \quad df = 7$$

- i. Interpret R^2
- ii. What can you say about the sign of wealth? What is its implication b. As a matter of fact, if we regress X_3 on X_2 , we obtain:

$$X_{3i} = 7.5454 + 10.1909X_{2i}$$

$$(29.4758) \quad (0.1643)$$

$$t = (0.2560) \quad (62.0405) \quad R^2 = 0.9979$$

What is the message behind the value of R^2 ?

1.3.4. Tests for detecting Multicollinearity

Dear students, so far we have studied the nature and consequences of multicollinearity. *The next question, then, is: How does one know that collinearity is present in any given situation?*

To detect whether there is collinearity in any given situation, one can use different tests, i.e., if one or more of the following conditions occur we suspect the problem of multicollinearity to be severe in the given situation.

I. A Method based on Frisch's Confluence Analysis

Frisch's Confluence Analysis estimates all possible regressions between the variables which are present in a relationship, taking each variable successively as the dependent variable and considering all possible regressions of each variable on all others which are gradually introduced into the analysis.

According to this approach, a recognizable set of symptoms for the existence of multicollinearity on which one can rely are:

- a. High coefficient of determination (R^2)
- b. High correlation coefficients among the explanatory variables ($r_{xi,xj}$'s)
- c. Large standard errors and smaller t-ratio of the regression parameters

Note: None of the symptoms by itself are a satisfactory indicator of multicollinearity.

Because:

- i. the overall R^2 may be high (relative to the $r_{xi,xj}$) and yet the results may be highly imprecise and insignificant (with wrong signs and/or large standard errors).
- ii. A high $r_{xi,xj}$ is only sufficient but not a necessary condition (adequate condition) for the existence of multicollinearity because multicollinearity can also exist even if the correlation coefficient is low.
- iii. Large standard errors may arise for various reasons and not only because of the presence so linear relationships among explanatory variables.

However, the combination of all these criteria should help the detection of multicollinearity.

II. Test Based on Auxiliary Regressions:

Since multicollinearity arises because one or more of the regressors are exact or approximately linear combinations of the other regressors, one may find out which X variable is related to other X variables to regress each X_i on the remaining X variables and compute the corresponding R^2 , which we designate as R_i^2 , each one of these regressions is called auxiliary to the main regression of Y on the X's. Then, following the relationship between F and R^2 established in chapter three of Econometrics I under overall significance, the variable,

$$R_i^2 \frac{R_{X_1, X_2, \dots, X_k}^2 / (k-2)}{1 - R_{X_1, X_2, \dots, X_k}^2 / (n-k-1)} \sim F_{(k-2, n-k-1)} \quad (1.3.12)$$

Where, n is number of observation and k is number of parameters including the intercept.

Interchangeably we can use the following formula to compute $F_{\text{calculated}}$.

$$F = \frac{R^2 / (k-1)}{1 - R^2 / (n-k)}$$

If the computed F exceeds the critical F at the chosen level of significance, it is taken to mean that the particular X_i is collinear with other X's; if it does not exceed the critical F,

we say that it is not collinear with other X 's in which case we may retain the variable in the model. If F_i is statistically significant, we will have to decide whether the particular X_i should be dropped from the model.

Note that according to Klieri's rule of thumb, which suggest that multicollinearity may be a troublesome problem only if R^2 obtained from an auxiliary regression is greater than the overall R^2 , that is obtained from the regression of Y on all regressors.

III. The Farrar-Glauber test

They developed a set of three tests statistics for testing for multicollinearity.

The first test is a Chi-Square (χ^2) test for the detection of the existence and the severity of multicollinearity in a function including several explanatory variables. The second test is an F test for locating which variables are multicollinear. The third test is a t test for finding out the pattern of multicollinearity, that is for determining which variables are responsible for the appearance of multicollinear variables.

A. χ^2 test (test of orthogonality)

This test enables the researcher to consider the whole extent and pattern of multicollinearity and hence helps to decide on the appropriate procedure for dealing with the problem.

Two variables are called **orthogonal** if $r_{x_i, x_j} = 0$, i.e., if there is no any colinearity between them. For detecting the degree of multicollinearity over the whole set of explanatory variables, Glauber and Farrar suggests χ^2 to test in the following way.

$$H_0 : x' \text{ s are orthogonal (i.e. } r_{x_i x_j} = 0)$$

$$H_1 : x' \text{ s are not orthogonal (i.e. } r_{x_i x_j} \neq 0)$$

If the computed χ^2 is greater than the critical value of χ^2 , reject H_0 in favour of multicollinearity. But if it is less, then accept H_0 .

B. t-test (test for the pattern of multicollinearity)

The t-test helps to detect those variables which are the cause of multicollinearity. This test is performed based on the partial correlation coefficients through the following procedure of hypothesis.

$$H_0 : r_{x_i x_j \cdot x_1, x_2, x_3, \dots, x_k} = 0$$

$$H_1 : r_{x_i x_j \cdot x_1, x_2, x_3, \dots, x_k} \neq 0$$

In the three variable model, we know that:

$$r_{x_1 x_2 \cdot x_3}^2 = \frac{(r_{12} - r_{13}r_{23})^2}{(1 - r_{13}^2)(1 - r_{23}^2)}, \quad r_{x_1 x_3 \cdot x_2}^2 = \frac{(r_{13} - r_{12}r_{23})^2}{(1 - r_{12}^2)(1 - r_{23}^2)} \quad \text{and} \quad r_{x_2 x_3 \cdot x_1}^2 = \frac{(r_{23} - r_{12}r_{13})^2}{(1 - r_{12}^2)(1 - r_{13}^2)}$$

$$t^* = \frac{(r_{x_i x_j \cdot x_1, x_2, x_3, \dots, x_k}^2 \sqrt{n - k})}{(1 - r_{x_i x_j \cdot x_1, x_2, x_3, \dots, x_k}^2)} \quad (1.3.13)$$

Where, n is number of observation and k is number of parameters including the intercept. If $t^* > t$ (tabulate), H_0 is rejected, but if $t^* < t$ (tabulated), H_0 is accepted, we accept X_i and X_j are not collinear since ($r_{x_i x_j \cdot x_1, x_2, x_3, \dots, x_k}$ is not significant).

IV. Test of multicollinearity using Tolerance and variance inflation factor

Dear students, in our previous section, we developed the relationship between variance of the parameter estimates and Variance Inflation Factor as:

$$\hat{\sigma}_{\beta_i}^2 = \frac{\sigma^2}{X_i' X_i} = \frac{1}{1 - R_i^2} \frac{\sigma^2}{X_i' X_i} = \frac{1}{1 - R_i^2} \text{VIF}$$

Where, R_i^2 is the R^2 in the auxiliary regression of X_j on the remaining regressors. That is, regressing one explanatory variable over all the remaining explanatory variables: suppose

$$\text{VIF}_i = \frac{1}{1 - R_i^2}, \quad \text{where } R_i^2 \text{ is the squared multiple-correlation coefficient.}$$

The squared multiple-correlation coefficient for a specific parameter is a measure of the linear strength between a variable X_1 and the rest of the variables included in the model. Assuming three variable model, the squared multiple-correlation coefficient is nothing else than the coefficient of determination received from a auxiliary regression made for each variable against the other variables in the model. That is

$$R_1^2 \text{ - is obtained from regression: } X_1 = \alpha_{10} + \alpha_{11}X_2 + \alpha_{12}X_3 + U: \quad VIF_1 = \frac{1}{1 - R_1^2} \quad (1.3.14a)$$

$$R_2^2 \text{ - is obtained from regression: } X_2 = \alpha_{20} + \alpha_{21}X_1 + \alpha_{22}X_3 + U: \quad VIF_2 = \frac{1}{1 - R_2^2} \quad (1.3.4b)$$

$$R_3^2 \text{ - is obtained from regression: } X_3 = \alpha_{30} + \alpha_{31}X_1 + \alpha_{32}X_2 + U: \quad VIF_3 = \frac{1}{1 - R_3^2} \quad (1.3.4c)$$

NB: When the model contains only two explanatory variables the squared multiple correlation will coincide with the squared bivariate correlation coefficient between the two variables in the model.

If a set of independent variables is uncorrelated, then VIF_i is equal to 1. If R_i^2 approaches to 1, both VIF and the standard deviations of the slopes(S^2) approach infinity.

The larger is the value of VIF_i , the more troublesome or collinear is the variable X_i .

However, how high should VIF be before a regressor becomes troublesome? Researchers have used $VIF_i = 10$ as a critical-value rule of thumb to determine whether too much correlation exists between the i^{th} independent variable and other independent variables.

Thus, if VIF of a variable exceeds 10 (this will happens if R_i^2 exceeds (0.9) the variable is said to be highly collinear.

Other researchers use the measure of tolerance to detect multicollinearity. It is defined as

$$TOL_i = (1 - R_i^2) = \text{---} \quad (1.3.15)$$

Clearly, $TOL_i = 1$ if X_i is not correlated with the other regressors, where as it is zero if it is perfectly related to other regressors. Thus, as a rule of thumb if Tolerance is less than 0.1, multicollinearity problem become sever.

VIF or its inverse (Tolerance) is commonly used measures of multicollinearity in applied research undertaking.

Most statistical softwares (like SPSS, STATA, etc) have routines for this, and hence you should not need to run the auxiliary regression by yourself.

For example, using one such routine in SPSS we obtained the following regression results and collinearity statistics:

	P.E (Coefficients)	S.E	VIF	Tolerance
Constant	117.085	99.782	-	-
X ₁	4.334	3.016	708.843	0.00141
X ₂	-2.857	2.582	564.343	0.00177
X ₃	-2.186	1.595	104.606	0.00956
R ²	0.801			
Test of overall significance :				
F=21.516	<0.0001			
P-value				

From the above table we see that none of the coefficients are significantly different from zero. We also observe that the coefficient of determination is above 80 percent. That implies that 80 percent of the variation in the dependent variable is explained by the three explanatory variables included in the model. Furthermore, the test of overall significance of the model is highly significant which is in line with the measure of fit.

When you are exposed to a situation like this, you must go on and calculate some multicollinearity statistics such as VIF and the tolerance for each variable.

In the table we see that VIF take large values for all variables if we compare to the case of no correlation that results in VIF=1.

From the analysis of the pair wise correlations we know that the reason for X₁ to have a relatively large VIF number is mainly due to the correlation with X₂, since the correlation with X₃ is very weak. Even so, VIF for X₃ is 104 and the corresponding tolerance is as low as 0.009 which means that X₁ and X₂ only leave 0.9 percent unique variation left of X₃'s total variation that can be used in explaining the variation in Y. The remaining variation was in common with the other two variables and hence must be disregarded. One should observe that even though the pair wise correlations are relatively low, their multiple-regression correlation is much higher, which emphasize the shortcomings of only looking at pair wise correlations.

Thus, as the value of VIF for all variables is very much higher in comparison with the critical value, multicollinearity is very severe problem of this particular regression.

However, VIF is commonly used in applied researches to detect multicollinearity, it is not free of criticism.

Firstly, as we have seen earlier, $\hat{\text{var}}(\beta_1)$, $\frac{\sigma^2}{\sum x_i^2}$ (VIF) depends on three factors: σ^2 , x_i^2 and VIF.

A high VIF can be counter balanced by low σ^2 or high x_i^2 . To put differently, a high VIF is neither necessary nor sufficient to get high variances and high standard errors. Therefore, high multicollinearity, as measured by a high VIF may not necessary cause high standard errors.

Secondly, strictly speaking, it is used to test multicollinearity for continuous variables, i.e., it should not be used for discrete or dummy variables.

V. Contingency Coefficient (CC) test

When the variables to be investigated are *discrete in nature*, Contingency Coefficient (CC) is used. It is a test based on *chi-square statistics*. It is given by the formula:

$$CC = \sqrt{\frac{\chi^2}{N}}$$
(1.3.16)

Where, N is the total sample size and χ^2 is the chi-square value.

As a rule of thumb, if CC is greater than 0.75, the variables are said to be collinear and multicollinearity seeks treatment or remedial measures.

1.3.5. Solutions for the Incidence of Multicollinearity

The solutions which may be adopted if multicollinearity exists in a function vary, depending on the severity of multicollinearity, on availability of other sources of data (larger samples, or cross-section samples, etc.), on the importance of factors which are multicollinear, on the purpose for which the function is being estimated and other considerations.

- I. Some writers have suggested that if multicollinearity does not seriously affect the estimates of the coefficients, one may tolerate its presence in the function, although the integrity of the least squares estimates is to a certain extent impaired.
- II. Others have suggested that if multicollinearity affects some of the unimportant factors one may exclude these factors from the function. Again specification error may well be expected to undermine the BLUE character of the ordinary least-squares.
- III. Multicollinearity may affect only a part of the 's, while other estimates may remain fairly stable and reliable. In this case: (a) the reliable 's may be used for any purpose, forecast or policy formulation (which requires reliable information about structural coefficients); (b) all the estimates may be used for forecasts, provided that the same multicollinearity pattern will continue to exist in the forecast period.

If, however, multicollinearity has serious effects on the coefficient estimates of important factors one should adopt one of the following *corrective* solutions.

1. *Do nothing*: Sometimes multicollinearity is not necessarily bad or unavoidable. If the R^2 of the regression exceeds the R^2 of the regression of any independent variable on other variables, there should not be much worry. Also, if the t-statistics are all greater than 2 there should not be much problem. If the estimation equation is used for prediction and the multicollinearity problem is expected to prevail in the situation to be predicted, we should not be concerned much about multicollinearity.
2. *Drop a variable(s) from the model*: This however could lead to *specification error and lose of very influential variable(s)*.

3. *Increase the size of the sample*: it is suggested that multicollinearity may be avoided or reduced if the size of the sample is increased. With increase in the size of the sample, the covariances are inversely related to the sample size. But we should remember that this will be true when intercorrelation happens to exist only in the sample but not in the population of the variables. If the variables are collinear in the population, the procedure of increasing the size of the sample will not help to reduce multicollinearity.

4. *Rethinking of the model*: revising incorrect choice of functional form, specification errors, etc... if any.

5. ***Transformation of variables to some forms*** (into natural logarithms, forming ratios, first differencing, etc...).

4. *Introduce additional equation in the model*: The problem of multicollinearity may be overcome by expressing explicitly the relationship between multicollinear variables. Such relation in a form of an equation may then be added to the original model. The addition of new equation transforms our single equation (original) model to simultaneous equation model. The reduced form method (which is usually applied for estimating simultaneous equation models) can then be applied to avoid multicollinearity.

5. *Combining cross-sectional and time series data*: A variant of the extraneous or a priori information technique is the combination of cross-sectional and time-series data, known as *pooling the data*. The other way of increasing sample size

6. *Substitution of lagged variables for other explanatory variables in distributed-lag models*

Suppose the original function for lag model may be written as

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{(t-1)} + \beta_3 X_{(t-2)} + \dots + u_t$$

For obvious reasons the successive values (X_t , X_{t-1} , X_{t-2} , etc.) of any explanatory variable X_i are often highly correlated. Multicollinearity may be avoided in this case by adopting Koyck's suggestion of substitution of the lagged values of X for a single lagged value of the dependent variable

$$Y_t = \beta_0 + \beta_1 X_t + \rho Y_{t-1} + (u_t - \rho U_{t-1})$$

In this model instead of having all the lagged values of X in the function we have only

, which in general are expected to be less correlated than the lagged values of X (For the derivation of this model see Chapter 3.).

7. Use partial correlation and stepwise regression

This involves the determination the relationship between a dependent variable and independent variable(s) by netting out the effect of other independent variable(s).

Suppose a dependent variable Y is regressed on two independent variables X_1 and X_2 .

Let's assume that the two independent variables are collinear.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$$

The partial correlation coefficient between Y and X_2 must be defined in such a way that it measures the effect of X_2 on Y which is not accounted for the other variables in the model. In the present regression equation, this is done by finding the partial correlation coefficient that is calculated by eliminating the linear effect of X_2 on Y as well as the linear effect of X_2 on X_1 and thus running the appropriate regression. The procedure can be described as follows:

Step 1: Run the regression of Y on X_2 and obtain the fitted values

$$\hat{Y}_{12} = \frac{\Lambda}{2}$$

Step 2: Run the regression of X_1 on X_2 and obtain fitted values

$$\hat{X}_{12}$$

Step 3: Remove the influence of X_2 on both Y and X_1

$$Y^* = Y - \hat{Y}_{12}, \quad X_1^* = X_1 - \hat{X}_{12}$$

Step 4: Compute the partial correlation coefficient

The partial correlation between X_1 and Y is then simple correlation between Y^* and X_1^* .

The partial correlation of Y on X_1 is represented as $r_{YX_1 \cdot X_2}$ (i.e. Controlling for X_2).

$$r_{YX_1 \cdot X_2} = \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{\sqrt{1 - r_{YX_2}^2} \sqrt{1 - r_{X_1X_2}^2}}$$

r_{YX_1} Simple correlation between Y and X_1
 r_{YX_2} Simple correlation between Y and X_2
 $r_{X_1X_2}$ Simple correlation between X_1 and X_2

Also the partial correlation of Y on X_2 keeping X_1 constant is represented as:

$$r_{YX_2 \cdot X_1} = \frac{r_{YX_2} r_{YX_1} - r_{X_1 X_2} r_{YX_1}}{\sqrt{1 - r_{X_1 X_2}^2} \sqrt{1 - r_{YX_1}^2}}$$

We can also establish relationship between the partial correlation coefficient and the multiple correlation coefficient R^2 .

$$r_{YX_2 \cdot X_1}^2 = \frac{R^2 - r_{YX_1}^2}{1 - r_{YX_1}^2} \quad \text{or} \quad R^2 = \frac{r_{YX_2}^2 - r_{YX_1}^2 r_{YX_2 \cdot X_1}^2}{1 - r_{YX_1}^2}$$

In stepwise regression procedure one adds variables to the model to maximize the adjusted coefficient of determination $\bar{R}^2$.

Chapter Summary

- A critical assumption of the classical linear regression model is that the disturbances u_i have all the same variance, σ^2 . If this assumption is not satisfied, there is heteroscedasticity of different nature.
- Heteroscedasticity does not destroy the unbiasedness and consistency properties of OLS estimators. But these estimators are no longer minimum variance or efficient. That is, they are not BLUE.
- In the presence of heteroscedasticity, the variances of OLS estimators are not provided by the usual OLS formulas. But if we persist in using the usual OLS formulas, the t and F tests based on them can be highly misleading, resulting in erroneous conclusions.
- There are several diagnostic tests available to detect heteroscedasticity: Informal methods and formal methods (statistical tests).
- The BLUE estimators are provided by the method of weighted least squares.
- Autocorrelation or serial correlation is the temporal dependence of successive values of the disturbance term U .

- Autocorrelation is a problem specific to time series data. Due to economic growth and business cycles autocorrelation is positive in most economic relationships. Autocorrelation does not arise in cross-section data (unless the sample is not random). Such data refer to a given 'point' of time; hence temporal dependence is ruled out by the nature itself of cross-section random samples.
- The most common sources of autocorrelation are: omission of some important variable, error in the mathematical form of the equation, and misspecification of the behavior of u .
- The consequences of autocorrelation are serious for the values and the standard errors of the parameter estimates. In the presence of autocorrelation:
 - i. The variance of u may be underestimated. The underestimation is serious in the case of positive autocorrelation of the u 's and of positively auto correlated X 's, which is the usual case in the real economic world.
 - ii. The variances of the parameter estimates are likely to be seriously underestimated when both the u 's and the X 's are positively auto correlated. Thus we run the danger of accepting as significant variables which in reality are not significant explanatory variables.
 - iii. The estimates 's are not best (they do not possess the minimum variance property) as compared with estimates from other econometric techniques.
 - iv. The predictions based on estimates obtained from OLS applied to a model with autocorrelated u 's are not efficient.
- If autocorrelation is due to one of the three causes mentioned above, the appropriate solution is clear: (1) inclusion of the missing important variable(s); (2) 'correction' of the mathematical form; (3) obtain an estimate of ρ (or ρ 's) by some method, transform the original data , respectively.
- The other most important assumptions of the classical linear regression model is that there is no multicollinearity among the explanatory variables, the X 's.

Broadly interpreted, multicollinearity refers to the situation where there is either an exact or approximately exact linear relationship among the X variables.

- If there is perfect collinearity among the X 's, their regression coefficients are indeterminate and their standard errors are not defined. If collinearity is high but not perfect, estimation of regression coefficients is possible but their standard errors tend to be large. As a result, the population values of the coefficients cannot be estimated precisely.
- Detection of multicollinearity is possible through various tests. VIF and CC are commonly used in applied research.
- After detection, the other important point is how to get rid of the problem. Some of the remedies are: (1) using extraneous or prior information, (2) combining cross-sectional and time series data, (3) omitting a highly collinear variable, (4) transforming data, and (5) obtaining additional or new data. Of course, which of these rules will work in practice will depend on the nature of the data, importance of the variable and severity of the collinearity problem.

Check Your Progress for Activities

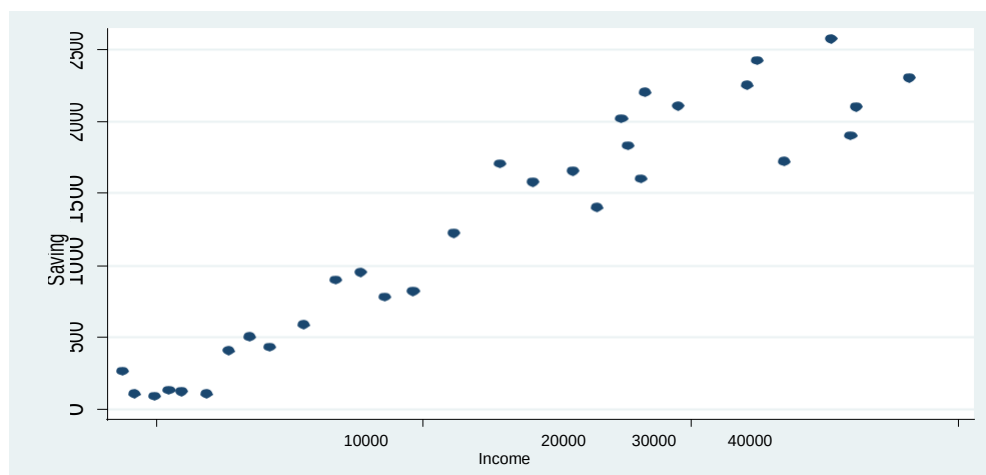
Activity 1.1.1

Applying OLS to the data of the table, we obtain:

$$\begin{aligned} & , = -644.1 + 0.085 X_t \\ & (117.6) (0.005) \qquad R^2 = 0.903 \end{aligned}$$

NB: Dear Learners, as a short cut, if you are able to execute STATA, copy the data on the first two columns and paste it in to STATA spreadsheet so that by using the command **reg saving income**, you can produce the above regression results. Similarly, in STATA using the command **scatter saving income**, you will have the following Scatter plot.

Thus, plotting the observations on a scatter diagram we observe a pattern of increasing dispersion of the savings at higher levels of income, which suggests that the u_i 's are heteroscedastic.



Test 1: Spearman's rank correlation test

To apply this test we rank the X 's and e 's in ascending order. The rankings are shown in following table:

Rank of X	Rank of e	D	D^2
1	16	-15	-225
2	3	-1	1
3	7	-4	16
4	8	-4	16
5	10	-5	25
6	23	-17	28
7	2	5	25
8	1	7	49
9	11	-2	4
10	5	-5	25
11	9	-2	4
12	6	6	36
13	15	-2	4
14	29	-6	36
15	-4	11	121
16	28	-12	144
17	18	-1	1
18	12	6	36
19	19	0	0
21	13	8	64
23	29	-6	36
20	27	-7	49
24	22	2	4
22	14	8	64
25	17	8	64
26	26	0	0
28	21	7	49
27	30	-3	9
29	31	-2	4
30	25	5	25
31	24	7	49
			$D^2 = 1,474$

The rank correlation coefficient estimated from the data of the table is:

$$r_s = 1 - \frac{\sum (R_i - \bar{R})^2}{n(n-1)} = 1 - \frac{(10)(10)}{(11)(10)} \cong 0.7$$

The standard error of r^s is $\frac{1}{\sqrt{n-2}} = \frac{1}{\sqrt{11-2}} \cong 0.14$

Since the estimated Spearman rank correlation coefficient (0.70) is greater than, the standard error of r^s (0.14), showing that the rank correlation is statistically significant and hence, we *reject the null hypothesis of homoscedasticity*.

Test 2: The Goldfeld and Quandt test

The Goldfeld and Quandt test yields the following results.

We order the observation in ascending order of the X 's, and omitting the nine central observations, we are left with two subsets of data, one with the lower values of X and one with the higher values of X .

Subset 1			Subset 2		
n_1	S	X_L	n_2	S	X_H
1	264	8,777	1	1,829	27,670
2	105	9,210	2	1,600	28,150
3	90	9,954	3	2,200	28,300
4	131	10,508	4	2,105	29,560
5	122	10,979	5	2,250	32,100
6	107	11,912	6	2,420	32,500
7	406	12,747	7	1,720	33,500
8	503	13,499	8	2,570	35,250
9	431	14,269	9	1,900	36,000
10	588	15,522	10	2,100	36,200
11	898	16,730	11	2,300	38,200

Applying OLS to each subset we obtain

(a) For subset 1

$$= -738.84 + 0.008X$$

$$(189.4) \quad (0.015)$$

With

$$R^2 = 0.787 \text{ and } \Sigma = 144,771.5$$

(b) For subset 2

$$= 1141.07 + 0.029X$$

$$(709.8) \quad (0.022)$$

With

$$R^2 = 0.152 \text{ and } \Sigma = 769,899.2$$

We form the ratio of the two unexplained variations

$$F = \frac{\frac{\Sigma}{n_1 - 2}}{\frac{\Sigma}{n_2 - 2}} = \frac{769,899.2}{144,771.6} \approx 5$$

The theoretical value of F at the 5 per cent level of significance with

$$F_{0.05, 9, 2} = \frac{31 - 9 - (2)(2)}{2} = 9$$

degrees of freedom is 3,18.

Given that the F-calculated value (5) is greater than F-tabulated value ($F_{0.05, 9, 2} = 3.18$), we reject the null hypothesis of homoscedasticity and conclude that there is heteroscedasticity problem in the saving function.

Thus both tests suggest the existence of heteroscedasticity in the savings function.

Activity 1.1.2

First compute $(4 d_L)$ and $(4 d_U)$ and compare the computed value of d

with d_L , d_U , $(4 d_L)$ and $(4 d_U)$

$$(4 d_L) = 4 - 1.37 = 2.63$$

$$(4 d_U) = 4 - 1.5 = 2.50$$

Since d is less than d_L , we reject the null hypothesis of no autocorrelation and conclude that there is positive autocorrelation problem.

Activity 1.1.3

ai. Regression result shows that income and wealth together explain about 96 percent of the variation in consumption expenditure, and yet neither of the slope coefficients is individually statistically significant.

aii. Moreover, not only is the wealth variable statistically insignificant but also it has the wrong sign. A priori, one would expect a positive relationship between consumption and wealth. This is an implication of multicollinearity.

c. It shows that there is almost perfect collinearity between X_3 and X_2 .

Review Exercises

1. Discuss the Causes and consequences of:
 - i. Multicollinearity
 - ii. Autocorrelation
 - iii. Heteroscedasticity
2. The following table presents hypothetical data on consumption expenditure, $Y(\$)$ in relation to income, $X(\$)$ for a cross-section of 30 families.

Y	X
55	80
65	100
70	85
80	110
79	120
84	115
98	130
95	140
90	125
75	90
74	105
110	160
113	150
125	165
108	145
115	180
140	225
120	200
145	240
130	185
152	220
144	210
175	245
180	260
135	190
140	205
178	265
191	270
137	230
189	250

Suppose we postulate that consumption expenditure is linearly related to income but that heteroscedasticity is present in the data. Using the Goldfeld-Quandt test, verify the hypothesis.

3. Suppose you have a sample of 40 observations and 3 explanatory variables and you want to test for autocorrelation. What can you say about autocorrelation if you have the following Durbin-Watson statistics?

- a. $d = 1.30$
- b. $d = 1.95$
- c. $d = 2.25$

4. Given the following variance-covariance matrices inspect each of them for the autocorrelation and heteroscedasticity problems.

3	0	0
0	5	0
0	0	3

1	1/2	0
1/2	1	1/2
1	1/2	1

3	1/2	1/4
1/2	1	1/2
1/4	1/2	2

5. A financial analyst is interested in the relationship between Earnings per share (EPS) and Sales(S). She collects information on these two variables and estimates the

$$\text{regression: } \text{EPS}_i = \quad + S_i + e_i$$

and, from it, she computes the error from the regression for each company.

Company	EPS	e
1	1200	400.05
2	2210	500.35
3	3201	-50.10
4	3400	-200.72
5	4525	300.25
6	5320	-100.21
7	6001	-87.25

Does heteroscedasticity appear to be a problem in this regression?

6. Consider the model $Y_t = \beta_0 + \beta_1 X_t + U_t$ with the following observation on X and Y

X	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Y	2	2	2	1	3	5	6	6	10	10	10	12	15	10

Test for autocorrelation using Durbin -Watson method

7. The following are error terms from a regression, where $n = 23$ and $k = 2$.

Year	e	Year	e
1970	1.2	1982	-.50
1971	-.3	1983	-.20
1972	2.4	1984	1.10
1973	-1.0	1985	2.10
1974	.4	1986	-1.50
1975	-.5	1987	2.20
1976	-.4	1988	.50
1977	2.3	1989	-3.10
1978	-2.7	1990	4.20
1979	.1	1991	-1.10
1980	-3.00	1992	1.80
1981	2.40		

Compute the Durbin-Watson d statistic. Does autocorrelation appear to be a problem?

8. An equation of demand for food estimated from 50 observations gave the following results:

$$\log q = -0.31 \log p_t + 0.45 \log y_t + 0.65 \log q_{t-1}, R^2 = 0.90, d = 1.8$$

Where: q_t – food consumption per capita, p_t – food price (retail price deflated by Consumption Price Index, CPI), and y_t – per capita disposable income deflated by the CPI. Use an appropriate test and check for the existence of autocorrelation problem

9. The following table shows the annual consumption and disposal income of a hypothetical economy (*in Million's dollar*).

Year	1953	1955	1958	1960	1961	1962	1963	1964	1965	1966	1967	1968
C_t	26.1	29.3	35.6	39.4	42.7	46.3	50.1	54.5	60.1	64.9	69.2	73.1
Y_{dt}	38.3	43.5	53.5	60.8	66.4	71.2	77.2	86.1	94.6	102.4	102.9	115.6

- Estimate the savings function $S_t = f(Y_{dt})$
- Test for heteroscedasticity using Spearman's rank correlation coefficient.
- Re-estimate the savings function with an appropriately transformed model.

CHAPTER TWO

REGRESSION ON DUMMY VARIABLES

Chapter Contents

Dear distance students, this chapter comprises the following major sub-topics:

Introduction

The nature of dummy variables

Dummy Variables for Differences in Intercept Terms

Regression on one quantitative variable and one qualitative variable with two classes, or categories

Regression on one quantitative variable and one qualitative variable with more than two classes

Regression on one quantitative variable and two qualitative variables

Testing for structural stability of regression models: The Slope dummy variables

Interaction effects

Dummy Variables in Seasonal Analysis

Chapter Objectives

At the end of this chapter, students will be able to:

- ✓ define dummy variables
- ✓ formulate dummy variables models
- ✓ know how to use dummy variables for the difference in intercept term
- ✓ estimate various regression equations with dummy variables
- ✓ apply dummy variables to represent the difference in slope coefficients
- ✓ test structural stability
- ✓ use dummy variables to explain seasonal differences

2.1. Introduction

Dear learners, as you might remind from Introduction to Business Statistics or Research Methods for Economists, in empirical analysis we may encounter four types of variables. These are ratio scale, interval scale, ordinal scale and nominal scale variables. Up until now we have been assumed that all variables used in regression analysis are quantitative in nature, which is to say that they have been continuous and hence, in the preceding chapters, in one way or another, we have discussed the ratio scale variables in detail. However, this does not mean that these variables are the only variables that the regression models can deal with. Regression analysis can deal with many interesting variables that are expressed in qualitative terms. These qualitative measures have to be transformed into some proxy so that it could be represented and used in a regression. Dummy variables are discrete transformations and used for this purpose. Thus, the purpose of this chapter is to describe different techniques on how to use dummy variables in regression analysis and how to interpret them.

2.2. The nature of dummy variables

It frequently happens that some of the factors that you would like to introduce into a regression model (that can influence the dependent variable) are qualitative in nature (e.g., sex, race, color, religion, nationality, wars, earthquakes, strikes, political upheavals, and changes in government economic policy) and therefore not measurable in numerical terms. Some illustrative examples are the following:

1. You are investigating the relationship between schooling and earnings, and you have both males and females in your sample. You would like to see if the sex of the respondent makes a difference.
2. You have data on the growth rate of GDP per capita and foreign aid per capita for a sample of developing countries, of which some are democracies and some are not. You would like to investigate whether the impact of foreign aid on growth is affected by the type of government.

In each of these examples, one solution would be to run separate regressions for the two categories and see if the coefficients are different. Alternatively, you could run a single regression using all the observations together, measuring the effect of the qualitative factor with what is known as dummy variables. Dummy variables are constructed by econometricians to be mainly used as proxies for qualitative variables which cannot be measured in any particular case for various reasons. They are artificial variables that work as proxies for qualitative variables and since they are discrete we need to be careful when working with and interpreting them. Dummy variables regression has the two important advantages of providing a simple way of testing whether the effect of the qualitative factor is significant and, provided that certain assumptions are valid, making the regression estimates more efficient. Hence, in this chapter we will mainly deal with nominal scale variables, which are also known **categorical/qualitative/binary or dummy variables**.

***Dear Distance learners,** how can you introduce these variables in your regression analysis?*

To include qualitative variables into empirical analysis, one has to change them into quantitative variables. One way to quantify nominal scale variables is by creating artificial variables that take on values of 1 or 0, where 1 indicates the presence of a certain attribute/quality and 0 indicates the absence of that attribute. This artificial representation is possible since nominal scale variables usually indicate the presence or absence of a certain quality or an attribute.

Gender is a typical example of a qualitative variable that need to be transformed into a numerical form so that it could be used in a regression. Since gender could be male or female it is a categorical variable with two categories. We therefore need to decide what category the dummy should represent and what category that should be used as a reference. If the dummy variable should represent men, the discrete variable D would take the value 1 for men and the value 0 for all other values in the data set. It is therefore important to be sure that all other observations really represent what you want it to represent. A dummy variable for men could therefore be expressed in this way:

$$D_i = \begin{cases} 1 & \text{if male} \\ 0 & \text{otherwise} \end{cases} \quad (2.1)$$

When running the regression you can treat the dummy variable D as any other variables included in the model. This does not change anything for the theory of univariate/multivariate regressions. However, the economic interpretation of the results can be a bit tricky.

The variable D could take other numerical values than 1 and 0, for instance 9 and 7, and it will not have any effect on its coefficient. However, the interpretation is easiest when using 1 and 0, which is the reason why you should follow the structure of (2.1) which is standard.

In passing not that dummy variables can also take other values than 0 and 1, where a linear transformation function $Y = \alpha + \beta D$ (where, α and β are constants, $\beta \neq 0$ and $D = 1$ or 0) can be used to transform the pair (0, 1) to other pairs. This is the reason why qualitative/dummy variables are called nominal scale variables; they have no natural scale of measurement.

By analogy dummy variables for other nominal variables can be created.

Variables that assume 0 and 1 artificially are called **dummy variables**. Hence, dummy variables can be considered as a device to classify data on qualitative variables into mutually exclusive categories such as presence or absence of an attribute.

Dummy variables can serve different purposes in regression analysis. They can be used to:

- i. Allow for differences in intercept terms
- ii. Allow for differences in slopes
- iii. Allow for seasonal differences

2.3. Dummy Variables for Differences in Intercept Terms

Sometimes an investigator may deal with a regression analysis to assess the statistical significance of the relationship between quantitative dependent variable and qualitative explanatory variables. In such cases one often uses dummy variables models, with some assumptions about their use. For instance, the implicit assumption in this case is that the regression lines for the different groups differ only in the intercept term but have the same slope coefficients. *The most basic form of application using dummy variables is when only the intercept is affected.*

Illustration: $Y = \alpha + \beta D_i + U_i$ (2.2) Where, Y is annual salary of accountant, D is 1 if male and 0 if female. So we will have one dummy variable, the omitted category(female) become a base or benchmark to compare with male counterpart

The mean salary of a female accountant by taking conditional expectation is:

$$E(Y/D=0) = \alpha + \beta (0) = \alpha \quad (2.3)$$

The mean salary of a male accountant is:

$$E(Y/D=1) = \alpha + \beta (1) = \alpha + \beta \quad (2.4)$$

The intercept term gives the mean salary of a female accountant while dummy coefficient indicates how much the mean salary of a male differs from a female. The mean salary of males is the intercept plus coefficient of dummy.

Note: We cannot put in 2 dummies (1 for male and 1 for female), else will have perfect collinearity. If we construct D_1 as 1 if male, 0 if female and D_2 as 1 if female, 0 if male, we are creating a common problem of dummy variables regression called the **dummy variable trap**.

Thus, for the dummy variable regression models we can note the following **four** features.

1. To distinguish the two categories, male and female, we have introduced only one dummy variable D_i . For if $D_i = 1$ always denotes a male, when $D_i = 0$ we know that it is a female since there are only two possible outcomes. Hence, one dummy variable suffices to distinguish two categories. The general rule is this: ***If a qualitative variable has 'm' categories, introduce only 'm-1' dummy variables.***
In our example, sex has two categories, and hence we introduced only a single dummy variable. If this rule is not followed, we shall fall into what might be called the dummy variable trap, that is, the situation of perfect multicollinearity.
2. The assignment of 1 and 0 values to two categories, such as male and female, is arbitrary in the sense that in our example we could have assigned $D=1$ for female and $D=0$ for male.
3. The group, category, or classification that is assigned the value of 0 is often referred to as the base, benchmark, control, comparison, reference, or omitted category. It is the base in the sense that comparisons are made with that category.
4. The coefficient attached to the dummy variable D can be called the **differential intercept coefficient** because it tells by how much the value of the intercept term of the category that receives the value of 1 differs from the intercept coefficient of the base category.

Example 2.1: Suppose we want to compare gender differences in salary of an accountant. Assume after regression, we found the following result:

$$Y = 35.20 + 10.25D$$

t

(43.82) (3.45)

Where, Y is income in thousands and D is a dummy variable taking on the value of 1 if male.

Interpretation

The estimated mean salary of female accountants is \$35,200 and the estimated mean salary of male accountants is \$45,450. Thus, t- statistics shows us gender differential is statistically significant since β is significant (because the t-calculated value $=\beta/se(\beta) = 10.25/3.45=2.97$ is greater than the t-tabulated value at 5% significance level for the two tailed test).

2.4. Regression on one quantitative variable and one qualitative variable with two classes, or categories

In similar fashion, using the categorical variable defined by (2.1) we can also form the following model with two explanatory variables (the dummy for gender and for years of experience).

$$Y_i=\beta_0+\beta_1D_i+\beta_2X_i+U_i$$

(2.5)

Where: Y_i annual salary of i^{th} accountant
 X_i Years of experience
 D_i 1 if male and =0, otherwise

Model (2.5) contains one quantitative variable (years of experience) and one qualitative variable (sex) that has two classes (or levels, classifications, or categories), namely, male and female. What is the meaning of this equation? Assuming, as usual, that $E(u_i) = 0$, we see that

$$\text{Mean salary of male accountant: } E(Y_i / X_i, D_i = 1) = (\beta_0 + \beta_1) + \beta_2 X_i$$

(2.6)

$$\text{Mean salary of female accountant: } E(Y_i / X_i, D_i = 0) = \beta_0 + \beta_2 X_i$$

(2.7)

In words, model 2.2 postulates that the male and female accountants' salary functions in relation to the years of working experience have the same slope but different intercepts. In other words, it is assumed that the level of the male accountant's mean salary is different from that of the female accountant's mean salary (by β_1) but the rate of change in the mean annual salary by years of experience is the same for both sexes. If the assumption of common slopes is valid, a test of the hypothesis that the two regressions (2.6) and (2.7) have the same intercept (i.e., there is no sex discrimination) can be made easily by running the regression (2.5) and noting the statistical significance of the estimated β_1 on the

basis of the traditional t test. If the t test shows that $\hat{\beta}_1$ is statistically significant, we reject the null hypothesis that the male and female accountants' levels of mean annual salary are the same.

The only thing that differs between the two expectations is the coefficient for the dummy variable. When $D=1$ we see that the conditional expectation in (2.6) consist of two constants β_0 and β_1 , which sum represents the intercept in that case. However, when $D=0$, the conditional expectation will be given by (2.7) which only contain one constant β_0 . Hence, the model as a whole contains two intercepts β_0 and $\beta_0+\beta_1$. If we take the difference between the two conditional expectations we

receive: $E[Y = 1,] - [= 0,] = (2.8)$ which equals the coefficient for the dummy variable. Since our binary variable D is discrete, we cannot take the derivative of Y with respect to D , since a derivative requires a continuous variable, and therefore is undefined here. In order to find the corresponding marginal effect (derivative) in this case we have to form the difference given by (2.8) and conclude that when D moves from 0 to 1, the conditional expectation of Y change by β_1 units, which represent the marginal effect for the linear model. When working with the linear model, it makes no difference if we treat the dummy variable as it was continuous when calculating the marginal effect since they become the same. But with other functional forms it makes a difference.

Example 2.2

Assume the following regression result from a model given by (2.2) with Y being the hourly wage rate, D a dummy for men, and X a variable for years of schooling. The dependent variable is expressed in USA dollar (\$). Standard errors are given within parenthesis:

$$= 55.9 + 21.9 D + 2.4 X \quad (8.16) \quad (4.30) \quad (0.63)$$

Use the regression results to calculate how much higher the average hourly wage rate is for men. First we have to check if the coefficient for the male dummy is significant. With a t-value ($\beta/\text{se}(\beta)=21.9/4.3$) equals to 5, the coefficient is significantly different from zero at any conventional significance levels. The marginal effect measured with this regression says that men earn 21.9 hours more than women do on average, maintaining years of schooling constant.

2.5. Regression on one quantitative variable and one qualitative variable with more than two classes

The human capital model described above includes a continuous variable for the number of years of schooling. When including a continuous variable for schooling it is under the belief that the hourly wages are set and determined based on this measure. An alternative approach would be to argue that it is the level of schooling, the received diploma that matters in the determination of the wage rate. That calls for a qualitative variable with more than two categories.

For instance:
$$= \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \beta_4 D_4 + \beta_5 D_5 + \beta_6 D_6 + \beta_7 D_7 + \beta_8 D_8 + \beta_9 D_9 + \beta_{10} D_{10} + \beta_{11} D_{11} + \beta_{12} D_{12} + \beta_{13} D_{13} + \beta_{14} D_{14} + \beta_{15} D_{15} + \beta_{16} D_{16} + \beta_{17} D_{17} + \beta_{18} D_{18} + \beta_{19} D_{19} + \beta_{20} D_{20} + \beta_{21} D_{21} + \beta_{22} D_{22} + \beta_{23} D_{23} + \beta_{24} D_{24} + \beta_{25} D_{25} + \beta_{26} D_{26} + \beta_{27} D_{27} + \beta_{28} D_{28} + \beta_{29} D_{29} + \beta_{30} D_{30} + \beta_{31} D_{31} + \beta_{32} D_{32} + \beta_{33} D_{33} + \beta_{34} D_{34} + \beta_{35} D_{35} + \beta_{36} D_{36} + \beta_{37} D_{37} + \beta_{38} D_{38} + \beta_{39} D_{39} + \beta_{40} D_{40} + \beta_{41} D_{41} + \beta_{42} D_{42} + \beta_{43} D_{43} + \beta_{44} D_{44} + \beta_{45} D_{45} + \beta_{46} D_{46} + \beta_{47} D_{47} + \beta_{48} D_{48} + \beta_{49} D_{49} + \beta_{50} D_{50} + \beta_{51} D_{51} + \beta_{52} D_{52} + \beta_{53} D_{53} + \beta_{54} D_{54} + \beta_{55} D_{55} + \beta_{56} D_{56} + \beta_{57} D_{57} + \beta_{58} D_{58} + \beta_{59} D_{59} + \beta_{60} D_{60} + \beta_{61} D_{61} + \beta_{62} D_{62} + \beta_{63} D_{63} + \beta_{64} D_{64} + \beta_{65} D_{65} + \beta_{66} D_{66} + \beta_{67} D_{67} + \beta_{68} D_{68} + \beta_{69} D_{69} + \beta_{70} D_{70} + \beta_{71} D_{71} + \beta_{72} D_{72} + \beta_{73} D_{73} + \beta_{74} D_{74} + \beta_{75} D_{75} + \beta_{76} D_{76} + \beta_{77} D_{77} + \beta_{78} D_{78} + \beta_{79} D_{79} + \beta_{80} D_{80} + \beta_{81} D_{81} + \beta_{82} D_{82} + \beta_{83} D_{83} + \beta_{84} D_{84} + \beta_{85} D_{85} + \beta_{86} D_{86} + \beta_{87} D_{87} + \beta_{88} D_{88} + \beta_{89} D_{89} + \beta_{90} D_{90} + \beta_{91} D_{91} + \beta_{92} D_{92} + \beta_{93} D_{93} + \beta_{94} D_{94} + \beta_{95} D_{95} + \beta_{96} D_{96} + \beta_{97} D_{97} + \beta_{98} D_{98} + \beta_{99} D_{99} + \beta_{100} D_{100} + \beta_{101} D_{101} + \beta_{102} D_{102} + \beta_{103} D_{103} + \beta_{104} D_{104} + \beta_{105} D_{105} + \beta_{106} D_{106} + \beta_{107} D_{107} + \beta_{108} D_{108} + \beta_{109} D_{109} + \beta_{110} D_{110} + \beta_{111} D_{111} + \beta_{112} D_{112} + \beta_{113} D_{113} + \beta_{114} D_{114} + \beta_{115} D_{115} + \beta_{116} D_{116} + \beta_{117} D_{117} + \beta_{118} D_{118} + \beta_{119} D_{119} + \beta_{120} D_{120} + \beta_{121} D_{121} + \beta_{122} D_{122} + \beta_{123} D_{123} + \beta_{124} D_{124} + \beta_{125} D_{125} + \beta_{126} D_{126} + \beta_{127} D_{127} + \beta_{128} D_{128} + \beta_{129} D_{129} + \beta_{130} D_{130} + \beta_{131} D_{131} + \beta_{132} D_{132} + \beta_{133} D_{133} + \beta_{134} D_{134} + \beta_{135} D_{135} + \beta_{136} D_{136} + \beta_{137} D_{137} + \beta_{138} D_{138} + \beta_{139} D_{139} + \beta_{140} D_{140} + \beta_{141} D_{141} + \beta_{142} D_{142} + \beta_{143} D_{143} + \beta_{144} D_{144} + \beta_{145} D_{145} + \beta_{146} D_{146} + \beta_{147} D_{147} + \beta_{148} D_{148} + \beta_{149} D_{149} + \beta_{150} D_{150} + \beta_{151} D_{151} + \beta_{152} D_{152} + \beta_{153} D_{153} + \beta_{154} D_{154} + \beta_{155} D_{155} + \beta_{156} D_{156} + \beta_{157} D_{157} + \beta_{158} D_{158} + \beta_{159} D_{159} + \beta_{160} D_{160} + \beta_{161} D_{161} + \beta_{162} D_{162} + \beta_{163} D_{163} + \beta_{164} D_{164} + \beta_{165} D_{165} + \beta_{166} D_{166} + \beta_{167} D_{167} + \beta_{168} D_{168} + \beta_{169} D_{169} + \beta_{170} D_{170} + \beta_{171} D_{171} + \beta_{172} D_{172} + \beta_{173} D_{173} + \beta_{174} D_{174} + \beta_{175} D_{175} + \beta_{176} D_{176} + \beta_{177} D_{177} + \beta_{178} D_{178} + \beta_{179} D_{179} + \beta_{180} D_{180} + \beta_{181} D_{181} + \beta_{182} D_{182} + \beta_{183} D_{183} + \beta_{184} D_{184} + \beta_{185} D_{185} + \beta_{186} D_{186} + \beta_{187} D_{187} + \beta_{188} D_{188} + \beta_{189} D_{189} + \beta_{190} D_{190} + \beta_{191} D_{191} + \beta_{192} D_{192} + \beta_{193} D_{193} + \beta_{194} D_{194} + \beta_{195} D_{195} + \beta_{196} D_{196} + \beta_{197} D_{197} + \beta_{198} D_{198} + \beta_{199} D_{199} + \beta_{200} D_{200} + \beta_{201} D_{201} + \beta_{202} D_{202} + \beta_{203} D_{203} + \beta_{204} D_{204} + \beta_{205} D_{205} + \beta_{206} D_{206} + \beta_{207} D_{207} + \beta_{208} D_{208} + \beta_{209} D_{209} + \beta_{210} D_{210} + \beta_{211} D_{211} + \beta_{212} D_{212} + \beta_{213} D_{213} + \beta_{214} D_{214} + \beta_{215} D_{215} + \beta_{216} D_{216} + \beta_{217} D_{217} + \beta_{218} D_{218} + \beta_{219} D_{219} + \beta_{220} D_{220} + \beta_{221} D_{221} + \beta_{222} D_{222} + \beta_{223} D_{223} + \beta_{224} D_{224} + \beta_{225} D_{225} + \beta_{226} D_{226} + \beta_{227} D_{227} + \beta_{228} D_{228} + \beta_{229} D_{229} + \beta_{230} D_{230} + \beta_{231} D_{231} + \beta_{232} D_{232} + \beta_{233} D_{233} + \beta_{234} D_{234} + \beta_{235} D_{235} + \beta_{236} D_{236} + \beta_{237} D_{237} + \beta_{238} D_{238} + \beta_{239} D_{239} + \beta_{240} D_{240} + \beta_{241} D_{241} + \beta_{242} D_{242} + \beta_{243} D_{243} + \beta_{244} D_{244} + \beta_{245} D_{245} + \beta_{246} D_{246} + \beta_{247} D_{247} + \beta_{248} D_{248} + \beta_{249} D_{249} + \beta_{250} D_{250} + \beta_{251} D_{251} + \beta_{252} D_{252} + \beta_{253} D_{253} + \beta_{254} D_{254} + \beta_{255} D_{255} + \beta_{256} D_{256} + \beta_{257} D_{257} + \beta_{258} D_{258} + \beta_{259} D_{259} + \beta_{260} D_{260} + \beta_{261} D_{261} + \beta_{262} D_{262} + \beta_{263} D_{263} + \beta_{264} D_{264} + \beta_{265} D_{265} + \beta_{266} D_{266} + \beta_{267} D_{267} + \beta_{268} D_{268} + \beta_{269} D_{269} + \beta_{270} D_{270} + \beta_{271} D_{271} + \beta_{272} D_{272} + \beta_{273} D_{273} + \beta_{274} D_{274} + \beta_{275} D_{275} + \beta_{276} D_{276} + \beta_{277} D_{277} + \beta_{278} D_{278} + \beta_{279} D_{279} + \beta_{280} D_{280} + \beta_{281} D_{281} + \beta_{282} D_{282} + \beta_{283} D_{283} + \beta_{284} D_{284} + \beta_{285} D_{285} + \beta_{286} D_{286} + \beta_{287} D_{287} + \beta_{288} D_{288} + \beta_{289} D_{289} + \beta_{290} D_{290} + \beta_{291} D_{291} + \beta_{292} D_{292} + \beta_{293} D_{293} + \beta_{294} D_{294} + \beta_{295} D_{295} + \beta_{296} D_{296} + \beta_{297} D_{297} + \beta_{298} D_{298} + \beta_{299} D_{299} + \beta_{300} D_{300} + \beta_{301} D_{301} + \beta_{302} D_{302} + \beta_{303} D_{303} + \beta_{304} D_{304} + \beta_{305} D_{305} + \beta_{306} D_{306} + \beta_{307} D_{307} + \beta_{308} D_{308} + \beta_{309} D_{309} + \beta_{310} D_{310} + \beta_{311} D_{311} + \beta_{312} D_{312} + \beta_{313} D_{313} + \beta_{314} D_{314} + \beta_{315} D_{315} + \beta_{316} D_{316} + \beta_{317} D_{317} + \beta_{318} D_{318} + \beta_{319} D_{319} + \beta_{320} D_{320} + \beta_{321} D_{321} + \beta_{322} D_{322} + \beta_{323} D_{323} + \beta_{324} D_{324} + \beta_{325} D_{325} + \beta_{326} D_{326} + \beta_{327} D_{327} + \beta_{328} D_{328} + \beta_{329} D_{329} + \beta_{330} D_{330} + \beta_{331} D_{331} + \beta_{332} D_{332} + \beta_{333} D_{333} + \beta_{334} D_{334} + \beta_{335} D_{335} + \beta_{336} D_{336} + \beta_{337} D_{337} + \beta_{338} D_{338} + \beta_{339} D_{339} + \beta_{340} D_{340} + \beta_{341} D_{341} + \beta_{342} D_{342} + \beta_{343} D_{343} + \beta_{344} D_{344} + \beta_{345} D_{345} + \beta_{346} D_{346} + \beta_{347} D_{347} + \beta_{348} D_{348} + \beta_{349} D_{349} + \beta_{350} D_{350} + \beta_{351} D_{351} + \beta_{352} D_{352} + \beta_{353} D_{353} + \beta_{354} D_{354} + \beta_{355} D_{355} + \beta_{356} D_{356} + \beta_{357} D_{357} + \beta_{358} D_{358} + \beta_{359} D_{359} + \beta_{360} D_{360} + \beta_{361} D_{361} + \beta_{362} D_{362} + \beta_{363} D_{363} + \beta_{364} D_{364} + \beta_{365} D_{365} + \beta_{366} D_{366} + \beta_{367} D_{367} + \beta_{368} D_{368} + \beta_{369} D_{369} + \beta_{370} D_{370} + \beta_{371} D_{371} + \beta_{372} D_{372} + \beta_{373} D_{373} + \beta_{374} D_{374} + \beta_{375} D_{375} + \beta_{376} D_{376} + \beta_{377} D_{377} + \beta_{378} D_{378} + \beta_{379} D_{379} + \beta_{380} D_{380} + \beta_{381} D_{381} + \beta_{382} D_{382} + \beta_{383} D_{383} + \beta_{384} D_{384} + \beta_{385} D_{385} + \beta_{386} D_{386} + \beta_{387} D_{387} + \beta_{388} D_{388} + \beta_{389} D_{389} + \beta_{390} D_{390} + \beta_{391} D_{391} + \beta_{392} D_{392} + \beta_{393} D_{393} + \beta_{394} D_{394} + \beta_{395} D_{395} + \beta_{396} D_{396} + \beta_{397} D_{397} + \beta_{398} D_{398} + \beta_{399} D_{399} + \beta_{400} D_{400} + \beta_{401} D_{401} + \beta_{402} D_{402} + \beta_{403} D_{403} + \beta_{404} D_{404} + \beta_{405} D_{405} + \beta_{406} D_{406} + \beta_{407} D_{407} + \beta_{408} D_{408} + \beta_{409} D_{409} + \beta_{410} D_{410} + \beta_{411} D_{411} + \beta_{412} D_{412} + \beta_{413} D_{413} + \beta_{414} D_{414} + \beta_{415} D_{415} + \beta_{416} D_{416} + \beta_{417} D_{417} + \beta_{418} D_{418} + \beta_{419} D_{419} + \beta_{420} D_{420} + \beta_{421} D_{421} + \beta_{422} D_{422} + \beta_{423} D_{423} + \beta_{424} D_{424} + \beta_{425} D_{425} + \beta_{426} D_{426} + \beta_{427} D_{427} + \beta_{428} D_{428} + \beta_{429} D_{429} + \beta_{430} D_{430} + \beta_{431} D_{431} + \beta_{432} D_{432} + \beta_{433} D_{433} + \beta_{434} D_{434} + \beta_{435} D_{435} + \beta_{436} D_{436} + \beta_{437} D_{437} + \beta_{438} D_{438} + \beta_{439} D_{439} + \beta_{440} D_{440} + \beta_{441} D_{441} + \beta_{442} D_{442} + \beta_{443} D_{443} + \beta_{444} D_{444} + \beta_{445} D_{445} + \beta_{446} D_{446} + \beta_{447} D_{447} + \beta_{448} D_{448} + \beta_{449} D_{449} + \beta_{450} D_{450} + \beta_{451} D_{451} + \beta_{452} D_{452} + \beta_{453} D_{453} + \beta_{454} D_{454} + \beta_{455} D_{455} + \beta_{456} D_{456} + \beta_{457} D_{457} + \beta_{458} D_{458} + \beta_{459} D_{459} + \beta_{460} D_{460} + \beta_{461} D_{461} + \beta_{462} D_{462} + \beta_{463} D_{463} + \beta_{464} D_{464} + \beta_{465} D_{465} + \beta_{466} D_{466} + \beta_{467} D_{467} + \beta_{468} D_{468} + \beta_{469} D_{469} + \beta_{470} D_{470} + \beta_{471} D_{471} + \beta_{472} D_{472} + \beta_{473} D_{473} + \beta_{474} D_{474} + \beta_{475} D_{475} + \beta_{476} D_{476} + \beta_{477} D_{477} + \beta_{478} D_{478} + \beta_{479} D_{479} + \beta_{480} D_{480} + \beta_{481} D_{481} + \beta_{482} D_{482} + \beta_{483} D_{483} + \beta_{484} D_{484} + \beta_{485} D_{485} + \beta_{486} D_{486} + \beta_{487} D_{487} + \beta_{488} D_{488} + \beta_{489} D_{489} + \beta_{490} D_{490} + \beta_{491} D_{491} + \beta_{492} D_{492} + \beta_{493} D_{493} + \beta_{494} D_{494} + \beta_{495} D_{495} + \beta_{496} D_{496} + \beta_{497} D_{497} + \beta_{498} D_{498} + \beta_{499} D_{499} + \beta_{500} D_{500} + \beta_{501} D_{501} + \beta_{502} D_{502} + \beta_{503} D_{503} + \beta_{504} D_{504} + \beta_{505} D_{505} + \beta_{506} D_{506} + \beta_{507} D_{507} + \beta_{508} D_{508} + \beta_{509} D_{509} + \beta_{510} D_{510} + \beta_{511} D_{511} + \beta_{512} D_{512} + \beta_{513} D_{513} + \beta_{514} D_{514} + \beta_{515} D_{515} + \beta_{516} D_{516} + \beta_{517} D_{517} + \beta_{518} D_{518} + \beta_{519} D_{519} + \beta_{520} D_{520} + \beta_{521} D_{521} + \beta_{522} D_{522} + \beta_{523} D_{523} + \beta_{524} D_{524} + \beta_{525} D_{525} + \beta_{526} D_{526} + \beta_{527} D_{527} + \beta_{528} D_{528} + \beta_{529} D_{529} + \beta_{530} D_{530} + \beta_{531} D_{531} + \beta_{532} D_{532} + \beta_{533} D_{533} + \beta_{534} D_{534} + \beta_{535} D_{535} + \beta_{536} D_{536} + \beta_{537} D_{537} + \beta_{538} D_{538} + \beta_{539} D_{539} + \beta_{540} D_{540} + \beta_{541} D_{541} + \beta_{542} D_{542} + \beta_{543} D_{543} + \beta_{544} D_{544} + \beta_{545} D_{545} + \beta_{546} D_{546} + \beta_{547} D_{547} + \beta_{548} D_{548} + \beta_{549} D_{549} + \beta_{550} D_{550} + \beta_{551} D_{551} + \beta_{552} D_{552} + \beta_{553} D_{553} + \beta_{554} D_{554} + \beta_{555} D_{555} + \beta_{556} D_{556} + \beta_{557} D_{557} + \beta_{558} D_{558} + \beta_{559} D_{559} + \beta_{560} D_{560} + \beta_{561} D_{561} + \beta_{562} D_{562} + \beta_{563} D_{563} + \beta_{564} D_{564} + \beta_{565} D_{565} + \beta_{566} D_{566} + \beta_{567} D_{567} + \beta_{568} D_{568} + \beta_{569} D_{569} + \beta_{570} D_{570} + \beta_{571} D_{571} + \beta_{572} D_{572} + \beta_{573} D_{573} + \beta_{574} D_{574} + \beta_{575} D_{575} + \beta_{576} D_{576} + \beta_{577} D_{577} + \beta_{578} D_{578} + \beta_{579} D_{579} + \beta_{580} D_{580} + \beta_{581} D_{581} + \beta_{582} D_{582} + \beta_{583} D_{583} + \beta_{584} D_{584} + \beta_{585} D_{585} + \beta_{586} D_{586} + \beta_{587} D_{587} + \beta_{588} D_{588} + \beta_{589} D_{589} + \beta_{590} D_{590} + \beta_{591} D_{591} + \beta_{592} D_{592} + \beta_{593} D_{593} + \beta_{594} D_{594} + \beta_{595} D_{595} + \beta_{596} D_{596} + \beta_{597} D_{597} + \beta_{598} D_{598} + \beta_{599} D_{599} + \beta_{600} D_{600} + \beta_{601} D_{601} + \beta_{602} D_{602} + \beta_{603} D_{603} + \beta_{604} D_{604} + \beta_{605} D_{605} + \beta_{606} D_{606} + \beta_{607} D_{607} + \beta_{608} D_{608} + \beta_{609} D_{609} + \beta_{610} D_{610} + \beta_{611} D_{611} + \beta_{612} D_{612} + \beta_{613} D_{613} + \beta_{614} D_{614} + \beta_{615} D_{615} + \beta_{616} D_{616} + \beta_{617} D_{617} + \beta_{618} D_{618} + \beta_{619} D_{619} + \beta_{620} D_{620} + \beta_{621} D_{621} + \beta_{622} D_{622} + \beta_{623} D_{623} + \beta_{624} D_{624} + \beta_{625} D_{625} + \beta_{626} D_{626} + \beta_{627} D_{627} + \beta_{628} D_{628} + \beta_{629} D_{629} + \beta_{630} D_{630} + \beta_{631} D_{631} + \beta_{632} D_{632} + \beta_{633} D_{633} + \beta_{634} D_{634} + \beta_{635} D_{635} + \beta_{636} D_{636} + \beta_{637} D_{637} + \beta_{638} D_{638} + \beta_{639} D_{639} + \beta_{640} D_{640} + \beta_{641} D_{641} + \beta_{642} D_{642} + \beta_{643} D_{643} + \beta_{644} D_{644} + \beta_{645} D_{645} + \beta_{646} D_{646} + \beta_{647} D_{647} + \beta_{648} D_{648} + \beta_{649} D_{649} + \beta_{650} D_{650} + \beta_{651} D_{651} + \beta_{652} D_{652} + \beta_{653} D_{653} + \beta_{654} D_{654} + \beta_{655} D_{655} + \beta_{656} D_{656} + \beta_{657} D_{657} + \beta_{658} D_{658} + \beta_{659} D_{659} + \beta_{660} D_{660} + \beta_{661} D_{661} + \beta_{662} D_{662} + \beta_{663} D_{663} + \beta_{664} D_{664} + \beta_{665} D_{665} + \beta_{666} D_{666} + \beta_{667} D_{667} + \beta_{668} D_{668} + \beta_{669} D_{669} + \beta_{670} D_{670} + \beta_{671} D_{671} + \beta_{672} D_{672} + \beta_{673} D_{673} + \beta_{674} D_{674} + \beta_{675} D_{675} + \beta_{676} D_{676} + \beta_{677} D_{677} + \beta_{678} D_{678} + \beta_{679} D_{679} + \beta_{680} D_{680} + \beta_{681} D_{681} + \beta_{682} D_{682} + \beta_{683} D_{683} + \beta_{684} D_{684} + \beta_{685} D_{685} + \beta_{686} D_{686} + \beta_{687} D_{687} + \beta_{688} D_{688} + \beta_{689} D_{689} + \beta_{690} D_{690} + \beta_{691} D_{691} + \beta_{692} D_{692} + \beta_{693} D_{693} + \beta_{694} D_{694} + \beta_{695} D_{695} + \beta_{696} D_{696} + \beta_{697} D_{697} + \beta_{698} D_{698} + \beta_{699} D_{699} + \beta_{700} D_{700} + \beta_{701} D_{701} + \beta_{702} D_{702} + \beta_{703} D_{703} + \beta_{704} D_{704} + \beta_{705} D_{705} + \beta_{706} D_{706} + \beta_{707} D_{707} + \beta_{708} D_{708} + \beta_{709} D_{709} + \beta_{710} D_{710} + \beta_{711} D_{711} + \beta_{712} D_{712} + \beta_{713} D_{713} + \beta_{714} D_{714} + \beta_{715} D_{715} + \beta_{716} D_{716} + \beta_{717} D_{717} + \beta_{718} D_{718} + \beta_{719} D_{719} + \beta_{720} D_{720} + \beta_{721} D_{721} + \beta_{722} D_{722} + \beta_{723} D_{723} + \beta_{724} D_{724} + \beta_{725} D_{725} + \beta_{726} D_{726} + \beta_{727} D_{727} + \beta_{728} D_{728} + \beta_{729} D_{729} + \beta_{730} D_{730} + \beta_{731} D_{731} + \beta_{732} D_{732} + \beta_{733} D_{733} + \beta_{734} D_{734} + \beta_{735} D_{735} + \beta_{736} D_{736} + \beta_{737} D_{737} + \beta_{738} D_{738} + \beta_{739} D_{739} + \beta_{740} D_{740} + \beta_{741} D_{741} + \beta_{742} D_{742} + \beta_{743} D_{743} + \beta_{744} D_{744} + \beta_{745} D_{745} + \beta_{746} D_{746} + \beta_{747} D_{747} + \beta_{748} D_{748} + \beta_{749} D_{749} + \beta_{750} D_{750} + \beta_{751} D_{751} + \beta_{752} D_{752} + \beta_{753} D_{753} + \beta_{754} D_{754} + \beta_{755} D_{755} + \beta_{756} D_{756} + \beta_{757} D_{757} + \beta_{758} D_{758} + \beta_{759} D_{759} + \beta_{760} D_{760} + \beta_{761} D_{761} + \beta_{762} D_{762} + \beta_{763} D_{763} + \beta_{764} D_{764} + \beta_{765} D_{765} + \beta_{766} D_{766} + \beta_{767} D_{767} + \beta_{768} D_{768} + \beta_{769} D_{769} + \beta_{770} D_{770} + \beta_{771} D_{771} + \beta_{772} D_{772} + \beta_{773} D_{773} + \beta_{774} D_{774} + \beta_{775} D_{775} + \beta_{776} D_{776} + \beta_{777} D_{777} + \beta_{778} D_{778} + \beta_{779} D_{779} + \beta_{780} D_{780} + \beta_{781} D_{781} + \beta_{782} D_{782} + \beta_{783} D_{783} + \beta_{784} D_{784} + \beta_{785} D_{785} + \beta_{786} D_{786} + \beta_{787} D_{787} + \beta_{788} D_{788} + \beta_{789} D_{789} + \beta_{790} D_{790} + \beta_{791} D_{791} + \beta_{792} D_{792} + \beta_{793} D_{793} + \beta_{794} D_{794} + \beta_{795} D_{795} + \beta_{796} D_{796} + \beta_{797} D_{797} + \beta_{798} D_{798} + \beta_{799} D_{799} + \beta_{800} D_{800} + \beta_{801} D_{801} + \beta_{802} D_{802} + \beta_{803} D_{803} + \beta_{804} D_{804} + \beta_{805} D_{805} + \beta_{806} D_{806} + \beta_{807} D_{807} + \beta_{808} D_{808} + \beta_{809} D_{809} + \beta_{810} D_{810} + \beta_{811} D_{811} + \beta_{812} D_{812} + \beta_{813} D_{813} + \beta_{814} D_{814} + \beta_{815} D_{815} + \beta_{816} D_{816} + \beta_{817} D_{817} + \beta_{818} D_{818} + \beta_{819} D_{819} + \beta_{820} D_{820} + \beta_{821} D_{821} + \beta_{822} D_{822} + \beta_{823} D_{823} +$$

The first and most basic approach is to create three binary variables; one for each educational level, in the following way:

$$D_1 = \begin{cases} 1 & \text{if } h = \text{primary} \\ 0 & \text{otherwise} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{if } h = \text{secondary} \\ 0 & \text{otherwise} \end{cases}$$

$$D_3 = \begin{cases} 1 & \text{if } h = \text{post secondary} \\ 0 & \text{otherwise} \end{cases}$$

We can now treat D_1 , D_2 and D_3 as three explanatory variables, and include them in the regression model. However, it is important to avoid the so called **dummy variable trap**.

Consequences of the Dummy Variable Trap

What would happen if you included a dummy variable for the reference category? There would be two consequences. First, was it possible to compute regression coefficients, you would not be able to give them an interpretation. The coefficient α is a basic estimate of the intercept, and the coefficients of the dummies are the estimates of the increase in the intercept from this basic level, but now there is no definition of what is basic, so the interpretation collapses. The other consequence is that the numerical procedure for calculating the regression coefficients will break down and the computer will simply send you an error message (or possibly, in sophisticated applications, drop one of the dummies for you).

Suppose that there are m dummy categories and you define dummy variables D_1 ----- D_m .
Then, in observation i , $\sum_{j=1}^m D_{ij} = 1$, because one of the dummy variables will be equal to 1 and all the others will be equal to 0. As a consequence the model is subject to a special case of exact multicollinearity, making it impossible to compute regression coefficients.

The dummy variable trap appears when the analyst-tries to specify and estimate the following model:

$$\ln Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \beta_4 X + U \tag{2.10}$$

It is a mathematical impossibility to estimate the parameters in (2.10) since there is no variation in the sum of the three dummy variables, since $D_1 + D_2 + D_3 = 1$ for all observations in the data set. Since the model only can contain one constant, in this case the intercept, we cannot include all three dummy variables. The easiest way to solve this is to exclude one of them and treat the excluded category as a-reference category. We re-specify the model in following way:

$$\ln Y = B_0 + \beta_2 D_2 + \beta_3 D_3 + \beta_4 X + U \tag{2.11}$$

That is, if D_1 , is excluded, the other categories will have D_1 , as reference. B_2 will therefore be interpreted as the wage effect of going from a primary schooling diploma to a secondary schooling diploma, and β_3 will represent the wage effect of going from a primary schooling diploma to a post secondary schooling diploma.

An alternative way to get out of this trap is by introducing as many dummy variables as the number of categories of that variable, *provided we do not introduce the intercept in such a model*. Thus, if we drop the intercept term from (2.11), and consider the following model,

$$\ln Y = \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \beta_4 X + U \tag{2.12}$$

we do not fall into the dummy variable trap, as there is no longer perfect collinearity. The three dummy variables will then work as three intercepts in this model; one for each educational level. The coefficients can therefore not be interpreted as relative changes in this case.

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

As Kennedy notes (Cited by **Gujarati**, 2004): Most researchers find the equation with an intercept more convenient because it allows them to address more easily the questions in which they usually have the most interest, namely, whether or not the categorization makes a difference, and if so, by how much. If the categorization does make a difference, by how much is measured directly by the dummy variable coefficient estimates. Testing whether or not the categorization is relevant can be done by running a t test of a dummy variable coefficient against zero (or, to be more general, an F test on the appropriate set of dummy variable coefficient estimates).

Example 2.3

Estimate the parameters of (2.13) and (2.14) and compare and interpret the results.

$$\text{Specification I: (2.13) } \ln Y = 3.929 + 0.154D_2 + 0.295D_3 + 0.009X$$

$$(0.043) \quad (0.024) \quad (0.022) \quad (0.001)$$

$$\text{Specification II: (2.14) } \ln Y = 3.929D_1 + 4.083D_2 + 4.224D_3 + 0.009X$$

$$(0.043) \quad (0.034) \quad (0.036) \quad (0.001)$$

The three dummy variables represent three educational levels, and X represents the age of the individual. The first thing to notice is that $\beta_0 = C_1$, $\beta_0 + \beta_2 = C_2$ and $\beta_0 + \beta_2 + \beta_3 = C_3$. Hence, the two specifications are very much related. Furthermore $C_2 - C_1 = \beta_2$ and $C_3 - C_2 = \beta_3$. With help from specification II, we can derive the effect of going from a high school diploma to a college diploma by taking the difference between C_3 and C_2 which turns out to be equal to 0.14, i.e. a 14 percent increase. However, that effect could also have been received by taking the difference between β_3 and β_2 . For the obvious reason there should be no change in the effect of the other variables included in the model (β_4 in this example) when alternating between specification I and II.

Example 2.4: If we want to look at the effect of location(Addis Ababa, Hawassa, Arba Minch etc) on Person's salary in thousands of Birr (Y). We create variable for towns- three variables:

- Adiss Ababa- $D_1 = 1$ if from Adiss Ababa, 0 otherwise
- Hawassa – $D_2 = 1$ if from Hawassa, 0 otherwise
- Arba Minch- $D_3 = 1$ if from Arba Minch, 0 otherwise
- Multiple Regression Model: $Y = \gamma_0 + \gamma_1 D_1 + \gamma_2 D_2 + e$
- **Result:** $Y = \gamma_0 + 3.8D_1 + 2.1D_2$

Activity 2.1: Why was Arba Minch dropped?

Interpretation: Adiss Ababa $= \gamma_1 = 3.8$ means that people in Adiss Ababa earn birr 3800 more on average relative to people in Arbaminch.

Activity 2.2: Interpret Hawassa $= \gamma_2 = 2.1$

2.6. Regression on one quantitative variable and two qualitative variables

As you might have noted, so far we have been dealing with regression analysis where there is only one qualitative variable. In practice, however, we may have more than one qualitative variable affecting the dependent variable. Then, how will you introduce more than one qualitative variable in your regression analysis? The introduction of qualitative variables in regression analysis is as direct as the introduction of quantitative explanatory variables. However, for each qualitative variable we introduce one or more than one dummy following the rule mentioned above.

Illustration 1: Suppose that we want to analyze the determinates of household consumption (C), and that we have data on the income of the household (Y), the sex of household head (male or female), the age of the household head (given in three categories such as less than 30 years, 30 to 60 years and greater than 60 years) and the education of the household head (given in three categories such as less than high school, greater than or equal to high school but less than collage degree and greater than or equal to college degree). Following the rule of dummy variables (**m-1**), for each variable the number of dummies included in the regression analysis is one less than the number of its categories. The coefficients of the dummies are interpreted as the differences between the average consumption of the omitted (base) category and the category represented by the dummy under consideration, keeping other things constant.

Illustration 2: Let us revert to the accountants' salary regression (2.5), but now assume that in addition to years of experience and sex, the skin color of the accountant is also an important determinant of salary. For simplicity, assume that color has two categories: black and white. We can now write (2.5) as:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i \quad (2.13)$$

Where Y_i annual salary, X_i years of working experience, $D_2 = 1$ if male and $=0$ otherwise, $D_3 = 1$ if white and $=0$ otherwise.

Note that: Each of the two qualitative variables, sex and color, have two categories and hence need one dummy variable for each. Note also that the omitted, or base, category now is "black female accountant."

Assuming $E(u_i) = 0$, we can obtain the following regression from (2.13)

Mean salary for black female accountant:

$$E(Y_i | D_2 = 0, D_3 = 0, X_i) = \beta_0 + \beta_1 X_i$$

Mean salary for black male accountant:

$$E(Y_i | D_2 = 1, D_3 = 0, X_i) = (\beta_0 + \beta_1) + \beta_1 X_i$$

Mean salary for white female accountant:

$$E(Y_i | D_2 = 0, D_3 = 1, X_i) = (\beta_0 + \beta_3) + \beta_1 X_i$$

Mean salary for white male accountant:

$$E(Y_i | D_2 = 1, D_3 = 1, X_i) = (\beta_0 + \beta_1 + \beta_3) + \beta_1 X_i$$

Once again, it is assumed that the preceding regressions differ only in the intercept coefficient but not in the slope coefficient β_1 .

An OLS estimation of (2.15) will enable us to test a variety of hypotheses. Thus, if β_3 is statistically significant, it will mean that color does affect an accountant's salary. Similarly, if β_2 is statistically significant, it will mean that sex also affects an accountant's salary. If both these differential intercepts are statistically significant, it would mean sex as well as color is an important determinant of accountants' salaries.

Example 2.5: Assume the house with two attributes that can further be two qualitative variables. Let us denote the first variable as D_1 . This variable can take values 1 (if the house has a driveway) or 0 (if it does not). And also denote the second variable as D_2 . This variable can also take values 1 (if the house has a recreation room) or 0 (if it does not). X_1 is the size of the garden and X_2 is land rent and Y is the price of the house.

Now, suppose we will run the regression of Y on the four explanatory variables and a constant. The result will be:

$$Y = 2736 + 12598D_1 + 10969D_2 + 5197.X_1 + 10562.X_2$$

That means that if the house has no driveway ($D_1 = 0$) and no recreation room ($D_2 = 0$), its value will be $Y = 2736 + 5197.X_1 + 10562.X_2$. If the house has a driveway, its value will be, holding all other variables constant (or *ceteris paribus*), \$12598 more. If the house has a recreation room, its value will be, *ceteris paribus*, \$10969 more. If the house has a driveway and a recreation room, its value will be, *ceteris paribus*, $12598 + 10969 = \$23567$ more.

What is peculiar in the above result is that the marginal effect of the size of the lot, or the number of bedrooms, is the same for the four kinds of houses. Increasing the size of the garden by 1 square foot will increase the price of the house by \$5.197 whether the house has or not a driveway or a recreation room. This assumption may be justified or not.

From the preceding discussion it follows that we can extend our model to include more than one quantitative variable and more than two qualitative variables. *The only precaution to be taken is that the number of dummies for each qualitative variable should be one less than the number of categories of that variable.*

Thus, you can see that dummy explanatory variables do not change the econometrics. In this chapter of module we have just introduced a few ideas on the best way to give good economic interpretation for the result. However, dealing with an explained or dependent dummy variable is a much more difficult problem. In this case the econometrics of OLS does not work anymore and we must use another econometrics, called discrete choice or binary models. Actually, the basis of this new econometrics is not very difficult. However, it is different and will not be investigated in this undergraduate course. Rather it will be covered in the graduate studies.

2.7. Testing for structural stability of regression models: The Slope dummy variables

As could be seen in the previous section, the dummy variable could work as an intercept shifter. Sometimes it is reasonable to believe that the shift should take place in the slope coefficient instead of the intercept. If the slopes are in fact different, testing for differences in the intercepts may be of little practical significance. Therefore, we need to develop a general methodology to find out whether two (or more) regressions are different, where the difference may be in the intercepts or the slopes or both.

If we go back to the human capital model it is possible to argue that the difference in wage rate between men and women could be due to differences in their return to education. This would mean that men and women have slope coefficients that are different in size.

A model that control for differences in the slope coefficient for different categories of the qualitative variable could be expressed in the following way:

$$\begin{aligned}\ln Y &= \beta_0 + (\beta_1 + \beta_2 D) X + U_i \\ &= \beta_0 + \beta_1 X + \beta_2 (DX) + U_i\end{aligned}\quad (2.14)$$

In this case the slope coefficient for X equals β_1 when $D=0$ and $\beta_1 + \beta_2$ when $D=1$. Hence, a way to test if the return to education differs between men and women would be to test if β_2 is different from zero, which should be tested before going on to test if $\beta_1 + \beta_2$ is different from zero. Observe that the coefficient for the cross product is interpreted differently if both variables had been continuous. Since D is binary, DX is only active when $D=1$, and the corresponding effect is therefore related to the category specified by $D=1$.

Example 2.6

Use the same data set as in Example 2.5 and estimate the coefficients in (2.14). The results are presented below with standard errors within parenthesis:

$$\begin{aligned}\ln Y &= 4.11 + 0.024X + 0.014DX \\ &\quad (0.031) (0.003) (0.001)\end{aligned}$$

Use the regression results to investigate if there is a difference in the return to education between men and women. To answer that question we simply test the estimated coefficient for the cross product. Doing that, we receive a t-value of 10.3 which is above any critical values of conventional significance level. Hence, if this specification is correct, we can conclude that the returns to education differ between men and women.

2.7.1. A model with Intercept and Slope dummy variable

Whenever working with cross products it is very important to always include the involved variables separately to separate that kind of effect from the cross product. If it is the case that D in itself has a positive effect on the dependent variable, that unique effect will be part of the cross effect otherwise. Hence whenever including a cross product the model should be specified in the following way:

$$\ln Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 (DX) + U \quad (2.15)$$

When we include the two variables, X and D separately and together with their product we allow for changes in both the intercept and the slope. If it turns out that the coefficient of β_2 is not significant one can go on and reduce the specification to (2.14), but not otherwise.

Example 2.7

Extend the specification of (2.14) by including D separately. That is, estimate the parameters of the model given by (2.15) and interpret the results. Doing that, we received the following results, with standard errors within parenthesis.

$$\ln Y = 4.006 + 0.033X + 0.210D - 0.002DX$$

$$(0.045) (0.004) (0.062) (0.005)$$

By investigating the t-values we see that β_1 , and β_2 are statistically significant from zero. But the t-value from the cross product is not significant any more. Since D alone has a significant effect on the dependent variable, there is little effect left from the cross product, and hence we conclude that there is no difference in the return to education between men and women.

Example 2.6 and 2.7 should convince you that it is very important to include the variables that appear in a cross product separately, since they might stand for the main effect. In Example 2.6 we did not include D even though it was relevant. Omitting relevant variables has consequences and biases the remaining coefficients. In this case it made us to draw the wrong conclusion about the return to education for men and women.

2.7.2. Test for structural differences

An important application using dummy variables is to test if the coefficients of the model differ for different sub groups of the population, or if they have changed over time. For instance, assume that we have the following wage equation expressed with a semi logarithmic (log-linear) functional form:

$$\ln Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U \quad (2.16)$$

with Y being the wage rate, X_1 the number of years of schooling, and X_2 the number of years of working experience. We would like to know if β_1 and β_2 differ between men (m) and women (w) simultaneously. That is, we would like to test the following hypothesis:

$$H_0: \beta_{1m} = \beta_{1w}, \beta_{2m} = \beta_{2w}$$

$$H_1: \beta_{1m} \neq \beta_{1w} \text{ and / or } \beta_{2m} \neq \beta_{2w}$$

In order to carry out this test using dummy variables we need to create an indicator variable D , for let's say for men, and then form the following regression model:

$$\ln Y = \beta_0 + \beta_1 D + \beta_2 X_1 + \beta_3 (X_1 D) + \beta_4 X_2 + \beta_5 (X_2 D) + U \quad (2.17)$$

Equation (2.17) will be representing the unrestricted model, where men and women are allowed to have different coefficients, and equation (2.16) will be representing the restricted case where men and women have the same coefficients. We will now compare the residual sums of squares (RSS) between the two models and use those in our test statistic given by:

$$F = \frac{(RSS_R - RSS_U) / df_1}{RSS_U / df_2} \sim F_{df_1, df_2} \quad (2.18)$$

If the RSS_R is very different from RSS_U we will reject the null hypothesis in favor of the alternative hypothesis. If they are similar in size, the test value will be very small and we say that the coefficients are the same for men and women.

Example 2.8

Assume that you would like to know if the coefficients of equation (2.16) differ between men and women in the population. In order to test the joint hypothesis we need to run two regression models; one restricted model given by (2.16) and one unrestricted model given by (2.17). Using a sample of 1483 randomly selected individuals we obtained the following results:

Table 2.1 Regression results

	Restricted model	Unrestricted model
Residual sum of squares (RSS)	145.603	140.265
Degrees of freedom (n – k)	1483 – 3 = 1480	1483 – 6 = 1477

Using the results given in Table 2.1 we may calculate the test value using (2.18). The degree of freedom for the numerator is calculated as the difference between degrees of freedom for the RSS from the restricted and unrestricted model. That is $1480 - 1477 = 3$. Another way to think about the degrees of freedom for the numerator is to express it in terms of the number of restrictions imposed by the restricted model compared to the unrestricted. The unrestricted model has 6 parameters, while the restricted model has only 3, which means that three parameters have been set to zero in the restricted model. Therefore we have 3 restrictions. The test value using the test statistic is therefore equal to:

$$F = \frac{(RSS_R - RSS_U) / (k_R - k_U)}{RSS_U / (n - k_U)} = \frac{(145.603 - 140.265) / (1480 - 1477)}{140.265 / 1477} = 18.73$$

The test value has to be compared with a critical value. Using a significance level of 5 percent the F-critical value equals 2.6. Hence, the test value is much larger than the critical value which means that we can reject the null hypothesis. We can therefore conclude that the coefficient of the regression model differ for men and women.

2.8. Interaction effects

Consider the following model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i \tag{2.19}$$

where Y_i = annual expenditure on clothing

X_i = income

$D_2 = 1$ if female
= 0 if male

$D_3 = 1$ if college graduate
= 0 otherwise

Implicit in this model is the assumption that the differential effect of the sex dummy D_2 is constant across the two levels of education and the differential effect of the education dummy D_3 is also constant across the two sexes. That is, if, say, the mean expenditure on clothing is higher for females than males this is so whether they are college graduates or not. Likewise, if, say, college graduates on the average spend more on clothing than non college graduates, this is so whether they are female or males.

In many applications such an assumption may be untenable. A female college graduate may spend more on clothing than a male graduate. In other words, there may be interaction between the two qualitative variables D_2 and D_3 and therefore their effect on mean Y may not be simply additive as in (2.19) but multiplicative as well, as in the following model:

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 (D_{2i} D_{3i}) + X_i \beta + u_i \quad (2.20)$$

From (2.20) we obtain

$$E(Y_i | D_2 = 1, D_3 = 1, X_i) = (\beta_1 + \beta_2 + \beta_3 + \beta_4) + X_i \beta \quad (2.21)$$

which is the mean clothing expenditure of graduate females.

Notice that:

β_2 differential effect of being a female

β_3 differential effect of being a college graduate

β_4 differential effect of being a female graduate which shows that the mean clothing expenditure of graduate females is different (by β_4) from the mean clothing expenditure of females or college graduates. If β_2 , β_3 , and β_4 are all positive, the average clothing expenditure of females is higher (than the base category, which here is male non graduate), but it is much more so if the females also happen to be graduates. Similarly, the average expenditure on clothing by a college graduate tends to be higher than the base category but much more so if the graduate happens to be a female. This shows how the **interaction dummy** modifies the effect of the two attributes considered individually. Whether the coefficient of the interaction dummy is statistically significant can be tested by the usual t test. If it turns out to be significant, the simultaneous presence of the two attributes will attenuate or reinforce the individual effects of these attributes. Needless to say, omitting a significant interaction term incorrectly will lead to a specification bias.

2.9. Dummy Variables in Seasonal Analysis

Many economic time series based on monthly or quarterly data exhibit seasonal patterns, i.e., they show regular oscillatory movements. For example, household demand for money at holiday times shows seasonal pattern. Hence, it is desirable to remove the seasonal component from a time series so that one can concentrate on the other components, such as the trend. The process of removing the seasonal component from a time series is known as **deseasonalization**, and the series thus obtained is called **deseasonalized or seasonally adjusted** time series.

Usually the important economic series such as the unemployment rate, the consumer price index (*CPI*), the producer price index (*PPI*), and the industrial production index that we see in different reports are published in their seasonally adjusted form.

One of the methods used to **deseasonalize** a time series is a method of dummy variables.

To illustrate the dummy variables technique of deseasonalization, suppose that we have quarterly data on consumption (C) and income (Y). The seasonal adjustment in this case can be estimated by including among the explanatory variables three seasonal dummy variables, Q_1 , Q_2 and Q_3 .

In order to deal with this, we fit the following the quarterly regression model:

$$C_t = \beta_0 + \beta_1 Y_{1t} + \dots + \beta_k Y_{kt} + \alpha_1 Q_{1t} + \alpha_2 Q_{2t} + \alpha_3 Q_{3t} + u_t \quad (2.22)$$

where

$$\begin{matrix} & 1 & h & h \\ = & 1 & h & h \\ = & 0 & h & h \\ = & \frac{1}{h} & h & h \end{matrix}$$

Note that in equation 2.22, we have treated the fourth quarter as the base category. Consequently, the coefficients attached to the various dummies are differential intercepts, showing by how much the average value of C in the quarter that receives a dummy value of 1 differs from that of the base category, holding the income level constant. In other words, the coefficients on the seasonal dummies will give seasonal increase or decrease in the average value of C relative to the base season, for a given value of Y . Furthermore, the coefficient of Y shows that, allowing for seasonal effects, if income (Y) changes by one unit, on average, consumption (C) will change by β units.

Dear distance learners and readers, so far we have seen how to deseasonalize the dependent variable (for example C in equation 2.22). Now, an interesting question is: *Just as the dependent variable, C , exhibits seasonal patterns would not the explanatory variable, Y also exhibit seasonal patterns? How, then, do we take into account seasonality in the explanatory variable Y ?*

Don't worry! Now you have got the key at your hand, that is, the dummy variables technique. The interesting thing about equation 2.22 is that the dummy variables in the model not only remove the seasonality in C but also the seasonality in Y , if any. It is just like killing two birds with a single stone.

Chapter Summary:

1. Dummy variables, taking values of 1 and zero (or their linear transforms), are a means of introducing qualitative regressors in regression models.
2. Follow either of the two methods to introduce a dummy variable:
 - (a) Introduce a dummy for each category of a qualitative regressor and omit the intercept term or
 - (b) Include the intercept term and introduce only $(m-1)$ dummies, where m is the number of categories of the qualitative variable.
3. The category for which no dummy variable is assigned is known as the **base, benchmark, control, comparison, reference, or omitted category**. And all comparisons are made in relation to the benchmark category.
4. Dummy variables are a data-classifying device in that they divide a sample into various subgroups based on qualities or attributes (sex, marital status, race, religion, etc.) and implicitly allow one to run individual regressions for each subgroup. If there are differences in the response of the regress and to the variation in the quantitative variables in the various subgroups, they will be reflected in the differences in the intercepts or slope coefficients, or both, of the various subgroup regressions.
5. The value of the intercept represents the *mean value* of the benchmark category
6. The coefficients attached to the dummy variables are known as the **differential intercept coefficients** because they tell by how much the value of the intercept that receives the value of 1 differs from the intercept coefficient of the benchmark category.
7. If a model has several qualitative variables with several classes, introduction of dummy variables can consume a large number of degrees of freedom. Therefore, one should always weigh the number of dummy variables to be introduced against the total number of observations available for analysis.
8. Since the dummy variables are non stochastic, they pose no special problems in the application of OLS. However, care must be exercised in transforming data involving dummy variables. In particular, the problems of autocorrelation and heteroscedasticity need to be handled very carefully.

Check Your Progress for Activities**Activity 2.1:**

Arba Minch is dropped because it serves as a benchmark or a base for comparison

Activity 2.2:

Hawassa $\gamma_2=2.1$ implies that people in Hawassa earn birr 2100 more on average relative to people in Arba Minch.

Review Exercises

1. Explain the following concepts
 - a. Dummy variables
 - b. Seasonal dummy variables
 - c. Dummy variables trap
2. Consider the following model: $Y_i = \alpha_1 + \alpha_2 D_i + \beta X_i + u_i$, where Y = annual salary of a college professor, X = years of teaching experience and D = dummy for gender
Consider three ways of defining the dummy variable.

- a. $D = 1$ for male, 0 for female.
- b. $D = 1$ for female, 2 for male.
- c. $D = 1$ for female, -1 for male.

Interpret the preceding regression model for each dummy assignment. Is one method preferable to another? Justify your answer.

3. Given the regression result: $Y = 25.20 + 8.25D + 1.25X$
t (32.85) (10.45) (15.11)

Where, Y is income in thousands, D is a dummy variable taking on the value of 1 if male 0 for female and X = years of experience.

- a. Calculate the mean salary of male and female assuming 0 years of experience
 - b. Test the significance of coefficient of years of experience and Calculate the salary addition for male or female accountant after one year experience
 - c. Holding experience constant, check whether the mean salaries of male and female accountants are different or not
4. If you have monthly data over 30 years, how many dummy variables will you introduce in your regression analysis to test the following hypotheses?
 - a. All the 12 months of the year exhibit seasonal patterns.
 - b. Only April, June and August and December exhibit seasonal patterns.
5. Suppose that the regression program of a graduating class Economics student of Arba Minch University refused to estimate four seasonal coefficients when he enters the quarterly data Including zero to one dummy variables for each quarter. What is he supposed to do?

CHAPTER THREE

DYNAMIC ECONOMETRIC MODELS

Chapter Contents

Dear distance students, this chapter consists of the following major sub-topics:

Introduction

Distributed lag and Autoregressive models

The Role of Lag in Economics

The Reasons for Lags

Infinite Distributed Lag

Estimation of Lag Models

The Granger Causality Test

Chapter Objectives

At the end of this chapter, students will be able to:

- ✓ define dynamic models
- ✓ differentiate distributed lag models from autoregressive models
- ✓ explain why lags are commonly used in economic analysis
- ✓ recognize the use of infinite distributed lags in applied economics
- ✓ know the various estimation methods commonly used for lag models
- ✓ apply the Granger Causality test for different macroeconomic variables

3.1. Introduction

So far in the standard regression model we ignored the effect of timing (lagged values) 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 economic models, 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**. Many economic models have lagged values of the regressors in the regression equation.

For example, it takes time to build roads and highways. Therefore, the effect of this public investment on growth in GNP will show up with a lag, and this effect will probably go on about for several years. It takes time before investment in research and development pays off in new inventions which in turn take time to develop into commercial products.

In studying consumption behavior, a change in income may affect consumption over several periods. This is true in the permanent income theory of consumption, where it may take the consumer several periods to determine whether the change in real disposable income was temporary or permanent. For example, is the extra consulting money earned this year going to continue next year?

Also, lagged values of real disposable income appear in the regression equation because the consumer takes into account his life time earnings in trying to smooth out his consumption behavior. In turn, one's life time income may be guessed by looking at past as well as current earnings.

Therefore, realistic formulations of economic relations often require the insertion of lagged values of the explanatory or insertion of lagged dependent variables.

Autoregressive and distributed-lag models are used extensively in econometric analysis, and in this chapter we take a close look at such models with a view to finding out the following:

- What is the role of lags in economics?
- What are the reasons for the lags?
- Is there any theoretical justification for the commonly used lagged models in empirical econometrics?
- What is the relationship, if any, between autoregressive and distributed-lag models? Can one be derived from the other?
- What are some of the statistical problems involved in estimating such models?
- Does a lead–lag relationship between variables imply causality? If so, how does one measure it?

3.2. Lagged Models: Distributed lag and Autoregressive models

Regression models that take in to account time lags are known as **dynamic or lagged regression models**.

In regression analysis involving time series data, the regression model may include both the current and the lagged (past) values of the explanatory variables (the X’s) as well as the lagged value of the dependent variable.

There are two types of lagged models. These are: **Distributed and Autoregressive lag models**.

Dear students, in this particular module we will invest more of our time in discussing distributed lag models since in one way or another, an autoregressive model was discussed in autocorrelation section of chapter one.

Time series data are any set of data from a quantifiable or qualitative event that are recorded overtime or

A time series is a sequence of data points/numerical data measured typically successive points in time spaced at uniform time intervals.

Time series econometrics/analysis compares methods for analyzing time series data inorder to extract meaningful statistics and other characteristics of the data.

3.2.1. Distributed lag models

Distributed-lag models are models which include lagged values of the exogenous variables and/or lagged values of the dependent variables among the set of explanatory variables.

The general form of a distributed-lag model with only lagged exogenous variables is:

$$Y_t = a + b_0X_t + b_1X_{t-1} + b_2X_{t-2} + ... + b_sX_{t-s} + ... + U_t \tag{3.1}$$

They are called **distributed lag models** because the influence of the explanatory variable on the dependent variable is distributed over a number of past values of X.

The coefficient b_0 is known as the *short run, or impact, multiplier* because it gives the change in mean value of Y following a unit change in X in the same time period t. If the change in X is maintained at the same level thereafter, then b_0+b_1 gives the change in the mean value of Y in the next period, and $b_0+b_1+ b_2$ in the following period, and so on. These partial sums are called *interim, or intermediate, multipliers*.

Suppose that we observe X_t from 1955 to 1995. X_{t-1} is the same variable but for the previous period, i.e., 1954-1994. Since 1954 is not observed, we start from 1955 for X_{t-1} , and end at 1994.

This means that when we lag once, the current X_t series will have to start at 1956 and end at 1995. For practical purposes, this means that when we lag once we loose one observation from the sample. So if we lag ‘s’ periods, we loose ‘s’ observations. Furthermore, we are estimating one extra ‘b’ with every lag.

Finally, after ‘s’ periods we obtain which is known as the *long run, distributed-lag multiplier* provided the sum b exists.

$$\sum_{i=0}^s b_i \tag{3.2}$$

The number of lags, s, may be either finite or infinite. However, in order to avoid explosive values of Y_t , we assume that the b’s have a finite sum: $\sum_{i=0}^{\infty} b_i < \infty$

The average lag is defined as the weighted average of all the lags involved, with weights being the relative size of the respective b coefficients

$$\bar{t} = \frac{\sum_{i=1}^n t_i b_i}{\sum_{i=1}^n b_i}$$

(3.3)

Some examples will illustrate the importance of lagged variables.

Example 1: The consumption function

The model: $C_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \dots + \alpha_n Y_{t-n} + U_t$ is a *distributed lag model* of consumption function. This means that the value of the consumption expenditure (C_t) at any given time depends on the current and past values of the disposable income (Y_t).

Recent versions of the consumption function postulate that the current level of consumption depends on past levels of consumption, due to 'habit persistence', on current income and past levels of income and other factors

$$C_t = f(C_{t-1}, C_{t-2}, \dots, Y_t, Y_{t-1}, Y_{t-2}, \dots, X_{1t}, X_{2t}, \dots)$$

(3.4)

Example 2: The demand for durables

The demand for durables (D_d) depends, among others, on present income (Y_t), past levels of income (Y_{t-s}) which determine the amount saved for the acquisition of the durables, on the stocks of durables or, equivalently, on past acquisitions of durables ($S_{d,t-1}$), prices (P_t), and so on

$$D_d = f(Y_t, Y_{t-1}, \dots, S_{d,t-1}, P_t, \dots)$$

(3.5)

Example 3: The investment function

Investment projects depend on past outputs, on expectations about future profits, on capital stock and other factors

$$I_t = f(X_t, X_{t-1}, X_{t-2}, \dots, K_{t-1}, K_{t-2}, \dots, r_t, \dots)$$

(3.6)

where X = level of output
 π = profit
 K = capital stock
 r = interest rate

Example 4: Demand for nondurables

The habit-persistence is a characteristic of human behavior. The quantity demanded for food, tobacco and other nondurables depends, among others, on the past levels of consumption of these commodities

$$Q_t = f(Q_{t-1}, Y_t, P_t, \dots)$$

(3.7)

3.2.2. Autoregressive models

It should be noted that distributed lag model is not to be confused with autoregressive model.

*If the model includes one or more lagged values of the dependent variable among its explanatory variables, it is called an **autoregressive model**.*

Thus, $Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + u_t$ is an example of distributed-lag model whereas

$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + u_t$ is an example of an autoregressive model.

In the former case the current and lagged values of regressors are explanatory variables while in the latter case the regressors can be both explanatory (current and lagged) and lagged values of the regressand; thus, the latter are also known as **dynamic models** since they portray the time path of the dependent variable in relation to its past value(s).

3.3. The Role of lag in Economics

In economics the dependence of a variable Y (the dependent variable) on another variable(s) X (the explanatory variable) is rarely instantaneous. Very often, Y responds to X with a lapse of time. Such a lapse of time is called a **lag**. Indeed lags are involved in all economic behavior. We live in a dynamic world of continuous adjustment. Clearly an adjustment process takes time, the length of the time period depending on the nature of the particular phenomenon. Yet economic theory is mostly static. In comparative statics where different equilibria are compared, adjustments are assumed instantaneous.

The process of adjustment and the lags involved are scarcely discussed. Yet these lags are of paramount importance for decision making. It is crucial for the government official to know how fast, after how many time periods the economic units will react to changes of various policy variables (instruments). How fast will consumers react to the imposition of a sales tax or a credit squeeze? How fast will firms react to tax concessions and other incentives for investment and innovation? How fast will be the effects of devaluation? How fast will investors react to changes in the interest rates?

Similarly the lags involved in the demand for consumers goods following a change in the 'policy instruments' (price, quality, style, advertising) of a firm are of crucial importance for managerial decisions. Again the micro-economic theory of the firm is mostly static. The lag patterns in the adjustment of buyers to changes in various variables are scarcely mentioned.

In view of the nature of economic behavior any realistic formulation of economic models should involve some lagged variables among the set of explanatory variables. Lagged variables are one way for taking into account the length of time in the adjustment processes of economic behavior and perhaps the most efficient way for rendering them dynamic.

Lagged models have become increasingly popular in applied econometric research. Their refinement has made possible the handling of expectations about future events, if only in a rigid and not very satisfactory way. In short, lagged models offer much flexibility to the formulation of models of economic behavior.

Distributed lag models involve a high degree of empiricism, which is due to the unsatisfactory state of economic theory with regard to the length of the adjustment processes of economic phenomena.

Economic theory, even where it recognizes the importance of time lags, never suggests the precise number of lags that should be included in a function. Rather the pattern of lags is explored and determined from the available sample observations, by adopting an experimental approach involving various lag patterns and then choosing among them the one that gives the best statistical fit.

3.4. The Reasons for Lags

There are three major reasons why lags may occur.

1. **Psychological reasons:** As a result of the force of habit (inertia), people, for example, do not change their consumption habits immediately following a price decrease or an income increase perhaps because the process of change may involve some immediate disutility. Thus, those who become instant millionaires by winning lotteries may not change the lifestyle to which they were accustomed for a long time because they may not know how to react to such windfall gain immediately. Of course, given reasonable time, they may learn to live with their newly acquired fortune. Also, people may not know whether a change is “permanent” or “transitory”. Thus, my reaction to an increase in my income will depend on whether or not the increase is permanent. If it is only a nonrecurring increase and in succeeding periods my income returns to its previous level, I may save the entire increase, whereas someone else in my position might decide to “live it up”.
2. **Technological reason:** suppose, for instance, the price of capital relative to labor declines making substitution of capital for labor economically feasible. Of course, addition of capital takes time (gestation period) moreover, if the drop in price is expected to be temporarily, firms may not rush to substitute capital for labor, especially if they expect that after temporarily drop the price of capital may increase beyond its previous level.
Sometimes, imperfect knowledge also accounts for lags. At present the market for personal computers is glutted with all kinds of computers with varying features and prices. Moreover, since their introduction in the late 1970s, the prices of most personal computers have dropped dramatically. As a result, prospective consumers for the personal computer may hesitate to buy until they have had time to look into the features and prices of all the competing brands. Moreover, they may hesitate to buy in the expectation of further decline in price or innovations.
3. **Institutional reasons:** These reasons also contribute to lags. For example, contractual obligations may prevent firms from switching from one source of labor or raw material to another. As another example, those who have placed funds in long term saving accounts for fixed durations such as one year, three years or seven years are essentially “locked” in even though money market conditions may be such that higher yields are available.

For the reasons just discussed, lag occupies a central role in economics. This is clearly reflected in the short-run–long-run methodology of economics. It is for this reason we say that short-run price or income elasticities are generally smaller (in absolute value) than the corresponding long-run elasticities or that short-run marginal propensity to consume is generally smaller than long-run marginal propensity to consume.

Activity 3.1

All econometric models are essentially dynamic. Argue for or against this statement

3.5. Infinite Distributed Lag

Though so far we assumed a finite number of lags imposed on X_t , some lags may be infinite and it may have its own practical importance. For example, the investment in building highways and roads several decades ago may still have an effect on today’s growth in GNP. In this case, we write equation (3.1) as:

$$Y_t = \alpha + \sum_{i=0}^{\infty} \beta_i X_{t-i} + u_t \quad t = 1, 2, \dots, T$$

(3.8)

There are an infinite number of β_i ’s to estimate with only T observations. This can only be feasible if more structure is imposed on the β_i ’s.

First, we normalize these β_i ’s by their sum, i.e., let $w_i = \beta_i/\beta$ where $\beta = \sum_{i=0}^{\infty} \beta_i$. If all the β_i ’s have the same sign, then the β_i ’s take the sign of β and $0 \leq w_i \leq 1$ for all i , with $\sum_{i=0}^{\infty} w_i = 1$. This means that

the w_i ’s can be interpreted as probabilities.

In fact, Koyck (1954) imposed the geometric lag on the w_i ’s, i.e., $w_i = (1 - \lambda)\lambda^i$ for $i = 0, 1, \dots, \infty$.

Substituting $\beta_i = \beta w_i = \beta(1 - \lambda)\lambda^i$ in (3.8) we get:

$$Y_t = \alpha + \beta(1 - \lambda) \sum_{i=0}^{\infty} \lambda^i X_{t-i} + u_t$$

(3.9)

Equation (3.9) is known as the **infinite distributed lag** form of the Koyck lag. The short-run effect of a unit change in X_t on Y_t is given by $\beta(1 - \lambda)$; whereas the long-run effect of a unit change in X_t on Y_t declines the further back we go in time.

For example, if $\lambda = 1/2$, then $\beta_0 = \beta/2$, $\beta_1 = \beta/4$, $\beta_2 = \beta/8$, etc. Defining $LX_t = X_{t-1}$, as the lag operator, we have $L^i X_t = X_{t-i}$, and (3.9) reduces to:

$$Y_t = \alpha + \beta(1 - \lambda) \sum_{i=0}^{\infty} (\lambda L)^i X_t + u_t = \alpha + \beta(1 - \lambda) X_t / (1 - \lambda L) + u_t$$

(3.10)

gets: $Y_t - \lambda Y_{t-1} = \alpha(1 - \lambda) + \beta(1 - \lambda)X_t + u_t - \lambda u_{t-1}$ or

$Y_t = \lambda Y_{t-1} + \alpha(1 - \lambda) + \beta(1 - \lambda)X_t + u_t - \lambda u_{t-1}$ (3.11) This is the *autoregressive form of the infinite distributed lag*. It is autoregressive because it includes the lagged value of Y_t as an explanatory variable.

NB: Thus, autoregressive models can be derived from distributed lag models.

Note that we have reduced the problem of estimating an infinite number of β_i ’s into estimating λ and β from (3.11). However, OLS would lead to biased and inconsistent estimates, because (3.11) contains a lagged dependent variable as well as serially correlated errors. In fact the error in (3.11) is **a Moving Average** process of order one, i.e., MA(1).

Moving Average Process:

Suppose that U_t is a purely random process with mean zero and variance σ^2 , then the process X_t defined by: $X_t = \beta_0 U_t + \beta_1 U_{t-1} + \dots + \beta_m U_{t-m}$ is called a **moving average process** of order m and denoted by MA(m).

Autoregressive process:

Suppose again that u_t is a purely random process with mean zero and variance δ^2 , then the process X_t given by: $X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \dots + \alpha_r X_{t-r} + U_t$ is called **autoregressive process** of order r and denoted by $AR(r)$.

Autoregressive Moving Average Process (ARMA) is a combination of the AR and MA models.

*Note that: $AR(r)$ models can be estimated using OLS, but $MA(m)$ models are complicated, because we do not observe the lags of the error terms for which we want to estimate the parameters; thus, $ARMA(r,m)$ models are estimated using **Maximum Likelihoods estimation method**.*

We digress at this stage to give two econometric models which would lead to equations resembling (3.11).

I. Adaptive Expectations Model (AEM)

The modeling of expectations is frequently an important and difficult task of the applied economist using time series data. This is especially true in macroeconomics, in that investment, saving, and the demands for assets are all sensitive to expectations about the future.

As a makeshift solution, some models use an indirect technique known as the adaptive expectations process. This involves a simple learning process in which, in each time period, the actual value of the variable is compared with the value that had been expected. If the actual value is greater, the expected value is adjusted upwards for the next period. If it is lower, the expected value is adjusted downwards. The size of the adjustment is hypothesized to be proportional to the discrepancy between the actual and expected value.

Suppose that output Y_t is a function of expected sales X_t^* and that the latter is unobservable, i.e.,

$Y_t = \alpha + \beta X_t^* + u_t$, where expected sales are updated according to the following method

$$X_t^* - X_{t-1}^* = \delta (X_t - X_{t-1}^*) \quad (3.12)$$

that is, expected sales at time t is a weighted combination of expected sales at time $t - 1$ and actual sales at time t .

$$\text{In fact, } X_t^* = \delta X_t + (1 - \delta) X_{t-1}^* \quad (3.13)$$

Equation (3.12) is also an **error learning model**, where one learns from past experience and adjusts expectations after observing current sales. Using the lag operator L , (3.14) can be rewritten as

$$X_t^* = \delta X_t / [1 - (1 - \delta)L].$$

Substituting this last expression in the above relationship, we get:

$$Y_t = \alpha + \beta \delta X_t / [1 - (1 - \delta)L] + u_t \quad (3.14)$$

Multiplying both sides of (3.14) by $[1 - (1 - \delta)L]$, we get:

$$Y_t - (1 - \delta)Y_{t-1} = \alpha[(1 - (1 - \delta))] + \beta \delta X_t + u_t - (1 - \delta)u_{t-1} \quad (3.15)$$

NB: (3.15) looks exactly like (3.11) with $\lambda = (1 - \delta)$.

II. Partial Adjustment Model (PAM)

In the partial adjustment model it is assumed that the behavioral equation determines the *desired* (or "target") value, Y_t^* , of the dependent variable, rather than the actual value, Y_t .

Under this model there is a cost of being out of equilibrium and a cost of adjusting to that

$$\text{equilibrium, i.e., Cost} = a(Y_t - Y_t^*)^2 + b(Y_t - Y_{t-1})^2 \quad (3.16)$$

where Y_t^* is the target or equilibrium level for Y , whereas Y_t is the current level of Y .

The first term of (3.16) gives a quadratic loss function proportional to the distance of Y_t from the equilibrium level Y_t^* . The second quadratic term represents the cost of adjustment. Minimizing this quadratic cost function with respect to Y , we get $Y_t = \gamma Y_t^* + (1-\gamma)Y_{t-1}$, where $\gamma = a/(a+b)$.

Note that if the cost of adjustment was zero, then $b = 0$, $\gamma = 1$, and the target is reached immediately.

However, there are costs of adjustment, especially in building the desired capital stock.

$$\text{Hence, } Y_t = \gamma Y_t^* + (1 - \gamma)Y_{t-1} + u_t \quad (3.17)$$

where we made this relationship stochastic.

If the true relationship is $Y_t^* = \alpha + \beta X_t$, then from (3.17)

$$Y_t = \gamma\alpha + \gamma\beta X_t + (1 - \gamma)Y_{t-1} + u_t \quad (3.18)$$

and this looks like (3.11) with $\lambda = (1 - \gamma)$, except for the error term, which is not necessarily MA(1) with the Moving Average parameter λ .

Draw backs of Lagged models

Lagged values of the variables are important explanatory variables in most economic relationships, because economic behavior in anyone period is to a great extent determined by past experience and past patterns of behavior.

However, there are some practical problems we may encounter using lag models. If we lag 's' periods, we lose 's' observations as we have stated earlier. Furthermore, we are estimating one extra 'b' with every lag. Therefore, there is double jeopardy with respect to loss of degrees of freedom. The number of observations fall (because we are lagging the same series), and the number of parameters to be estimated increase with every lagging variable introduced. Besides the loss of degrees of freedom, the regressors in (3.1) are likely to be highly correlated with each other. In fact most economic time series are usually trended and very highly correlated with their lagged values. This introduces the problem of multicollinearity among the regressors and as we saw in Chapter 1, the higher the multicollinearity among these regressors, the lower is the reliability of the regression estimates.

In this model, OLS is still BLUE because the classical assumptions are still satisfied. All we have done in (3.1) is introduce the additional regressors ($X_{t-1}, \dots, X_{t-s}$). These regressors are uncorrelated with the disturbances since they are lagged values of X_t , which are by assumption not correlated with u_t for every t .

3.6. Estimation of Lag Models

3.6.1. Estimation of Autoregressive Models

Both the AEM and the PAM give equations resembling the autoregressive form of the infinite distributed lag. In all cases, we end up with a lagged dependent variable and an error term that is either Moving Average of order one as in (3.11), or just classical or autoregressive as in (3.18). In this section we study the testing and estimation of such autoregressive or dynamic models.

If there is a Y_{t-1} in the regression equation and the u_t 's are classical disturbances, as may be the case in equation (3.18), then Y_{t-1} is said to be *contemporaneously uncorrelated with the disturbance term* u_t . In fact, the disturbances satisfy assumptions 1-4 of classical linear regression and $E(Y_{t-1} u_t) = 0$ even though $E(Y_{t-1} u_{t-1}) \neq 0$. In other words, Y_{t-1} is not correlated with the current disturbance u_t but it is correlated with the lagged disturbance u_{t-1} . In this case, as long as the disturbances are not serially correlated, OLS will be biased, but remains **consistent and asymptotically efficient**. This case is unlikely with economic data given that most macro time-series variables are highly trended. More likely, the u_t 's are serially correlated. In this case, OLS is **biased and inconsistent**.

Intuitively, Y_t is related to u_t , so Y_{t-1} is related to u_{t-1} . If u_t and u_{t-1} are correlated, then Y_{t-1} and u_t are correlated. This means that one of the regressors, lagged Y , is correlated with u_t and we have the problem of **endogeneity**.

A model with a lagged dependent variable and an autoregressive error term is estimated using

Instrumental Variables (IV). This method will be studied extensively in your post graduate studies.

In short, the IV method corrects for the correlation between Y_{t-1} and the error term by replacing Y_{t-1} with its predicted value $\hat{Y}_{t-1}$. The latter is obtained by regressing Y_{t-1} on some exogenous variables say a set of Z 's, which are called a set of instruments for Y_{t-1} . Since these variables are exogenous and uncorrelated with u_t , $\hat{Y}_{t-1}$ will not be correlated with u_t .

3.6.2. Estimation of Distributed Lag Models

Granted that distributed-lag models play a highly useful role in economics, how does one estimate such models?

Specifically, suppose we have the following distributed lag model in one explanatory variable:

$$Y_t = \alpha_0 + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \dots + \alpha_s X_{t-s} + u_t \quad (3.19)$$

In (3.19) the length of the lag, that is, how far back into the past we want to go hasn't been defined. Such a model is called an *infinite (lag) model*, whereas models with specified lags are called a *finite (lag) distributed-lag model*.

How do we estimate α and s in (3.19)? We may adopt two approaches:

- A. *Ad Hoc estimation of distributed-lag models*
- B. *A priori restriction on α 's by assuming that the α 's follow some systematic pattern.*

A. Ad Hoc estimation of distributed lag models

Since the explanatory variable X_t is assumed to be non-stochastic (or at least uncorrelated with the disturbance term U_t), X_{t-1} , X_{t-2} and so on, are non-stochastic too. Therefore in principle, OLS can be applied to the above model (3.19).

This is the approach taken by Alt and Tinbergen. They suggest that to estimate (3.19) one may proceed *sequentially*; that is:

First, regress/estimate Y on X_t , then Y_t on X_t, X_{t-1} , then Y_t on X_t, X_{t-1}, X_{t-2} , and so on.

This sequential procedure will continue until the regression coefficients of the lagged variables start becoming statistically insignificant and/or the coefficient of at least one of the variables changes signs from positive to negative vice versa.

Consider the following hypothetical example.

$$\hat{Y}_t = 8.37 + 0.17X_t$$

$$\hat{Y}_t = 8.27 + 0.111X_t + 0.06X_{t-1}$$

$$\hat{Y}_t = 8.27 + 0.109X_t + 0.071X_{t-1} + 0.055X_{t-2}$$

$$\hat{Y}_t = 8.32 + 0.108X_t + 0.063X_{t-1} + 0.022X_{t-2} + 0.020X_{t-3}$$

Alt choose the second regression as the “best” one because in the last two equation the sign of X_{t-2} was not stable and in the last equation the sign of X_{t-3} was negative, which may be difficult to interpret economically.

Although seemingly straight forward, ad hoc estimation suffers from many drawbacks, such as the following:

- There is no guide as to what is the maximum lag length
- As one estimates successive lags, there are fewer degrees of freedom left, making statistical inference somewhat unstable
- More importantly, in economic time series data, successive values (lags) tend to be highly correlated; hence multicollinearity rears its hideous head.

As noted in Chapter 1, multicollinearity leads to imprecise estimation; that is, the standard errors tend to be large in relation to the estimated coefficients. As a result, based on the routinely computed t ratios, we may tend to declare (erroneously), that a lagged coefficient(s) is statistically insignificant.

- The sequential search for the lengths of lags opens the researcher to the charge of *data mining*.

In view of the preceding problems, the ad hoc estimation procedure has very little to recommend it. Some prior or theoretical considerations are brought to bear upon the various 's if we are to make headway with the estimation problem.

B. Methods based on A priori restriction on β 's

Method I: The koyck approach to distributed lag models

Koyck has proposed an ingenious method of estimating distributed-lag models. Suppose we start with the infinite lag distributed-lag model (3.19).

Assuming that the β 's are all of the same sign, Koyck assumes that they decline geometrically as follows.

$$\beta_k = \beta_0 \lambda^k \qquad k = 0, 1, \dots$$

(3.20)

where λ , such that $0 < \lambda < 1$, is known as the *rate of decline*, or *decay*, of the distributed lag and where $1 - \lambda$ is known as the *speed of adjustment*.

What (3.20) postulates is that each successive β coefficient is numerically less than each preceding β (this statement follows since $\lambda < 1$), implying that as one goes back into the distant past, the effect of that lag on Y_t becomes progressively smaller, a quite plausible assumption. After all, current and recent past incomes are expected to affect current consumption expenditure more heavily than income in the distant past.

In order to reduce the number of lags in the given distributed lag model, the present model makes an assumption that the impact of explanatory variables (on the dependent variable) in the most distant past is less than what is in most recent periods.

More specifically, the Koyck lag formulation assumes that the weights (impacts) are declining continuously.

Assume the original model is:

$$Y_t = \alpha_0 + \alpha_1 X_t + \alpha_2 X_{t-1} + \alpha_3 X_{t-2} + \dots + \alpha_s X_{t-s} + U_t$$

(3.21)

$$U_t \sim N(0, \sigma^2)$$

$$(u_i - u_j) = 0$$

$$(u_i - x_i) = 0$$

According to Koyck:

$$\alpha_i = \beta_i \lambda^i$$

$$\beta_1 = 0, \quad \beta_2 = 0, \quad \dots$$

As we saw in our previous section, the partial sums of the standardized β_i tell us the proportion of the long-run, or total, impact felt by a certain time period. In practice, though, the **mean** or **median lag** is often used to characterize the nature of the lag structure of a distributed lag model.

The Median Lag

The median lag is the time required for the first half, or 50 percent, of the total change in Y following a unit sustained change in X . For the Koyck model, the median lag is as follows:

Koyck model: Median lag = $-\log 2 / \log \lambda$

(3.22)

Thus, if $\lambda = 0.2$ the median lag is 0.4306, but if $\lambda = 0.8$ the median lag is 3.1067. Verbally, in the former case 50 percent of the total change in Y is accomplished in less than half a period, whereas in the latter case it takes more than 3 periods to accomplish the 50 percent change. But this contrast should not be surprising, for as we know, the higher the value of λ the lower the speed of adjustment, and the lower the value of λ the greater the speed of adjustment.

The Mean Lag

Provided all β_k are positive, the mean, or average, lag is defined as:

$$\bar{\lambda} = \frac{\sum_{k=0}^{\infty} k \beta_k}{\sum_{k=0}^{\infty} \beta_k} \quad (3.23)$$

which is simply the weighted average of all the lags involved, with the respective β coefficients serving as weights. In short, it is a **lag-weighted average** of time.

For the Koyck model the mean lag is Koyck model: Mean lag = $\lambda / (1 - \lambda)$ (3.24)

Thus, if $\lambda = 0.12$, the mean lag is 1.

From the preceding discussion it is clear that the median and mean lags serve as a summary measure of the speed with which Y responds to X .

Note the following features of the Koyck scheme:

1. Assuming non-negative values for λ , Koyck rules out the λ 's from changing sign,
2. by assuming $\lambda < 1$, lesser weight is assigned to the decline λ 's than current one; and
3. The sum of the β 's, which gives the long run multiplier, is a finite amount in Koyck scheme and is.

$$\sum_{i=0}^{\infty} \lambda^i = \frac{1}{1 - \lambda} \quad (3.25)$$

Substituting the values of λ 's in the original model (3.19) we obtain:

$$Y_t = \alpha_0 + \alpha_1 X_t + \lambda(Y_{t-1} - \alpha_0 - \alpha_1 X_{t-1}) + U_t \quad (3.26)$$

As it stands, the model is still not amenable to easy estimation since a large (literally infinite) number of parameters remain to be estimated and the parameter λ enters in a highly nonlinear form: Strictly speaking, the method of linear (in the parameters) regression analysis cannot be applied to such a model.

Lagging by one period and multiply by λ we get:

$$\lambda Y_{t-1} = \lambda \alpha_0 + \lambda \alpha_1 X_{t-1} + \lambda^2(Y_{t-2} - \alpha_0 - \alpha_1 X_{t-2}) + \lambda U_{t-1} \quad (3.27)$$

Subtracting (3.27) from (3.26) and rearranging we obtain:

$$Y_t - \lambda Y_{t-1} = (1 - \lambda)\alpha_0 + (1 - \lambda)\alpha_1 X_t + (U_t - \lambda U_{t-1})$$

Let $V_t = (1 - \lambda)U_t - \lambda U_{t-1}$

$$Y_t^* = (1 - \lambda)X_t + V_t \quad (3.28)$$

where $V_t = (u_t - \lambda u_{t-1})$, a **moving average of u_t and u_{t-1}** .

The above procedure of transformation is known as **koyck transformation**. If the Koyck hypothesis concerning the lag scheme and assumptions concerning V_t are accepted, ordinary least square can be applied to obtain estimators of α_0 and α_1 . From these estimates, the estimates of the original parameters $\alpha_0, \alpha_1, \alpha_2, \dots, \alpha_k$ can be easily obtained through:

$$\hat{\alpha}_i = \frac{\sum_{t=1}^n V_t^* X_{t,i}}{\sum_{t=1}^n V_t^* X_{t,i}^2}$$

Comparing (3.25) with (3.19), we see the tremendous simplification accomplished by Koyck. Whereas before we had to estimate α and an infinite number of β 's, we now have to estimate only three unknowns: α , β_0 , and λ .

Now there is no reason to expect multicollinearity. In a sense multicollinearity is resolved by replacing $X_{t-1}, X_{t-2}, \dots$, by a single variable, namely, Y_{t-1} .

However, the following features of the Koyck transformation may be taken into account:

1. We started with a distributed-lag model but ended up with an autoregressive model because Y_{t-1} appears as one of the explanatory variables. This transformation shows how one can “convert” a distributed-lag model into an autoregressive model.
2. The appearance of Y_{t-1} is likely to create some statistical problems. Y_{t-1} , like Y_t , is stochastic, which means that we have a stochastic explanatory variable in the model. Recall that the classical least-squares theory is predicated on the assumption that the explanatory variables either are non-stochastic or, if stochastic, are distributed independently of the stochastic disturbance term. The lagged variable Y_{t-1} is not also independent of the error term V_t i.e. $(U_t Y_{t-1}) \neq 0$ this is because Y_t is directly dependent on V_t . Similarly Y_{t-1} on V_{t-1} . But since V_t and V_{t-1} are not independent, Y_{t-1} will obviously be related to V_t . Hence, we must find out if Y_{t-1} satisfies this assumption.
3. In the original model (3.19) the disturbance term was u_t , whereas in the transformed model it is $v_t = (u_t - \lambda u_{t-1})$. The statistical properties of v_t depend on what is assumed about the statistical properties of u_t , for, as shown later, if the original u_t 's are serially uncorrelated, the v_t 's are serially correlated.

It can be seen as under

$$\begin{aligned} (V_t V_{t-1}) &= (U_t - U_{t-1})(U_{t-1} - U_{t-2}) \\ &= (U_t U_{t-1} - U_t U_{t-2} - U_{t-1}^2 + U_{t-1} U_{t-2}) \\ &= (U_{t-1}^2) \\ &= \frac{1}{n} \sum u^2 \neq 0 \end{aligned}$$

Therefore, we may have to face up to the serial correlation problem in addition to the stochastic explanatory variable Y_{t-1} .

4. The presence of lagged Y violates one of the assumptions underlying the Durbin-Watson d test. Therefore, we will have to develop an alternative to test for serial correlation in the presence of lagged Y . One alternative is the **Durbin h test**, which we discussed in previously.

Due to these problems, the Koyck transformation of the distributed lag model will give rise to biased and inconsistent estimates. In addition to these estimation problem, the Koyck hypothesis is quite restrictive in the sense that it assumes that impact of past periods decline successively in a specific way.

Method II: Rationalization of the Koyck Model

A. Adaptive Expectations (AE) Model

The Koyck model is ad hoc since it was obtained by a purely algebraic process; it is devoid of any theoretical foundation. But this gap can be filled if we start from a different perspective.

Illustration:

Suppose we postulate the following model:

$$Y_t = X_t^* + u_t \quad (3.29)$$

Where: Y = demand for money (real cash balances)

X^* = equilibrium, optimum, expected long-run or normal rate of interest

u = error term

Equation (3.29) postulates that the demand for money is a function of expected (in the sense of anticipation) rate of interest.

Since the expectational variable X^* is not directly observable, let's propose the following hypothesis (the hypothesis we made so far) about how expectations are formed as:

$X_t^* = \lambda X_t + (1 - \lambda) X_{t-1}^*$, where λ , such that $0 < \lambda < 1$, is known as the coefficient of expectation. The hypothesis is known as the *adaptive expectation*, *progressive expectation*, or *error learning hypothesis*, popularized by Cagan and Friedman.

The hypothesized equation states that expectations are revised each period by a fraction of the gap between the current value of the variable and its previous expected value. Thus, for our model this would mean that expectations about interest rates are revised each period by a fraction of the discrepancy between the rate of interest observed in the current period and what its anticipated value had been in the previous period. Another way of stating this would be to write the hypothesized equation as:

$$X_t^* = \lambda X_t + (1 - \lambda) X_{t-1}^* \quad (3.30)$$

which shows that the expected value of the rate of interest at time t is a weighted average of the actual value of the interest rate at time ' t ' and its value expected in the previous period, with weights of ' λ ' and ' $1 - \lambda$ ', respectively.

- ❖ If $\lambda = 1$, $X_t^* = X_t$, meaning that expectations are realized immediately and fully, that is, in the same time period.
- ❖ If, on the other hand, $\lambda = 0$, $X_t^* = X_{t-1}^*$, meaning that expectations are static, that is, conditions prevailing today will be maintained in all subsequent periods. Expected future values then become identified with current values." Substituting (3.27) into (3.26), we obtain

$$Y_t = \lambda X_t + (1 - \lambda) X_{t-1}^* + u_t$$

$$X_t^* = \lambda X_t + (1 - \lambda) X_{t-1}^* + u_t \quad (3.31)$$

Now, lag equation (3.29) by one period, multiply it by $1 - \lambda$, and subtract the product from (3.31).

After simple algebraic manipulations, we obtain:

$$\lambda Y_t = \lambda X_t - (1 - \lambda) Y_{t-1} + u_t - (1 - \lambda) u_{t-1}$$

$$\lambda Y_t = (1 - \lambda) Y_{t-1} + v_t \quad (3.32)$$

where $v_t = u_t - (1 - \lambda) u_{t-1}$.

Let us note the difference between (3.29) and (3.32). In the former, β_1 measures the average response of Y to a unit change in X^* , the equilibrium or long-run value of X . In (3.29), on the other hand, β_1 measures the average response of Y to a unit change in the actual or observed value of X . These responses will not be the same unless, of course, $\beta_1 = 1$, that is, the current and long-run values of X are the same. In practice, we first estimate (3.32). Once an estimate of β_1 is obtained from the coefficient of lagged Y , we can easily compute β_1 by simply dividing the coefficient of X_t (β_1) by β_1 .

The similarity between the adaptive expectation model and the Koyck model should be readily apparent although the interpretations of the coefficients in the two models are different.

Note that like the Koyck model, the adaptive expectations model is autoregressive and its error term is similar to the Koyck error term.

B. Stock Adjustment, or Partial Adjustment Model

The adaptive expectation model is one way of rationalizing the Koyck model. Another rationalization is provided by Marc Nerlove in the so-called stock adjustment or partial adjustment model. We developed this model in our *previous section*, but for the sake of elaboration we will try to discuss it with the use of some economic theories.

To illustrate this model, consider the **flexible accelerator model** of economic theory, which assumes that there is equilibrium, optimal, desired, or long-run amount of capital stock needed to produce a given output under the given state of technology, rate of interest, etc. For simplicity assume that this desired level of capital K_t^* is a linear function of output X as follows:

$$K_t^* = \beta_0 + \beta_1 X_t + u_t \quad (3.33)$$

Since the desired level of capital is not directly observable, Nerlove postulates the following hypothesis, known as the partial adjustment, or stock adjustment, hypothesis:

$$K_t - K_{t-1} = \alpha (K_t^* - K_{t-1}) \quad (3.34)$$

Where, α such that $0 < \alpha < 1$, is known as the coefficient of adjustment and where $K_t - K_{t-1}$ is the actual change in capital stock and $(K_t^* - K_{t-1})$ is the desired change. Since $K_t - K_{t-1}$, the change in capital stock between two periods, it is nothing but investment, (3.31) can alternatively be written as:

$$I_t = (Y_t^* - Y_{t-1}) \quad (3.35)$$

Where, I_t = Investment in time period t .

Equation (3.34) postulates that the actual change in capital stock (investment) in any given time period t is some fraction of the desired change for that period. If $\alpha = 1$, it means that the actual stock of capital is equal to the desired stock; that is, actual stock adjusts to the desired stock instantaneously (in the same period). However, if $\alpha = 0$, it means that nothing changes since actual stock at time t is the same as that observed in the previous time period. Typically, α is expected to

lie between these extremes since adjustment to the desired stock of capital is likely to be incomplete because of rigidity, inertia, contractual obligations, etc. – hence the name partial adjustment model.

Note that the adjustment mechanism (3.34) alternatively can be written as:

$$K_t = K_t^* - (1 - \alpha)K_{t-1} \quad (3.36)$$

showing that the observed capital stock at time t is a weighted average of the desired capital stock at that time and the capital stock existing in the previous time period, and $(1 - \gamma)$ being the weights. Now substitution of (3.33) into (3.36) gives:

$$K_t = (\gamma + (1 - \gamma)X_t) + (1 - \gamma)K_{t-1} \quad (3.37)$$

This model is called the **Partial Adjustment Model**.

Since (3.33) represents the *long-run*, or equilibrium, demand for capital stock, (3.37) can be called the *short-run demand function for capital stock* since in the short run the existing capital stock may not necessarily be equal to its long-run level.

Once we estimate the short-run function (3.37) and obtain the estimate of the adjustment coefficient (from the coefficient of K_{t-1}), we can easily derive the long-run function by simply dividing γ and $(1 - \gamma)$ by γ and omitting the lagged K term, which will then give (3.33).

The partial adjustment model resembles both the Koyck and adaptive expectation models in that it is **autoregressive**. But it has a much simpler disturbance term: the original disturbance term u_t multiplied by a constant γ . But bear in mind that although similar in appearance, the adaptive expectation and partial adjustment models are conceptually very different.

The former is based on uncertainty (about the future course of prices, interest rates, etc.), whereas the latter is due to technical or institutional rigidities, inertia, cost of change, etc. However, both of these models are theoretically much sounder than the Koyck model.

Since in appearance the adaptive expectations and partial adjustment models are indistinguishable, the γ coefficient of the adaptive expectations model can also be interpreted as the δ coefficient of the stock adjustment model if we assume that the latter model is operative in the present case.

The important point to keep in mind is that since *Koyck*, *adaptive expectations*, and *stock adjustment models* – apart from the difference in the appearance of the error term – yield the same final estimating model.

Method III. Combination of Adaptive Expectations and Partial Adjustment Models

Consider the following model:

$$Y_t^* = \gamma X_t^* + (1 - \gamma)Y_{t-1} + u_t \quad (3.38)$$

where Y_t^* = desired stock of capital and X_t^* = expected level of output.

Since both Y_t^* and X_t^* are not directly observable, one could use the partial adjustment mechanism for Y_t^* and the adaptive expectations model for X_t^* to arrive at the following estimating equation.

$$Y_t - \gamma X_t = (1 - \gamma)Y_{t-1} + (1 - \gamma)(1 - \gamma)Y_{t-2} + \gamma u_t \quad (3.39)$$

where $v_t = \gamma u_t + (1 - \gamma)u_{t-1}$.

This model too is autoregressive, the only difference from the purely adaptive expectations model being that Y_{t-2} appears along with Y_{t-1} as an explanatory variable. Like Koyck and the AE models, the error term in (3.39) follows a **moving average process**.

Another feature of this model is that although the model is linear in the β 's, it is nonlinear in the original parameters.

A celebrated application of (3.38) has been Friedman's *permanent income hypothesis*, which states that "*permanent*" or *long-run consumption is a function of "permanent" or long-run income*.

The estimation of (3.39) presents the same estimation problems as the Koyck's or the AE model in that all these models are autoregressive with similar error structures.

Summary of Estimation of Autoregressive Models

From our discussion so far we have the following three models:

Koyck

$$Y_t = \alpha(1 - \lambda) + \beta_0 X_t + \lambda Y_{t-1} + (u_t - \lambda u_{t-1})$$

Adaptive expectation

$$Y_t = \gamma \beta_0 + \gamma \beta_1 X_t + (1 - \gamma) Y_{t-1} + [u_t - (1 - \gamma) u_{t-1}]$$

Partial adjustment

$$Y_t = \delta \beta_0 + \delta \beta_1 X_t + (1 - \delta) Y_{t-1} + \delta u_t$$

All these models have the following common form:

$$Y_t = \alpha_0 + \alpha_1 X_t + \alpha_2 Y_{t-1} + v_t$$

that is, they are all autoregressive in nature.

Therefore, we must now look at the estimation problem of such models, because the classical least-squares may not be directly applicable to them. **The reason is two-fold: the presence of stochastic explanatory variables and the possibility of serial correlation.**

Now, as noted previously, for the application of the classical least-squares theory, it must be shown that the stochastic explanatory variable Y_{t-1} is distributed independently of the disturbance term v_t . To determine whether this is so, it is essential to know the properties of v_t . If we assume that the original disturbance term u_t satisfies all the classical assumptions, such as $E(u_t) = 0$, $\text{var}(u_t) = \sigma^2$ (the assumption of homoscedasticity), and $\text{cov}(u_t, u_{t+s}) = 0$ for $s \neq 0$ (the assumption of no autocorrelation), v_t may not inherit all these properties.

Consider, for example, the error term in the Koyck model, which is $v_t = (u_t - \lambda u_{t-1})$.

Given the assumptions about u_t , we can easily show that v_t is serially correlated because $E(v_t v_{t-1}) = -\lambda \sigma^2$ which is nonzero (unless λ happens to be zero). And since Y_{t-1} appears in the Koyck model as an explanatory variable, it is bound to be correlated with v_t (via the presence of u_{t-1} in it). As a matter of fact, it can be shown that $\text{cov}[Y_{t-1}, (u_t - \lambda u_{t-1})] = -\lambda \sigma^2$ which is the same as previous proof. The reader can verify that the same holds true of the adaptive expectations model.

What is the implication of the finding that in the Koyck model as well as the adaptive expectations model the stochastic explanatory variable Y_{t-1} is correlated with the error term v_t ?

As noted previously, *if an explanatory variable in a regression model is correlated with the stochastic disturbance term, the OLS estimators are not only biased but also not even consistent; that is, even if the sample size is increased indefinitely, the estimators do not approximate their true population values. Therefore, estimation of the Koyck and adaptive expectation models by the usual OLS procedure may yield seriously misleading results.*

However, the partial adjustment model is different. In this model $v_t = \delta u_t$, where $0 < \delta \leq 1$. Therefore, if u_t satisfies the assumptions of the classical linear regression model given previously, so will δu_t . Thus, OLS estimation of the partial adjustment model will yield consistent estimates although the estimates tend to be biased (in finite or small samples). Intuitively, the reason for consistency is this: Although Y_{t-1} depends on u_{t-1} and all the previous disturbance terms, it is not related to the current error term u_t .

Therefore, as long as u_t is serially independent, Y_{t-1} will also be independent or at least uncorrelated with u_t , thereby satisfying an important assumption of OLS, namely, non-correlation between the explanatory variable(s) and the stochastic disturbance term.

Activity 3.2: Check whether the following question is correct or wrong and also substantiate your answer with possible justifications.

The Koyck model will not make much sense if some of the distributed lag coefficients are positive and some negative.

Method IV: The Almon approach to Distributed Lag Models

Although used extensively in practice, the Koyck distributed-lag model is based on the assumption that the β coefficients decline geometrically as the lag lengthens. This assumption may be too restrictive in some situations.

The Almon lag model possesses two advantages over the Koyck procedure:

First, it does not violate any of the ordinary least square basic assumptions concerning the disturbance term.

Second, it is far more flexible than Koyck method in terms of the form of lag scheme. It is because; this method does not hypothesize any form of lag in advance.

This model assumes that any pattern of lag scheme among 's can be described by polynomial. This idea is based on a theorem in mathematics known as *Weierstrass's theorem*, which states that under general conditions a curve may be approximated by a polynomial whose degree is one more than the number of turning points in the curve. Suppose that the 's in a given distributed lag model are expected to decrease first, then increase and again decrease

Suppose our original model to be estimated is:

$$Y_t = \alpha_0 + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \dots + \alpha_s X_{t-s} + U_t \quad (3.37)$$

assuming the above polynomial: $\alpha_i = a_0 + a_1 i + a_2 i^2 + a_3 i^3$ where a_0, a_1, a_2 , and a_3 are parameters to be estimated.

We are, now in a position to obtain all 's by setting i equal to the value of the subscript of the particular coefficient.

$$\begin{aligned} \alpha_0 &= a_0 \\ \alpha_1 &= a_0 + a_1 + a_2 + a_3 \\ \alpha_2 &= a_0 + 2a_1 + 4a_2 + 8a_3 \\ \alpha_3 &= a_0 + 3a_1 + 9a_2 + 27a_3 \\ \alpha_k &= a_0 + ka_1 + k^2 a_2 + k^3 a_3 \end{aligned}$$

Naturally therefore, what is needed to be estimated is only four parameters of the polynomial function: $a_0, a_1, a_2, \text{ and } a_3$.

Obtaining the values of $a_0, a_1, a_2, \text{ and } a_3$, we are able to estimate all the parameters of the original distributed lag model.

Substituting values of λ 's in the original model,

$$Y_t = a_0 + a_1 X_{t-1} + a_2 X_{t-2} + a_3 X_{t-3} + U_t$$

$$Y_t = a_0 + a_1 \sum_{i=1}^k \lambda_i^k X_{t-i} + a_2 \sum_{i=1}^k \lambda_i^{k-1} X_{t-i} + a_3 \sum_{i=1}^k \lambda_i^{k-2} X_{t-i} + U_t$$

$$Y_t = a_0 + a_1 w_1 + a_2 w_2 + a_3 w_3 + U_t \quad (3.38)$$

where: $w_1 = \sum_{i=1}^k \lambda_i^k X_{t-i}$, $w_2 = \sum_{i=1}^k \lambda_i^{k-1} X_{t-i}$, $w_3 = \sum_{i=1}^k \lambda_i^{k-2} X_{t-i}$

This is the final form (or transformed form) of Almon Lag model. We can now apply OLS method to estimate $\hat{a}_0, \hat{a}_1, \hat{a}_2, \text{ and } \hat{a}_3$ to obtain $\hat{\beta}$ s in the original form.

Note that: v_t remains in its original form.

3.7. The Granger Causality Test

Although regression analysis deals with the dependence of one variable on other variables, it does not necessarily imply **causation**. In other words, the existence of a relationship between variables does not prove causality or the direction of influence.

But in regressions involving time series data, the situation may be somewhat different because time does not run backward. That is, if event A happens before event B , then it is *possible* that A is causing B . However, it is not *possible* that B is causing A . In other words, events in the past can cause events to happen today. Future events cannot.

This is roughly the idea behind the so-called **Granger causality test**. But it should be noted clearly that the question of causality is deeply philosophical with all kinds of controversies. At one extreme are people who believe that “everything causes everything,” and at the other extreme are people who deny the existence of causation whatsoever.

To explain the Granger test, we will consider the often asked question in macroeconomics: Is it GDP that “causes” the money supply M ($GDP \rightarrow M$) or is it the money supply M that causes GDP ($M \rightarrow GDP$), where the arrow points to the direction of causality.

The Granger causality test assumes that the information relevant to the prediction of the respective variables, GDP and M , is contained solely in the time series data on these variables. The test involves estimating the following pair of regressions:

$$GDP_t = \alpha_0 + \alpha_1 GDP_{t-1} + \alpha_2 GDP_{t-2} + \alpha_3 GDP_{t-3} + u_{1t} \quad (3.39)$$

$$M_t = \beta_0 + \beta_1 M_{t-1} + \beta_2 M_{t-2} + \beta_3 M_{t-3} + u_{2t} \quad (3.40)$$

Where: it is assumed that the disturbances u_{1t} and u_{2t} are uncorrelated. Since we have two variables,

we are dealing with **bilateral causality**.

Equation (3.39) postulates that current GDP is related to past values of itself as well as that of M , and (3.40) postulates a similar behavior for M .

Note that these regressions can be cast in growth forms, $\dot{GDP}$ and $\dot{M}$, where a dot over a variable indicates its growth rate. We now distinguish **four** cases by regressing/ estimating the above two equations and check for the significance of respective coefficients:

1. **Unidirectional causality from M to GDP** is indicated if the estimated coefficients on the lagged M in (3.39) are statistically different from zero as a group (i.e., $\sum_{i=1}^m \alpha_i \neq 0$) and the set of estimated coefficients on the lagged GDP in (3.40) is not statistically different from zero (i.e., $\sum_{j=1}^n \beta_j = 0$).
2. Conversely, **unidirectional causality from GDP to M** exists if the set of lagged M coefficients in (3.39) is not statistically different from zero (i.e., $\sum_{i=1}^m \alpha_i = 0$) and the set of the lagged GDP coefficients in (3.40) is statistically different from zero (i.e., $\sum_{j=1}^n \beta_j \neq 0$).
3. **Feedback, or bilateral causality**, is suggested when the sets of M and GDP coefficients are statistically significantly different from zero in both regressions.
4. Finally, **independence** is suggested when the sets of M and GDP coefficients are not statistically significant in both the regressions.

More generally, since the future cannot predict the past, if variable X (Granger) causes variable Y , then changes in X should *precede* changes in Y .

Therefore, in a regression of Y on other variables (including its own past values) if we include past or lagged values of X and it significantly improves the prediction of Y , then we can say that X (Granger) causes Y . A similar definition applies if Y (Granger) causes X .

The steps involved in implementing the Granger causality test are as follows. We illustrate these steps with the GDP-money example given in equation (3.39):

- Step 1:** Regress/ Estimate current GDP on all lagged GDP terms and other variables, if any, but *do not* include the lagged M variables in this regression. This is the restricted regression. From this regression obtain the restricted residual sum of squares, RSSR.
- Step 2:** Now run the regression including the lagged M terms. In the language of Econometrics, this is the unrestricted regression.

From this regression obtain the unrestricted residual sum of squares, RSSUR.

Step 3: The null hypothesis is $H_0: \sum_{i=1}^m \alpha_i = 0$, that is, lagged M terms do not belong in the regression.

Step 4: To test this hypothesis, we apply the F test given by (3.41), namely,

$$F = \frac{(RSS - RSSUR)/m}{RSSUR/(n - k)}$$

(3.41)

which follows the F distribution with m and $(n-k)$ degree of freedom for numerator and denominator, respectively. In the present case m is equal to the number of lagged M terms and k is the number of parameters estimated in the unrestricted regression.

- Step 5: Decision:** If the computed F value exceeds the critical F value at the chosen level of significance, we reject the null hypothesis, in which case the lagged M terms belong in the regression. This is another way of saying that M causes GDP.
- Step 6:** Steps 1 to 5 can be repeated to test the model (3.40), that is, whether GDP causes M or not.

Before we illustrate the Granger causality test, there are several things that need to be noted:

1. It is assumed that the two variables, GDP and M , are *stationary*.

Stationarity is the most vital and common assumption in *Time series analysis*. A stationary process defined as a series/sequence without trend and seasonality as well as constant mean, variance, and autocorrelation structure over time. If either of these changes over time, the series is **non-stationary**.

Empirics and visual inspection tells us that most macroeconomic variables are non-stationary (either consistently increasing or decreasing on average). If we apply OLS for non-stationary series, the estimated parameters become significant and contain trend (a pattern that exhibits a tendency either to grow or to decrease fairly steadily over time); which is called **spurious regression** which is resulted in wrong conclusions and policy implications.

In standard linear regression modeling, stationarity is important to ensure consistent coefficient estimates and avoid spurious results. Thus, before estimation, we have to assure stationarity of all variables.

Sometimes taking the first differences of the variables makes them stationary, if they are not already stationary in the level form.

2. The number of lagged terms to be introduced in the causality tests is an important practical question. It should be added that the *direction of causality may depend critically on the number of lagged terms included*.
3. We have assumed that the error terms entering the causality test are uncorrelated. If this is not the case, appropriate transformation, may have to be taken.
4. Since our interest is in testing for causality, one need not present the estimated coefficients of models (3.39) and (3.40) explicitly, rather we just stick to the results of the *F* test given in (3.41).

Example

R. W. Hafer used the Granger test to find out the nature of causality between GNP (rather than GDP) and *M* for the United States for the period 1960-I to 1980-IV (as cited by Gujarati).

Instead of using the gross values of these variables, he used their growth rates, GNP and M and used

four lags of each variable in the two regressions given previously. The results were as follows:
The null hypothesis in each case is that the variable under consideration does not “Granger cause” the other variable.

Direction of causality	<i>F</i> value	Decision
M →GNP	2.68	Reject
GNP→M	0.56	Do not reject

These results suggest that the direction of causality is from money growth to GNP growth since the estimated *F* is significant at the 5 percent significance level; the critical *F* value is 2.50 (for 4 and 71 df). On the other hand, there is no “reverse causation” from GNP growth to money growth, since the *F* value is statistically insignificant.

Chapter Summary

1. For psychological, technological, and institutional reasons, a regressand may respond to a regressor(s) with a time lag. Regression models that take into account time lags are known as **dynamic** or **lagged regression models**.

2. There are two types of lagged models: **distributed-lag** and **autoregressive**.

In the former, the current and lagged values of regressors are explanatory variables. In the latter, the lagged value(s) of the regressand appear as explanatory variables.

3. A purely distributed-lag model can be estimated by OLS, but in that case there is the problem of multicollinearity since successive lagged values of a regressor tend to be correlated.

4. As a result, some shortcut methods have been devised. These include the Koyck, the adaptive expectations, and partial adjustment mechanisms, the first being a purely algebraic approach and the other two being based on economic principles.

5. But a unique feature of the **Koyck**, **adaptive expectations**, and **partial adjustment models** is that they all are autoregressive in nature in that the lagged value(s) of the regressand appear as one of the explanatory variables.

6. Autoregressiveness poses estimation challenges; if the lagged regressand is correlated with the error term, OLS estimators of such models are not only biased but also are inconsistent. Bias and inconsistency are the case with the Koyck and the adaptive expectations models; the partial adjustment model is different in that it can be consistently estimated by OLS despite the presence of the lagged regressand.

7. To estimate the Koyck and adaptive expectations models consistently, the most popular method is the **method of instrumental variable**. The instrumental variable is a proxy variable for the lagged regressand but with the property that it is uncorrelated with the error term.

8. An alternative to the lagged regression models just discussed is the **Almon polynomial distributed-lag model**, which avoids the estimation problems associated with the autoregressive models. The major problem with the Almon approach, however, is that one must *pre-specify* both the lag length and the degree of the polynomial. There are both formal and informal methods of resolving the choice of the lag length and the degree of the polynomial.

Check Your Progress for Activities

Activity 3.1

False

Econometric models are dynamic if they portray the time path of the dependent variable in relation to its past values. Models using cross-sectional data are not dynamic, unless one uses time series or panel regression models with lagged values of the regressand.

Activity 3.2

True

The Koyck model assumes that all the distributed lag coefficients have the same sign.

Review Exercises

1. Differentiate the following concepts
- a. Moving Average Process vs Autoregressive process

b. Autoregressive vs Distributed lag models

c. Adaptive Expectation (AE) vs Partial Adjustment(PA) Models
2. Explain with a brief reason whether the following statements are true, false, or uncertain:
- a. If the Koyck and adaptive expectations models are estimated by OLS, the estimators will be biased but consistent.

b. In the partial adjustment model, OLS estimators are biased in finite samples.

c. In the presence of a stochastic regressor(s) and an autocorrelated error term, the method of instrumental variables will produce unbiased as well as consistent estimates.

d. In the presence of a lagged regressand as a regressor, the Durbin– Watson d statistic to detect autocorrelation is practically useless.

e. The Granger test is a test of precedence rather than a test of causality.

3. Consider the model: $Y_t = \alpha + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 Y_{t-1} + v_t$
- Suppose Y_{t-1} and v_t are correlated. To remove the correlation, suppose we use the following instrumental variable approach:

First regress Y_t on X_{1t} and X_{2t} and obtain the estimated $\hat{Y}_t$ from this regression. Then regress

$$Y_t = \alpha + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 \hat{Y}_{t-1} + v_t$$

where $\hat{Y}_{t-1}$ are estimated from the first-stage regression.

How does this procedure remove the correlation between Y_{t-1} and v_t in the original model?

4. Assume that prices are formed according to the following adaptive expectations hypothesis:
- $$P_t^* = \gamma P_{t-1} + (1 - \gamma) P_{t-1}^*$$

where P^* is the expected price and P the actual price.

Complete the following table, assuming $\gamma = 0.5$:

Period	P^*	P
$t - 3$	100	110
$t - 2$	-	125
$t - 1$	-	155
t	-	185
$t + 1$	-	—

5. What are the advantages of using Almon Polynomial Distributed-lag Model over that of Koyck Model and what are its draw backs
6. Consider the Koyck (or for that matter the adaptive expectation) model given in (3.28), namely,
$$Y_t = \alpha(1 - \lambda) + \beta_0 X_t + \lambda Y_{t-1} + (u_t - \lambda u_{t-1})$$

Suppose in the original model u_t follows the first-order autoregressive scheme $u_t - \rho u_{1-t} = \varepsilon_t$, where ρ is the coefficient of autocorrelation and where ε_t satisfies all the classical OLS assumptions.
- a. If $\rho = \lambda$, can the Koyck model be estimated by OLS?
- b. Will the estimates thus obtained be unbiased? Consistent? Why or why not?
- c. How reasonable is it to assume that $\rho = \lambda$?
7. Evaluate the median lag for $\lambda = 0.2, 0.4, 0.6, 0.8$.

CHAPTER FOUR

INTRODUCTION TO SIMULTANEOUS EQUATION MODELS

Chapter Contents

Dear distance students, this chapter consists of the following major sub-topics:

Introduction

Simultaneous Dependence of Economic

variables Definitions of Some Concepts

Consequences of Simultaneity bias

Problems of simultaneous equation models

The Identification Problem

Chapter Objectives

At completion of this chapter, students will be able to:

- ✓ recognize the nature of simultaneous equations
- ✓ define the basic concepts commonly used in simultaneous equation models
- ✓ prove the consequences of simultaneity bias
- ✓ know problems of simultaneous equation models
- ✓ verify the identifiability of simultaneous equations using order and rank conditions

4.1. Introduction

One important assumption of the basic linear regression model or the application of least squares to a single equation assumes, among others, that the explanatory variables are truly exogenous (i.e., the error term has to be uncorrelated with the explanatory variables); thus, there is one-way causation between the dependent variable Y and the explanatory X 's. If the explanatory variables are in fact correlated with the error term, one of the classical linear regression (OLS) assumption: $Cov(Xu)=0$ is violated and then become $Cov(Xu) \neq 0$, which is called **endogeneity problem**. This is a very serious problem in model estimation even severer than multicollinearity, heteroscedasticity and autocorrelation problems. If we have a two-way causation in a function this implies that the function cannot be treated in isolation as a single equation model, but belongs to a wider system of equations which describes the relationships among all the relevant variables. If $Y = f(X)$, but also $X = f(Y)$ we are not allowed to use a single-equation model for the description of the relationship between Y and X . We must use a multi-equation model, which would include separate equations in which Y and X would appear as **endogenous variables**, although they might appear as explanatory in other equations of the model. A system describing the joint dependence of variables is called a **system of simultaneous equation or simultaneous equations model**. The number of equations in such models is equal to the number of jointly dependent or endogenous variables involved in the phenomenon under analysis. Unlike the single equation models, in simultaneous equation models it is not usually possible (possible only under specific assumptions) to estimate a single equation of the model without taking into account the information provided by other equation of the system.

Thus, the application of Ordinary Least Squares (OLS) estimation for the model with the endogeneity problem, caused by **simultaneity** yields biased and inconsistent estimates of the parameters.

The bias arising from application of such procedure of estimation which treats each equation of the simultaneous equations model as though it were a single model is known as *simultaneity bias or simultaneous equation bias*.

In this chapter we will relax this assumption by including additional equations to the model that explains where the correlation is coming from, and discuss the conditions that need to be fulfilled to receive consistent estimate.

4.2. Simultaneous Dependence of Economic variables

Given the nature of economic phenomena it is evident that in reality there might be bidirectional causation among different economic variables since most economic variables are interrelated. Economists formulate models for consumption, production, investment, money demand and money supply, labor demand and labor supply to attempt to explain the workings of the economy. These behavioral equations are estimated equations by equation or jointly as a system of equations.

Some examples will illustrate the meaning of simultaneous relationships and the violation of assumption of ordinary least squares: $Cov(Xu) \neq 0$, which creates what is known as *simultaneous equations bias*.

Example 1: Suppose we want to estimate the demand for food. We know from economic theory that the demand for any particular commodity depends on its price, P , on other prices, P_o , and on income, Y , *ceteris paribus*. Thus we may write the demand function for food as

$$Q = \beta_0 + \beta_1 P + \beta_2 P_o + \beta_3 Y + u \quad (4.1) \text{ where } Q =$$

quantity demanded, P = price of food, P_o = price of other commodities, Y = income and u = random variable.

If we apply least squares to this equation we will obtain biased estimates of β_0 and β_1 , because P and u are not independent. The demand for any commodity is a function of its price, but at the same time the price in the market is influenced by the quantity demanded of that commodity. Consequently the above single equation cannot be treated as a complete (single-equation) model. There should be at least one more equation in the model giving the relationship between P and Q , for example:

$$P = \alpha_0 + \alpha_1 Q + \alpha_2 W + v \quad (4.2)$$

Where, W = index of weather conditions.

Substituting Q in this equation with its equal, we obtain

$$P = \alpha_0 + \alpha_1 (\beta_0 + \beta_1 P + \beta_2 P_o + \beta_3 Y + u) + \alpha_2 W + v \quad (4.3)$$

Obviously P is dependent on u and hence we have violation of Assumption 6 of the method of least squares. P is not an exogenous variable in the demand function.

Example 2: Suppose we want to estimate the supply of money. This of course is regulated by the government in an attempt to avoid inflation. Thus we may say that the main determinant of the decision of the government about the supply of money is the real level of income. Hence we may write the supply function of money as

$$M = b_0 + b_1 Y + u \quad (4.4)$$

Where, M = money supply and Y = level of real income.

However, the level of real income is in turn influenced by the supply of money as well as by other real forces, like the investment decisions of businessmen, the welfare policies of the government, and so on. Consequently the supply of money cannot be treated as a single-equation model. Y is not truly exogenous. There is a joint dependence between M and Y and hence we must construct a model with simultaneous equations, one of which would be

$$Y = a_0 + a_1 M + a_2 I_t + \dots + v \quad (4.5)$$

Substituting M , we obtain

$$Y = a_0 + a_1 (b_0 + b_1 Y + u) + a_2 I_t + \dots + v \quad (3.6)$$

Obviously $Y = f(u)$ and hence in the function of the supply of money the explanatory variable Y is not independent of the random variable u :

As stated earlier, the bias arising from the application of classical least squares to an equation belonging to a system of simultaneous relations is called *simultaneous equations bias*. It originates from the violation of Assumption 6 of OLS, that is it arises from the dependence of the explanatory variables and u , [$Cov(UX) \neq 0$]. This creates several problems. Firstly, there arises the problem of identification of the parameters of individual relationships. Secondly, there arise problems of estimation. The application of OLS yields biased and inconsistent estimates. One should therefore choose other estimation methods stated under introductory part above.

4.3. Definitions of Some Concepts Definition

1: Endogenous and exogenous variables

In simultaneous equation models variables are classified as endogenous and exogenous. The traditional definition of these terms is that endogenous variables are variables that are determined by the economic model (within the system) and exogenous variables are those determined from outside. However, from Econometrics view point in the case of Simultaneous Equation Models (SEMs), endogenous variable is a variable that is correlated with error term (i.e., the covariance between explanatory variable and error term is different from zero, $Cov(XU) \neq 0$) while Exogenous variable is a variable that is independent from or uncorrelated with the error term ($Cov(XU)=0$). .

Exogenous variables are also called *predetermined*.

Consider the following simple Keynesian macroeconomic model of income determination:

$$Y_t = C_t + I_t \quad (4.7)$$

$$C_t = \alpha_0 + \alpha_1 Y_t + U_t \quad (4.8)$$

with Y_t , being the national income, I_t investments, C_t , the consumption expenditure and U_t a stochastic term. Equation (4.7) is an identity and **an equilibrium condition**. Hence this model is formulated under the condition of being in equilibrium, and the equation show how national income is related to consumption and investment in equilibrium. The second, equation given by (4.8), is a **behavioral equation** since it defines the behavior of the consumption expenditure in this economy. Equations with stochastic error terms are to be considered behavioral. Since Y_t and C_t are left hand variables in this system of equations, their values are determined by the model. We therefore say that Y_t and C_t are **endogenous variables**. We have an additional variable included in the model which is the investment. Since it is a right hand variable in the consumption function we say that it is an **exogenous variable**, which is to say that the value of investments are determined outside the model, it is pre determined. Since investment is determined outside the model it is also uncorrelated with the stochastic term U_t .

Predetermined groups can be divided into two categories which are considered in general as exogenous variables. These are: *current and lagged exogenous* and *lagged endogenous*. For instance; X_t and X_{t-1} depict the current and lagged exogenous variables and Y_{t-1} depicts lagged endogenous variable. This is on the assumption that X 's symbolize the exogenous variables and Y 's symbolize the endogenous variables. Thus, X_t , X_{t-1} and Y_{t-1} are regarded as predetermined (exogenous) variables.

Since the exogenous variables are predetermined (determined outside the model), they are supposed to be independent of the error terms/stochastic term, U_t in the model with the (uncorrelated).

Consider the demand and supply functions.

$$Q^d = \alpha_0 - \alpha_1 P + \alpha_2 Y + U_1 \quad (4.9)$$

$$Q^s = \beta_0 + \beta_1 P + \beta_2 R + U_2 \quad (4.10)$$

Where, Q =quantity, Y =income, P =price, R =Rainfalls, U_1 & U_2 are error terms and we call them **unobserved shifters** of demand and supply curves, respectively.

Here P and Q are endogenous variables and Y and R are exogenous variables and are called **observed shifters** of demand and supply curves, respectively.

Under the assumption that markets clear, what we actually observe is equilibrium values of price and quantity demanded or supplied. The two equations are linked in econometric analysis only because observed price and quantity in the “i” units are simultaneously determined by the interaction of supply and demand by the condition: $Q^d = Q^s$.

Thus, the two equations constitute the *simultaneous equation models (SEMs)*.

Definition 2: Structural models

A structural model describes the complete structure of the relationships among the economic variables. Structural equations of the model may be expressed in terms of *endogenous variables, exogenous variables and disturbances (random variables)*.

The model is a *structural equation* in that it is derivable from economic theory and has a causal interpretation. The parameters of structural model express the *direct effect* of each explanatory variable on the dependent variable. Variables not appearing in any function explicitly may have an *indirect effect* and is taken into account by the simultaneous solution of the system.

For instance, in the simple Keynesian model of (4.7 and 4.8) a change in consumption affects the investment indirectly and is not considered in the consumption function. The effect of consumption on investment cannot be measured directly by any structural parameter, but is measured indirectly by considering the system as a whole.

As an illustration for structural models, we will use the following simple model for a closed economy.

$$C_t = a_0 + a_1 Y_t + U_1 \quad (4.11)$$

$$I_t = b_0 + b_1 Y_t + b_2 Y_{t-1} + U_2 \quad (4.12)$$

$$Y_t = C_t + I_t + G_t \quad (4.13)$$

The first equation is a consumption function, the second is an investment function, and the third is a definitional equation. The system is complete in that it contains three equations in three endogenous variables, C_t , I_t , Y_t . The model contains two predetermined variables; government expenditure, G , and lagged income, Y_{t-1} . *The above equations are called structural equations.*

The structural parameters (α 's and β 's) are, in general, propensities, elasticities or other parameters of economic theory. A structural parameter expresses the *direct effect* of each explanatory variable on the dependent variable. Indirect effects can be computed only by the solution 'of the structural system, but not by the individual structural parameters.

For example a change in consumption will affect investment indirectly, through the increase that the consumption, C , will produce on income, Y , which is a determinant of investment. The effect of C on I cannot be measured directly by any of the structural parameters, but it will be taken into account by the simultaneous solution of the system.

Traditionally the structural parameters are represented by β 's when they refer to endogenous variables and by γ 's when they are attached to a pre-determined variable. Similarly endogenous variables are denoted by y 's while exogenous variables are represented by x 's.

Using the conventional notation (and ignoring the constant intercepts) the structural system above becomes:

$$y_l = \quad + \quad (4.14)$$

$$y_2 = \begin{matrix} & + & & + & \\ & & & & \end{matrix} \quad (4.15)$$

$$y_3 = y_1 + y_2 + x_2 \quad (4.16)$$

Where $y_1 = C$ $y_2 = I$ $y_3 = Y$
 $x_1 = Y_{t-1}$ $x_2 = G$

Transferring all the observable variables to the left-hand side we may obtain the complete table of structural parameters as follows:

$$\begin{array}{ccccccc} & +0 & - & & +0 & & +0 \\ & & & & & & - \\ 0 & + & - & & - & & +0 = \\ & & & & & & \\ - & - & + & +0 & - & & =0 \end{array}$$

Note: We introduced zeros for missed variables from the order of system of linear equations

Similarly transferring equations 3.11-3.13, we will have:

$$\begin{aligned} C_t + 0I_t - a_l Y_t + 0Y_{t-1} + 0G_t &= U_1 \\ 0C_t + I_t - b_1 Y_t - b_2 Y_{t-1} + 0G_t &= U_2 \\ -C_t - I_t + Y_t + 0Y_{t-1} - G_t &= 0 \end{aligned}$$

Table of structural coefficients				
1	0	-a ₁	0	0
0	1	-b ₁	-b ₂	0
-1	-1	1	0	-1

Table of structural coefficients in standard notation				
1	0	-	0	0
0	1	-	-	0
-1	-1	1	0	-1

Values of the structural parameters may be obtained by using sample observations on the variables of the model and applying an appropriate econometric method

Activity 4.1

Consider a simple Keynesian model with no government

$$C_t = \alpha + \beta Y_t + u_t, t = 1, 2, \dots, T$$

$$Y_t = C_t + I_t$$

Where C_t denotes consumption, Y_t denotes disposable income, and I_t denotes autonomous investment. α and β are constant parameters such that $\alpha > 0$ & $0 < \beta < 1$.

- Comment on system of equations
- Identify the endogenous and exogenous variables

Definition 3: Reduced form of the model:

The reduced form of a structural model is the model in which the endogenous variables are expressed as a function of the predetermined variables and the error term only.

The reduced form equations can be obtained in two ways:

- 1. To express the endogenous variables directly as a function of the predetermined variables as:

$$Y_i = \alpha_i + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + u_i \quad (i = 1, 2, \dots, n) \tag{4.17}$$

and proceed with the estimation of the parameters $\alpha_i, \beta_1, \beta_2, \dots, \beta_k$ by applying some appropriate technique to this expression.

- 2. To solve the structural system of endogenous variables in terms of the predetermined variables, the structural parameters, and the disturbance terms.

Let us see one by one:

The first is to express the endogenous variables directly as functions of the predetermined variables. This is called *Direct Method*.

Consider in our example of the simple three- equation closed economy model of equations (4.11- 4.13). As we have said this model has three equations in three endogenous variables (C_t, I_t ,and Y_t) and two predetermined variables (G_t , and Y_{t-1}). Thus, express the endogenous variables (C_t, I_t , and Y_t) directly as functions of the predetermined variables (G_t , and Y_{t-1}), ignoring the constant terms of each equations) and use α_i 's as the parameters of the reduced form model.

Thus, we can re-write the model as:

$$C_t = \alpha_1 + \beta_1 Y_{t-1} + \beta_2 G_t + u_{1t} \tag{4.18}$$

$$I_t = \alpha_2 + \beta_3 Y_{t-1} + \beta_4 G_t + u_{2t} \tag{4.19}$$

$$Y_t = \alpha_3 + \beta_5 Y_{t-1} + \beta_6 G_t + u_{3t} \tag{4.20}$$

Note: $\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$, and β_7 are reduced form parameters.

The second method for obtaining the reduced form of a model is to solve the structural system of endogenous variables in terms of the predetermined variables, the structural parameters and the disturbances.

To derive of the reduced form model for each equation, we can follow the following procedures:

- (a) Substitute C_t and I_t in the third structural equation

$$Y_t = (a_1 Y_t + u_1) + (b_1 Y_t + b_2 Y_{t-1} + u_2) + G_t$$

By rearranging we obtain

$$Y_t = \frac{a_1 b_2}{1 - a_1 - b_1} Y_{t-1} + \frac{a_1}{1 - a_1 - b_1} G_t + \frac{u_1 + a_1 u_2 - b_1 u_1}{1 - a_1 - b_1} \tag{4.21}$$

This is the reduced form of the third structural equation.

- (b) Substitute Y_t into the consumption function

$$C_t = \frac{a_1 b_2}{1 - a_1 - b_1} Y_{t-1} + \frac{a_1}{1 - a_1 - b_1} G_t + \frac{u_1 + a_1 u_2 - b_1 u_1}{1 - a_1 - b_1} \tag{4.22}$$

This is the reduced form of the consumption function.

(c) Substitute Y_t into the investment function

$$I_t = b_1 \left[\frac{b_2}{1 - a_1 - b_1} Y_{t-1} + \frac{1}{1 - a_1 - b_1} G_t + \frac{u_1 + u_2}{1 - a_1 - b_1} \right] + b_2 Y_{t-1} + u_2$$

Or

$$= \frac{b_1 b_2}{1 - a_1 - b_1} Y_{t-1} + \frac{b_1}{1 - a_1 - b_1} G_t + \frac{b_1 (u_1 + u_2)}{1 - a_1 - b_1} + b_2 Y_{t-1} + u_2 \tag{4.23}$$

This is the reduced form of the investment function.

Thus, the structural system of our example gives the following reduced form model:

$$\begin{aligned} &= \frac{b_1 b_2}{1 - a_1 - b_1} Y_{t-1} + \frac{b_1}{1 - a_1 - b_1} G_t + \frac{b_1 (u_1 + u_2)}{1 - a_1 - b_1} + b_2 Y_{t-1} + u_2 \\ &= \frac{0}{1 - a_1 - b_1} + \frac{b_1}{1 - a_1 - b_1} G_t + \frac{b_1 (u_1 + u_2)}{1 - a_1 - b_1} + b_2 Y_{t-1} + u_2 \\ &= \frac{0}{1 - a_1 - b_1} + \frac{b_1}{1 - a_1 - b_1} G_t + \frac{b_1 (u_1 + u_2)}{1 - a_1 - b_1} + b_2 Y_{t-1} + u_2 \end{aligned}$$

Clearly for the two reduced forms to be consistent- the following relationships between the Π 's and the structural parameters must hold

$$\begin{aligned} \Pi_1 &= \frac{b_1 b_2}{1 - a_1 - b_1} & \Pi_2 &= \frac{b_1}{1 - a_1 - b_1} \\ \Pi_3 &= \frac{0}{1 - a_1 - b_1} & \Pi_4 &= \frac{b_1 (u_1 + u_2)}{1 - a_1 - b_1} \\ \Pi_5 &= \frac{0}{1 - a_1 - b_1} & \Pi_6 &= \frac{b_1 (u_1 + u_2)}{1 - a_1 - b_1} \end{aligned}$$

It should be clear that there is a definite relationship between the reduced- form parameters and the structural parameters: the Π 's are functions of the structural parameters.

The constants and coefficients of exogenous variables are called reduced form equation parameters and are non-linear functions of the structural parameters if the constant term included in the model.

The reduced-form parameters measure the *total effect*, direct and indirect, of a change in the predetermined variable on the endogenous variables, after taking account of the interdependences among the jointly dependent endogenous variables, while a structural parameter indicates only the direct effect within a single sector of the economy.

For example Π_1 measures the effect of a unit increase in Y_{t-1} on the value of investment.

This effect consists of two parts: firstly, there is the direct effect on I through the coefficient b_2 as set out in the structural equation of investment; secondly there is the additional effect due to the fact that an increase in Y_{t-1} affects $I_t \rightarrow$ and I_t influences $Y_t \rightarrow$ which in turn affects I_t ; finally Y_t affects $C_t \rightarrow$ which in turn affects Y_t and hence I_t .

Thus the total effect (measured by Π_1) of Y_{t-1} on I_t may be split into the following components

$$\Pi_1 = \frac{b_1 b_2}{1 - a_1 - b_1} = \frac{b_2}{1 + \frac{b_1}{1 - a_1 - b_1}}$$

Or

$$\begin{aligned} \Pi_1 &= \frac{b_2}{1 + \frac{b_1}{1 - a_1 - b_1}} + \frac{b_2 b_1}{1 - a_1 - b_1} \\ &= \frac{b_2}{1 + \frac{b_1}{1 - a_1 - b_1}} + \frac{b_2 b_1}{1 - a_1 - b_1} \end{aligned} \tag{4.24}$$

The reduced form coefficients are used for forecasting and policy analysis, since it is the total effect of a change in the exogenous variables on the dependent variable(s) that is of interest to the policy maker.

In summary, the above two ways of defining the reduced form model suggests that estimates of the reduced form coefficients may be obtained in two ways.

Firstly. Direct estimation of the reduced-form coefficients.

We express all the endogenous variables as functions of all the pre-determined variables of the system and we apply ordinary least squares to these reduced-form functions. This method of obtaining the Π 's is called least-squares- no-restrictions (LSNR), because it does not take into account any information on the structural parameters, that is, it does not use any restrictions imposed by the form of the structural system.

For example the structural equations define that some coefficients are zero if the respective variables are not included in a function; this information is not taken into account by the method of LSNR. This method does not require complete knowledge of the structural system. What is required is knowledge of the predetermined variables appearing in the whole system.

Secondly. Indirect estimation of the reduced-form coefficients.

We saw that there is a definite relationship between the reduced-form coefficients and the structural parameters. It is thus possible first to obtain estimates of the structural parameters by any appropriate econometric technique and then substitute these estimates into the *system of parameters' relationships* to obtain (indirectly) values for the Π 's. This indirect method involves three steps:

- (1) Solve the system of endogenous variables so that each equation contains only predetermined explanatory variables. This, as we saw, may be done by continuous substitution of variables, until we arrive at the reduced-form of all the equations. In this way we obtain *the system of parameters' relations*, that is to say the system which defines the relations between the β 's, the γ 's and γ 's.
- (2) Obtain estimates of the structural parameters by any appropriate econometric method.
- (3) Substitute the estimates β 's and γ 's into the system of parameters' relations to find the estimates of the reduced-form coefficients.

- This method is more complicated, but it has several advantages over the direct estimation of β 's from LSNR. (a) The derivation of reduced-form β 's from the structural β 's and γ 's is more efficient because in this way we take into account all the information (that is all the *a priori* restrictions imposed by the structure on the parameters) incorporated into the structural model. (b) Structural changes occur continuously over time. If we know these changes in the β 's and γ 's we may easily re compute the β 's. While if the β 's are computed with the LSNR method it will not, in general, be possible to take this information into account, because each is a function of several structural parameters, and if the exact relationship between β 's, β 's and γ 's has not been established, we cannot incorporate into the former the changes that may have occurred to the latter. (c) Extraneous information on some structural parameters may become available from other studies; such information again will be useless if we have not estimated the β 's from previous estimates of β 's and γ 's.

Activity 4.2

- 1. Refer to activity 3.1, solve for Y_t and C_t in terms of the constant and I_t . What are the resulting two equations?
- 2. Find the reduced form of the following structural model:

$C = \gamma_C + \gamma_Y Y + \gamma_Z Z + u_C$
 $Y = \gamma_Y C + \gamma_Z Z + u_Y$, for $\gamma_Y > 0$ and $0 < \gamma_Z < 1$

Where: C =consumption expenditure
 Z =non-consumption expenditure
 Y =national income

C and Y are endogenous variables while Z is exogenous variable.

Definition 4: Recursive models

A model is called recursive if its structural equations can be ordered in such a way that the first equation includes only the predetermined variables in the right hand side; the second equation contains predetermined variables and the first endogenous variable (of the first equation) in the right hand side and so on. For example:

$y_1 = f(x_1, x_2, \dots, x_k; u_1)$
 $y_2 = f(x_1, x_2, \dots, x_k; y_1; u_2)$
 $y_3 = f(x_1, x_2, \dots, x_k; y_1, y_2; u_3)$ and soon.

The special feature of recursive model is that its equations may be estimated, one at a time, by OLS without simultaneous equations bias.

To understand this, let us write the above recursive model in its complete form. Assume that there are G endogenous variables k exogenous variables in the model

$y_1 = \gamma_{11}x_1 + \gamma_{12}x_2 + \dots + \gamma_{1k}x_k + u_1$
 $y_2 = \gamma_{21}x_1 + \gamma_{22}x_2 + \dots + \gamma_{2k}x_k + \gamma_{21}y_1 + u_2$
 $y_3 = \gamma_{31}x_1 + \gamma_{32}x_2 + \dots + \gamma_{3k}x_k + \gamma_{31}y_1 + \gamma_{32}y_2 + u_3$
 $\vdots$
 $y_G = \gamma_{G1}x_1 + \gamma_{G2}x_2 + \dots + \gamma_{Gk}x_k + \gamma_{G1}y_1 + \gamma_{G2}y_2 + \dots + u_G$

Given values of the exogenous variables (X_i) we may apply OLS to each equation individually, because by assumption the distribution variables u_i and u_j are independent, and hence the y 's appearing in the right-hand side of each equation are independent of this equation's error term.

For example, in the second equation y_1 is independent of u_2 , given u_1 and u_2 are *ex-hypothesi* independent (uncorrelated).

Recursive systems are also called *triangular systems* because the coefficients of the endogenous variables (the 's) form a triangular array: the main diagonal of the array of β 's contains units, and no coefficients appear above the **main diagonal**.

For example assume that we have a model with four endogenous and five predetermined variables

$$y_1 = \gamma_{11}x_1 + \gamma_{12}x_2 + u_1$$
$$y_2 = \alpha_{21}y_1 + \gamma_{21}x_1 + \gamma_{22}x_2 + \gamma_{23}x_3 + u_2$$
$$y_3 = \alpha_{31}y_1 + \alpha_{32}y_2 + \gamma_{31}x_1 + \gamma_{34}x_4 + u_3$$
$$y_4 = \alpha_{41}y_1 + \alpha_{42}y_2 + \alpha_{43}y_3 + \gamma_{44}x_4 + \gamma_{45}x_5 + u_4$$

In the above illustration, as usual, the x 's and y 's are exogenous and endogenous variables respectively. The disturbance terms follow the following assumptions:

$$(u_1u_2) \quad (u_1u_3) \quad (u_1u_4) \quad E(u_2u_3) \quad E(u_2u_4) \quad E(u_3u_4) = 0$$

(4.25)

The above assumption is the most crucial assumption that defines the recursive model. If this does not hold, the above system is no longer recursive and OLS is also no longer valid. The first equation of the above system contains only the exogenous variables on the right hand side.

Since by assumption, the exogenous variable is independent of u_1 , the first equation satisfies the critical assumption of the OLS procedure. Hence OLS can be applied straight forwardly to this equation.

Consider the second equation. It contains the endogenous variable y_1 as one of the explanatory variables along with non-stochastic x 's. OLS can be applied to this equation only if it can be shown that y_1 and u_2 are independent of each other. This is true because u_1 , which affects y_1 is by assumption uncorrelated with u_2 , i.e. $(u_1u_2) = 0$. y_1 acts as a predetermined variable in so far as y_2 is concerned. Hence OLS can be applied to this equation. Similar argument can be extended to the 3rd and 4th equations because right hand side y 's are independent of respective equations u 's.

In this way, in the recursive system OLS can be applied to each equation separately.

To see whether this model is recursive it suffices to examine the form of the array of the α 's. If it is triangular the system is **recursive**. The system may be rewritten in the form where we formulated earlier as:

$$y_1 - \gamma_{11}x_1 - \gamma_{12}x_2 = u_1$$
$$y_2 - \alpha_{21}y_1 - \gamma_{21}x_1 - \gamma_{22}x_2 - \gamma_{23}x_3 = u_2$$
$$y_3 - \alpha_{31}y_1 - \alpha_{32}y_2 - \gamma_{31}x_1 - \gamma_{34}x_4 = u_3$$
$$y_4 - \alpha_{41}y_1 - \alpha_{42}y_2 - \alpha_{43}y_3 - \gamma_{44}x_4 - \gamma_{45}x_5 = u_4$$

The array of structural parameters is shown below.

α of endogenous variables				γ 's of exogenous variables				
y_1	y_2	y_3	y_4	x_1	x_2	x_3	x_4	x_5
1	0	0	0	-	-	0	0	0
-	1	0	0	-	-	-	0	0
-	-	1	0	-	0	0	-	0
-	-	-	1	0	0	0	-	-

Obviously the array of 's is triangular: the main diagonal contains units and all the coefficients above it are zero (the lower triangular matrix). Hence the system is **recursive**, and its equations can be estimated by OLS, which will yield **unbiased and consistent estimates**. Thus, in a very special type of simultaneous equations model called *Recursive, Triangular or Causal model*, the use of OLS procedure of estimation is appropriate.

Activity 4.3

Let us build a hypothetical recursive model for an agricultural commodity, say wheat. The production of wheat Y_1 , may be assumed to depend on exogenous factors: X_2 = climatic conditions; and X_3 =last season's price. The retail rice $=Y_2$ may be assumed to be the function of production level $=Y_1$ and exogenous factor X_4 = disposable income. Finally the price obtained by the producer $=Y_3$ can be expressed in terms of the retail price Y_2 and exogenous factor X_j = the cost of marketing the producer.

- a. Describe the relevant equations of the model
- b. Re-write the stated equation in a matrix form

4.4. Consequences of Simultaneity bias

Most of the time in the simultaneous equation models, the endogenous variables may depend on the error terms of the model (if there is interdependence between the explanatory variables and the error term); hence the OLS technique is not appropriate for estimation of an equation in a simulations equations model. *In such a case OLS suffers from **simultaneity bias**: results in biased and inconsistent estimators of α 's and β 's.*

To show simultaneity bias problem and its consequences, let's consider the following simple simultaneous equation model:

$$\begin{matrix} Y_1 & X & U \\ & & Y & Z & V \end{matrix}$$

(4.26)

Suppose that the following assumptions hold.

$$\begin{matrix} (U) & 0, & (V) & 0 \\ (U^2) & u^2, & (V^2) & u^2 \\ (U_i U_j) & 0, & (V_i V_j) & 0, \text{ also } (U_i V_i) & 0; \end{matrix}$$

where X and Y are endogenous variables and Z is an exogenous variable.

The *reduced form* of X of the above model is obtained by substituting Y in the equation of X.

$$X = \frac{0}{1} + \frac{0}{1} \frac{1}{1} + \frac{2}{1} \frac{1}{1} \frac{1}{1} Z + \frac{1}{1} \frac{U}{1} \frac{V}{1}$$

(4.27)

Applying OLS to the first equation of the above *structural model* will result in biased estimator because $cov(X_i U_i) (X_i U_j) \neq 0$.

Now, let's proof whether this expression holds or not.

$$\text{cov}(XU) = X' (X'X)^{-1} X' U = (U)' X (X'X)^{-1} X' U, \text{ since } E(U)=0 \quad (4.28)$$

$$\frac{\begin{smallmatrix} 0 & 0 & 1 \\ & 1 & \end{smallmatrix}}{\begin{smallmatrix} & & \\ & 1 & \end{smallmatrix}} \quad \frac{\begin{smallmatrix} 2 \\ & 1 & \end{smallmatrix}}{\begin{smallmatrix} & & \\ & 1 & \end{smallmatrix}}^Z \quad \frac{\begin{smallmatrix} U & V \\ & 1 & \end{smallmatrix}}{\begin{smallmatrix} & & \\ & 1 & \end{smallmatrix}} \quad \frac{\begin{smallmatrix} 0 & 0 & 1 \\ & 1 & \end{smallmatrix}}{\begin{smallmatrix} & & \\ & 1 & \end{smallmatrix}} \quad \frac{\begin{smallmatrix} 2 \\ & 1 & \end{smallmatrix}}{\begin{smallmatrix} & & \\ & 1 & \end{smallmatrix}}^{ZU}$$

Substituting the value of X in equation (3.27) into equation (3.28)

$$\frac{1}{1} \frac{U}{1} (0 \quad 012Z_1UV_{01} \quad 0 \quad 2^L)$$

That is, covariance between X and U is not zero. As a consequence, if OLS is applied to each equation of the model separately the coefficients will turn out to be biased. Now, let's examine how the non-zero co-variance of the error term and the explanatory variable will lead to biasness in OLS estimates of the parameters.

If we apply OLS to the first equation of the above structural equation (3.26) $Y_{01} X U$, we obtain:

$$\hat{1} \frac{xy}{x^2} - \frac{x(Y - \bar{Y})}{x^2} = \frac{xY}{x^2} - \frac{\bar{Y}x}{x^2} \quad ; \text{ (since } \frac{\bar{Y}x}{x^2} \text{ is zero)}$$

$$\frac{x(0 - 1X - U)}{x^2} = \frac{0 - x}{x^2} - \frac{xU}{x^2} = -\frac{xU}{x^2}$$

But, we know that $x > 0$ and $\frac{xX}{x^2} > 1$, hence

$$^1 \frac{xU}{x^2} \quad (4.29)$$

Taking the expected values on both sides;

$$\hat{()}\frac{xU}{x_2} \quad 1$$

Since, we have already proved that $(XU) = 0$; which is the same as $(XU) = 0$. Consequently,

0 ; (^) , that is ^_1 will be **biased by the amount equivalent to** $\frac{xu}{x^2}$ —

NB: With correlation between the error term and the explanatory variables in a single equation model (endogeneity problem due to simultaneity bias), using OLS would lead to:

Biased and inconsistent parameter estimates

Invalid tests of hypothesis

Biased and inconsistent forecasts

4.5. Problems of simultaneous equation models

Simultaneous equation models create three distinct problems. These are:

1. **Mathematical completeness of the model:** any model is said to be (mathematically) complete only when it possesses as many independent equations as endogenous variables. In other words if we happen to know values of disturbance terms, exogenous variables and structural parameters, then all the endogenous variables are uniquely determined.
2. **Identification of each equation of the model:** Many times it so happens that a given set of values of disturbance terms and exogenous variables yield the same values of different endogenous variables included in the model. It is because the equations are observationally indistinguishable, what is needed is that the parameters of each equation in the system should be uniquely determined. Hence, certain tests are required to examine the identification of each equation before its estimation.
3. **Statistical estimation of each equation of the model:** Since application of OLS yield biased and inconsistent estimates, different statistical techniques are to be developed to estimate the structural parameters. Some of the most common simultaneous methods of estimation/estimation methods to avoid simultaneity bias are:
 - i) The indirect least square method (ILS)
 - ii) The two-stage least square method (2SLS)
 - iii) The three-stage least square method (3SLS)
 - iv) Limited information maximum likelihood method (LIML)
 - v) The instrumental variable method (IV)
 - vi) The mixed estimation method; and
 - vii) The full information maximum likelihood method (FIML)

Of these three problems, we are going to discuss the second problem (the **identification problem**) in the following section.

However, these estimation methods are beyond the scope of this module due to the fact that this chapter will only scratch the surface of the issues involved in estimating simultaneously equations and should therefore be seen as an introduction to the subject.

4.6. The Identification Problem

In general, we can think of any structural equation, as having one left hand side endogenous variable y_1 , g_1 right hand side endogenous variables, and k_1 right hand side exogenous variables. The right hand side endogenous variables are correlated with the error term rendering OLS on this equation biased and inconsistent. In order to be able to estimate the structural equation coefficients they need to be identified. So, what do we mean by that?

In simultaneous equation models, the problem of identification is a **problem of model formulation**; it does not concern with the estimation of the model. The estimation of the model depends up on the empirical data and the form of the model.

If the model is not in the proper statistical form, it may turn out that the parameters may not uniquely estimated even though adequate and relevant data are available.

In a language of econometrics, *a model is said to be identified only when it is in unique statistical form to enable us to obtain unique estimates of its parameters from the sample data.*

Normally, for each endogenous variable, there exists a corresponding structural equation explaining its behavior in the model. We say that a system of simultaneous equations is complete if there are as many endogenous variables as there are equations.

To give an intuitive feeling for meaning of identification, we will give an example before going into any formal and mechanical tests.

Consider the following two equation system:

$$Q = A_0 + A_1P + A_2X_1 + U_1 \quad (\text{Supply}) \quad (4.30)$$

$$Q = B_0 + B_1P + U_2 \quad (\text{Demand}) \quad (4.31)$$

This system contains two endogenous variables (P and Q) and one exogenous (X_1) variable. These two equations represent a supply and demand system for a given market, respectively. The question is if any of these two equations are identified? That is to ask if the parameters of the two equations can be estimated consistently. It turns out that the demand function is identified while the supply function is not. The supply function contains an exogenous variable X_1 , and the supply function takes a new position for each value of X_1 . In that process we identify the demand function. But in the demand function we have nothing unique that does not appear in the supply function so it is impossible to move the demand function while holding the supply function fixed. Hence it is the presence of an exogenous variable in one equation that allows us to estimate the parameters of the other equation. If X_1 had been included in both equations there would have been no unique variation in any of the equations and hence no equation had been identified. However, if another exogenous variable, X_2 , had been introduced and placed in the demand function, we would receive some exogenous variation that could help us to identify the supply function. In that case both equations would have been identified.

In econometric theory two possible situations of identifiability are traditionally distinguished:

1. Equation under identified.
2. Equation identified
 - a. Exactly identified.
 - c. Over identified.

An *equation* is under identified if its statistical form is not unique. A *system* is under identified when one or more of its equations are under identified.

If an *equation* has a unique statistical form we say that it is identified. It may be exactly identified or over identified. But in both cases it is identified. A *system* is identified if *all* its equations are identified.

It should be noted that identification problems arise only for those equations which contain coefficients which must be estimated statistically (from sample data). Identification difficulties do not arise for definitional equations, identities, or statements of equilibrium conditions, because such relationships do not require measurement.

Though identification doesn't concern with estimation of the model, it is closely related to it.

(1) If an equation (or a model) is under identified it is impossible to estimate all its parameters with any econometric technique.

(2) If an equation is identified, its coefficient can, in general, be statistically estimated.

In particular: (a) If the equation is exactly identified, the appropriate method to be used for its estimation is the method of indirect least squares. (b) If the equation is over identified, indirect least squares cannot be applied, because it will not yield unique estimates of the structural parameters. There are various other methods which can be used in this case, for example two-stage least squares (2SLS), or maximum likelihood methods.

4.6.1. Formal Rules (Conditions) for identification

Identification may be established either by the examination of the specification of the structural model, or by the examination of the reduced form of the model.

Traditionally identification has been approached via the reduced form. Actually the term 'identification' was originally used to denote the possibility (or impossibility) of deducing the values of the parameters of the structural relations from a knowledge of the reduced-form parameters. In this section we will examine both approaches. However, we think that the reduced form approach is conceptually confusing and computationally more difficult than the structural model approach, because it requires the derivation of the reduced form first and then examination of the values of the determinant formed from some of the reduced form coefficients. The structural form approach is simpler and more useful.

In applying the identification rules we should either ignore the constant term, or, if we want to retain it, we must include in the set of variables a dummy variable (say X_0) which would always take on the value 1. Either convention leads to the same results as far as identification is concerned.

The process of identifying equations can be formalized in a decision rule that specify the conditions that have to be fulfilled in order to identify one or several equations in a system. In the literature, two rules are described and one is slightly easier to use than the other.

4.6.1.1. The Order Condition of Identification

The first decision rule for identification is the so called order condition. This condition is based on a counting rule of the variables included and excluded from the particular equation. It is **a necessary** but not sufficient condition for the identification of an equation, which means that it is possible that the equation is undefined even though the order condition says it is identified. However, in a system with only two equations, the order condition will work well and can be trusted. The order condition may be stated as follows:

For an equation to be identified the total number of variables (endogenous and exogenous) excluded from it must be equal to or greater than the number of endogenous variables in the model less one. Given that in a complete model the number of endogenous variables is equal to the number of equations of the model, the order condition for identification is sometimes stated in the following equivalent form.

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. or

For an equation to be identified the total number of variables excluded from it but included in other equations must be at least as great as the number of equations of the system less one.

Let G = total number of equations (= total number of endogenous variables)
 K = number of total variables in the model (endogenous and predetermined)
 M = number of variables, endogenous and exogenous, included in a particular equation.

Then the order condition for identification may be symbolically expressed as

$$(K-M) \geq (G-1)$$

$K-M$ = the number of variables (endogenous and exogenous) in the model excluded from the equation under consideration.

The order condition states that:

- 1) If $K-M = G - 1 \Rightarrow$ the equation is exactly identified
- 2) If $K-M > G-1 \Rightarrow$ the equation is over identified
- 3) If $K-M < G - 1 \Rightarrow$ the equation is under identified

NB: When checking the order condition you have to do it for each equation in the system.

Example 1

Consider the following system:

$$\begin{aligned} Y_1 &= A_0 + A_1Y_2 + A_2X_1 + U_1 \\ Y_2 &= B_0 + B_1Y_1 + B_2X_2 + U_2 \end{aligned}$$

Use the order condition to check if the equations are identified.
In order to do that, we need to determine the value of K , M and G . This system contains two endogenous variables and the total number of variables, endogenous as well as exogenous, is 4. The number of endogenous and exogenous variables in each equation is 3.
For the first equation we have $G-1 = 2-1=1$ and $K-M = 4-3=1$ since X_2 is excluded from the first equation. Since $G-1=K-M$, we have that the first equation is **exactly identified**.

For the second equation we have $G-1=1$ and $K-M=1$ since X_1 is excluded from the second equation. Since $G-1=1$ and $K-M=1$ we have that also the second equation is exactly identified. When all the equations of the model are identified we say that the model is identified; thus, we are able to **estimate all the structural parameters**.

Example 2

Consider the following system:

$$\begin{aligned} Y_1 &= A_0 + A_1 Y_2 + A_2 X_1 + A_3 X_3 + U_1 \\ Y_2 &= B_0 + B_1 Y_1 + B_2 X_2 + U_2 \end{aligned}$$

In this example we have two endogenous variables and three exogenous variables with a total of five variables. $G-1$ will in this example equal 1 as before since we still have only two endogenous variables. Will the equations be identified in this case?

The first equation contains four variables which means that one variable has been excluded from the equation, that is, X_2 does not appear in equation 1 and $K-M=1$. Since $G-1 = K-M$ the equation is exactly identified.

The second equation includes three variables which mean that two variables have been excluded. That is, X_1 and X_3 are not included in equation 2. That is $K-M = 2$, which means that $G-1 < K-M$ which leads to the conclusion that equation 2 is **over identified**.

The order condition for identification is necessary for a relation to be identified, but it is not sufficient, that is, it may be fulfilled in any particular equation and yet the relation may not be identified.

Activity 4.4

Given a system contains 10 equations with 15 variables, ten endogenous and five exogenous, an equation containing 11 variables is not identified, while another equation containing 5 variables is identified. Use the order condition to check whether the equations are identified or not.

4.6.1.2. The Rank Condition of Identification

The rank condition is slightly more complicated when dealing with larger systems of equations, but when using only two equations it is as easy as the order condition. The rank condition is **a necessary and sufficient condition**, which means that if we can identify the equations using the rank condition we can be sure that the equation really is identified.

The rank condition states that: *in a system of 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*. This condition is called rank condition because it refers to the rank of the matrix of parameters of excluded variables. The rank of a matrix is the order of the largest non-zero determinant which can be formed from the matrix. In our case the relevant matrix is the sub matrix of coefficients of the absent variables.

Hence the rank condition may be also stated as follows: a sufficient condition for identification of a relationship is that the rank of the matrix of parameters of all the excluded variables (endogenous and predetermined) from that equation be equal to (G - 1).

The rank condition investigates whether two or more equations are linearly dependent on each other, which would be the case if the sum of two equations would equal a third equation in the model. If that is the case it is impossible to identify all structural parameters.

The basic steps in this decision rule can be best described using illustrative examples. The practical steps for tracing the identifiability of an equation of a structural model may be outlined as follows:

Firstly: Write the parameters of all the equations of the model in a separate table, noting that the parameter of a variable excluded from an equation is equal to zero.

For example let a structural model be

$$y_1 = 3y_2 - 2x_1 + x_2 + u_1$$
$$y_2 = y_3 + x_3 + u_2$$
$$y_3 = y_1 - y_2 - 2x_3 + u_3$$

where the y's are the endogenous variables and the x's are the predetermined variables.

This model may be rewritten in the form

$$-y_1 + 3y_2 + 0y_3 - 2x_1 + x_2 + 0x_3 + u_1 = 0$$
$$0y_1 - y_2 + y_3 + 0x_1 + 0x_2 + x_3 + u_2 = 0$$
$$y_1 - y_2 - y_3 + 0x_1 + 0x_2 - 2x_3 + u_3 = 0$$

Ignoring the random disturbances the table of the parameters of the model is as follows.

Equations	Variables					
	Y ₁	Y ₂	Y ₃	X ₁	X ₂	X ₃
1 st equation	-1	3	0	-2	1	0
2 nd equation	0	-1	1	0	0	1
3 rd equation	1	-1	-1	0	0	-2

Secondly, Strike out the row of coefficients of the equation which is being examined for identification.

For example if we want to examine the identifiability of the **second equation** of the model we strike out the second row of the table of coefficients.

Thirdly, Strike out the columns in which a non-zero coefficient of the equation being examined appears. By deleting the relevant row and columns we are left with the coefficients of variables *not included* in the particular equation, but contained in the other equations of the model.

For example if we are examining for identification the second equation of the system, we will strike out the second, third and the sixth columns of the above table, thus obtaining the following tables.

Table of structural parameters

	Y ₁	Y ₂	Y ₃	X ₁	X ₂	X ₃
1 st	-1	3	0	-2	1	0
2 nd	0	-1	1	0	0	1
3 rd	1	-1	-1	0	0	-2

Table of parameters of excluded variables

Y ₁	X ₁	X ₂
-1	-2	1
1	0	0

Fourthly, Form the determinant(s) of order (G - 1) and examine their value. If at least one of these determinants is non-zero, the equation is' identified. If all the determinants of order (G - 1) are zero, the equation is under identified.

In the above example of exploration of the identifiability of the second structural equation we have three determinants of order (G - 1) = 3 - 1 = 2. They are:

$$\Delta = \begin{vmatrix} 1 & 0 \\ -1 & -2 \end{vmatrix} \neq 0$$
$$\Delta = \begin{vmatrix} -2 & 1 \\ 0 & 0 \end{vmatrix} = 0$$
$$\Delta = \begin{vmatrix} -1 & 1 \\ 1 & 0 \end{vmatrix} \neq 0$$

(the symbol "Δ" stands for 'determinant').

We see that we can form two non-zero determinants of order G - 1 = 3 - 1 = 2; hence the second equation of our system is identified.

Fifthly, to see whether the equation is exactly identified or over identified we use the order condition $(K - M) \geq (G - 1)$. With this criterion, if the equality sign is satisfied, that is if $(K - M) = (G - 1)$, the equation is exactly identified. If the inequality sign holds, that is, if $(K - M) > (G - 1)$, the equation is over identified and if $(K - M) < (G - 1)$, the equation is underidentified.

In the case of the second equation we have: G=3 K=6 M=3 and the counting rule $(K - M) \geq (G - 1)$ gives $(6 - 3) > (3 - 1)$

Therefore the second equation of the model is over identified.

The identification of a function is achieved by assuming that some variables of the model have a zero coefficient in this equation, that is, we assume that some variables do not directly affect the dependent variable in this equation. This, however, is an *assumption* which can be tested with the sample data.

NB: In many applications, an equation that satisfies order condition is identified.

Dear Learners, the above basic steps are relatively time consuming and a bit tedious; hence, we can have another method that we can follow. Let us see it in the following section as usual by the help of illustrations:

Example 2

Consider the following system of equations:

$$Y_1 \quad A_0 \quad A_1 Y_3 \quad A_2 X_1 \quad A_3 X_3 \quad U_1$$
$$Y_2 \quad B_0 \quad B_1 Y_1 \quad B_2 X_2 \quad B_3 X_3 \quad U_2$$
$$Y_3 \quad C_0 \quad C_1 Y_2 \quad A_2 X_1 \quad A_3 X_2 \quad U_3$$

This system contains three endogenous variables (Y₁, Y₂, Y₃) and three exogenous variables (X₁, X₂, X₃), which means that we in total has six variables.

The first step in checking the rank condition is to put up a matrix that for each equation mark which of the six variables that are included (marked with 1) and which that are excluded (marked with 0) from the equation as usual. For our system we receive the following matrix:

Matrix for the rank condition

	Y ₁	Y ₂	Y ₃	X ₁	X ₂	X ₃
Equation1	1	0	1	1	0	1
Equation 2	1	1	0	0	1	1
Equation 3	0	1	1	1	1	0

In order to check the rank condition for the first equation we have to proceed as follows:

Delete the first row and collect the columns for those variables of the first equation that were marked with zero.

For equation 1: Y_2 and X_2 was marked with zero, and if we collect those two columns we have:

$$\begin{matrix} 1 & 1 \\ 1 & 1 \end{matrix}$$

If this matrix contains less than $G-1$ rows or columns where all elements are zero, equation 1 will not be identified. G refers to the number of equation just as in the order condition, which means that $G-1=2$.

Since we have two rows and two columns and none of them contains only zeros we conclude that equation 1 is identified.

For equation 2 we proceed in the same way. We delete the second row and collect those columns where the elements of the second row were marked with a zero. For equation 2 that was the case for Y_3 and X_1 , which is to say that these two variables was not included in equation 2. The resulting matrix for this case then becomes:

$$\begin{matrix} 1 & 1 \\ 1 & 1 \end{matrix}$$

It looks in the same way as for equation 1, which means that we have two rows and two columns that is not only zeros. The same procedure should be done for the third equation and if you do that you will see that it is **identified** as well.

When using larger systems it is quite possible that the order condition says that a particular equation is identified even though the rank condition says it is not. When that happens it might still be possible to generate estimates, but those estimates will not have any economic meaning since they will represent averages of those equations that are linear combinations of each other. Hence, you should not be content that you have received identified results just because the order condition says so and the econometric software generates results for you. When using systems of more than two equations you should also confirm the identification using the rank condition.

Activity 4.5

Assume that we have a model describing the market of an agricultural product. From the theory of partial equilibrium we know that the price in a market is determined by the forces of demand and supply. The main determinants of the demand are the price of the commodity, the prices of other commodities, incomes and tastes of consumers. Similarly, the most important determinants of the supply are the price of the commodity, other prices, technology, the prices of factors of production, and weather conditions. The equilibrium condition is that demand be equal to supply. The above theoretical information may be expressed in the form of the following mathematical model.

$$\begin{matrix} Q_t & a_0 & a_1 P_1 & a_2 P_2 & a_3 Y & a_4 t & u \\ Q_t & b_0 & b_1 P_1 & b_2 P_2 & b_3 C & b_4 t & w \end{matrix}$$

$$D \quad S$$

Where: D= quantity demanded

S= quantity supplied

P_1 Price of the given commodity

P_2 Price of other commodities

Y= income

C= costs (index of prices of factors of production)

t= time trend. In the demand function it stands for 'tastes'; in the supply function it stands for 'technology'.

The above model is mathematically complete in the sense that it contains three equations in three endogenous variables, D, S and P_1 . The remaining variables, Y, P_2 , C, t are exogenous.

Identify the supply function using:

- Order condition
- Rank criteria for identification

4.6.1.3. Establishing Identification from the reduced form

Like that of the identification conditions from structural equations, there are two conditions for identification based on the reduced form of the model, *an order condition and a rank condition*. The order condition is the same as in the structural model. The rank condition here refers to the value of the determinant formed from some of the reduced form parameters, π 's.

1. Order condition (necessary condition), as applied to the reduced form

An equation belonging to a system of simultaneous equations is identified if

$$\begin{array}{cc} (K - M) & (G - 1) \\ \text{Total number of} & \text{number of} \end{array}$$

excluded variables *equations less one*

where, K, M and G have the same meaning as before:

K= total number of variables, endogenous and exogenous, in the entire model

M= number of variables, endogenous and exogenous, in any particular equation

G= number of structural equations=number of all endogenous variables in the model

If $(K-M) = (G-1)$, the equation is exactly identified, provided that the rank condition set out below is also satisfied. If $(K-M) > (G-1)$, the equation is over identified, while if $(K-M) < (G-1)$, the equation is under identified, under the same condition.

2. Rank condition as applied to the reduced form

Let G^* stand for the number of endogenous variables contained in a particular equation. The rank condition as applied to the reduced form may be stated as follows:

An equation containing G^* endogenous variables is identified if and only if it is possible to construct at least one non-zero determinant of order G^*-1 from the reduced form coefficients of the exogenous (predetermined) variables excluded from that particular equation.

The practical steps involved in this method of identification may be outlined as:

Firstly, obtain the reduced form of structural model.

For example assume that the original model is:

$$y_1 = b_{12} y_2 + y_{11} x_1 + y_{12} x_2 + u_1$$
$$y_2 = b_{23} y_3 + y_{23} x_3 + u_2$$
$$y_3 = b_{31} y_1 + b_{32} y_2 + y_{33} x_3 + u_3$$

This model is complete in the sense that it contains three equations in three endogenous variables. The model contains altogether six variables, three endogenous (y_1, y_2, y_3) and three exogenous (x_1, x_2, x_3).

The reduced form of the model is obtained by solving the original equations for the exogenous variables. The reduced form in the above example is:

$$y_1 = \pi_{11} x_1 + \pi_{12} x_2 + \pi_{13} x_3 + v_1$$
$$y_2 = \pi_{21} x_1 + \pi_{22} x_2 + \pi_{23} x_3 + v_2$$
$$y_3 = \pi_{31} x_1 + \pi_{32} x_2 + \pi_{33} x_3 + v_3$$

where the π 's are functions of the structural parameters.

Secondly, form the complete table of the reduced form coefficients.

Equations	Exogenous Variables		
	x_1	x_2	x_3
1 st equation: y_1	π_{11}	π_{12}	π_{13}
2 nd equation: y_2	π_{21}	π_{22}	π_{23}
3 rd equation: y_3	π_{31}	π_{32}	π_{33}

Strike out the rows corresponding to endogenous variables excluded from the particular equation being examined for identifiability.

Also strike out all the columns referring to exogenous variables included in the structural form of the particular equation.

After these deletions we are left with the reduced form coefficients of exogenous variables excluded (absent) from the structural equation.

For example, assume that we are investigating the identification procedure are found by striking out the first row (since y_1 , does not appear in the second equation) and the third column (since x_3 , is included in this equation).

Complete table of reduced form coefficients

	x_1	x_2	x_3
11	~~π_{11}~~	~~π_{12}~~	~~π_{13}~~
21	π_{21}	π_{22}	~~π_{23}~~
31	π_{31}	π_{32}	~~π_{33}~~

Table of reduced form coefficients of excluded exogenous variables

	~~π_{21}~~	~~π_{22}~~
~~π_{31}~~		~~π_{32}~~

Thirdly, examine the order of the determinants of the π 's of excluded exogenous variables and evaluate them. If the order of the largest non-zero determinant is G^*-1 , the equation is identified. Otherwise the equation is not identified.

Example:

Assume that a structural model is of the form

$$y_1 = 3y_2 - 2x_1 + x_2 + u_1$$
$$y_2 = y_3 + x_3 + u_2$$
$$y_3 = y_1 - y_2 - 2x_3 + u_3$$

We will examine the identification of each structural equation using the reduced-form approach.

The reduced form of the structural model is

$$y_1 = 3y_2 - 2x_1 + x_2 + u_1$$
$$y_2 = 2y_3 - x_3 + u_2$$
$$y_3 = 2y_1 - y_2 - 2x_3 + u_3$$

First equation: $y_1 = 3y_2 - 2x_1 + x_2 + u_1$

1. Order condition: $(K-M) \geq (G-1)$

$$K=6 \quad M=4 \quad G=3$$

So that $(K-M)=(G-1)$

Or $(6-4)=(3-1)$

Thus the equation satisfies the necessary condition for identification.

2. Rank condition

Excluded endogenous variables: y_3

Included exogenous variables: $x_1; x_2$

Table of all reduced form coefficients

4	-2	3
2	-1	1
2	1	0

Table of π 's of excluded exogenous variables

3
1

From the table of π 's of excluded exogenous variables we can form two non-zero determinants of order 1×1 .

Now $G^* = 2$, since the first equation contains two endogenous variables. Thus the order of the determinant of π 's relating to excluded X 's is one:

$$(G^* - 1) = 1$$

The rank condition is satisfied and the equation is identified.

Furthermore, by virtue of the order condition we see that the first equation is exactly identified since the equality sign holds for this condition (that is $K - M = G - 1$).

Second equation: $y_2 = y_3 + x_3 + u_2$

1. Order condition

$$K=6 \quad M=3 \quad G=3$$

$$(K-M) > (G-1)$$

$(6 - 3) > (3 - 1)$. Thus the order condition is satisfied.

2. Rank condition

Excluded endogenous variables: y_1

Included exogenous variables: x_3

Table of all γ

4	-2	3
2	-1	1
2	-1	0

Table of γ of excluded exogenous variables

2	-1
2	-1

The order of the determinant of γ of the excluded exogenous variables is 2 x 2; however its value is zero:

$$\Delta = \begin{vmatrix} 2 & -1 \\ 2 & -1 \end{vmatrix} = 0$$

Hence the highest non-zero determinant is of order 1 x 1, which we may form from the γ is 1x1.
The second equation contains two endogenous variables, that is $G^* = 2$. Therefore the order of highest non-zero determinant is $G^* - 1 = 1$.

Consequently the second equation is identified. Furthermore, using the order condition $(6 - 3) > (3 - 1)$ we conclude that the second structural equation is **over identified**.
Third equation: $y_3 = y_1 - y_2 - 2x_3 + u_3$

1. Order condition

$$K=6 \qquad M=4 \qquad G = 3$$

so that $(K - M) = (G - 1)$

or $(6-4)=(3-1)$

The order condition is satisfied

2. Rank condition

Excluded endogenous variables: none

Included exogenous variables: x_3

Table of all γ

4	-2	3
2	-1	1
2	-1	0

Table of γ of excluded exogenous variables

4	-2
2	-1
2	-1

The highest non-zero determinant is of order 1 x 1, since

$$\Delta = \begin{vmatrix} 4 & -2 \\ 2 & -1 \end{vmatrix} = 0 \qquad \Delta = \begin{vmatrix} 2 & -1 \\ 2 & -1 \end{vmatrix} = 0$$

The equation contains three endogenous variables, that is $G^*=3$. Hence order of largest non-zero determinant) $< (G^* - 1$.

The rank condition is not satisfied and therefore the third equation is not identified, despite the fulfillment of the first condition for identification.

Chapter Summary:

1. In contrast to single-equation models, in simultaneous-equation models more than one dependent, or **endogenous**, variable is involved, necessitating as many equations as the number of endogenous variables.
2. A unique feature of simultaneous-equation models is that the endogenous variable (i.e., regressand) in one equation may appear as an explanatory variable (i.e., regressor) in another equation of the system. As a consequence, such an **endogenous explanatory variable** becomes stochastic and is usually correlated with the disturbance term of the equation in which it appears as an explanatory variable.
3. In the above situation the classical OLS method may not be applied because the estimators thus obtained are not consistent, that is, they do not converge to their true population values no matter how large the sample size.
4. The problem of identification precedes the problem of estimation. The identification problem asks whether one can obtain unique numerical estimates of the structural coefficients from the estimated reduced form coefficients. If this can be done, an equation in a system of simultaneous equations is identified. If this cannot be done, that equation is un- or under identified.
5. An identified equation can be **just/exact/fully identified** or **over identified**. In the former case, unique values of structural coefficients can be obtained; in the latter, there may be more than one value for one or more structural parameters.
6. The identification problem arises because the same set of data may be compatible with different sets of structural coefficients, that is, different models. Thus, in the regression of price on quantity only, it is difficult to tell whether one is estimating the supply function or the demand function, because price and quantity enter both equations.
7. To assess the identifiability of a structural equation, one may apply the technique of **reduced-form equations**, which expresses an endogenous variable solely as a function of predetermined variables. However, this time-consuming procedure can be avoided by resorting to either the **order condition** or the **rank condition** of identification.
8. Although the order condition is easy to apply, it provides only a necessary condition for identification. On the other hand, the rank condition is both a necessary and sufficient condition for identification. If the rank condition is satisfied, the order condition is satisfied, too, although the converse is not true.

In practice, though, the order condition is generally adequate to ensure identifiability.

Check Your Progress for Activities

Activity 4.1

- a. This is a system of two simultaneous equations, also known as structural equations with the second equation being an identity. The first equation can be estimated by OLS
- b. Since I_t is autonomous, it is an exogenous variable determined outside the system, whereas C_t and Y_t are endogenous variables determined by the system.

Activity 4.2

1. The resulting two equations are known as the **reduced form equations**

$$C_t = \alpha/(1 - \beta) + \beta I_t(1 - \beta) + u_t/(1 - \beta)$$
$$Y_t = \alpha/(1 - \beta) + I_t/(1 - \beta) + u_t/(1 - \beta)$$

2. Since C and Y are endogenous variables and only Z is the exogenous variables, we have to express C and Y in terms of Z. To do this substitute $Y=C+Z$ into consumption equation

$$C = (C + Z) + U$$
$$C - Z = U$$

$$C = \frac{U}{1 - \beta} + \frac{Z}{1 - \beta}$$

**

Substituting again the above equation into income identity, we get;

$$Y = \frac{1}{1 - \beta} + \frac{Z}{1 - \beta} + \frac{U}{1 - \beta}$$

Equation (**) and (***) are called the reduced form of the structural model of the above. We can write this more formally as:

Structural form equations	Reduced form equations
$C = \alpha + \beta Y + u$	$C = \frac{\alpha}{1 - \beta} + \frac{\beta Z}{1 - \beta} + \frac{U}{1 - \beta}$
$Y = C + Z$	$Y = \frac{1}{1 - \beta} + \frac{Z}{1 - \beta} + \frac{U}{1 - \beta}$

Parameters of the reduced form measure the *total effect (direct and indirect)* of a change in exogenous variables on the endogenous variable.

For instance, in the above reduced form equation (**), $\frac{1}{1 - \beta}$ measures the total effect of a unit change in the non-consumption expenditure on consumption. This total effect is, the direct effect ($\frac{1}{1 - \beta}$) times the indirect effect $\frac{\beta}{1 - \beta}$.

Activity 4.3

a. The relevant equations of the model may be described as under:

$$\begin{matrix} Y_1 & 1 & X_2 & X_3 & U_1 \\ 2 & 4 & 1 & 1 & 5 & 4 & 2 \\ Y_3 & 6 & 2Y_2 & 7X_5 & U_3 \end{matrix}$$

In the first equation, there are only exogenous variables and are assumed to be independent of U_1 .
In the second equation, the causal relation between Y_1 and Y_2 is in one direction.
Also Y_1 is independent of U_2 and can be treated just like exogenous variable. Similarly since Y_2 is independent of U_3 , OLS can be applied to the third equation.

b. we can rewrite the above equations as follows:

$$\begin{matrix} Y_1 & 1 & Y_2 & X_2 & X_3 & U_1 \\ 1 & 1 & 2 & 4 & 5 & 4 & 2 \\ 2Y_2 & Y_3 & 6 & 7X_5 & U_3 \end{matrix}$$

We can again rewrite this in matrix form as follows:

1	0	0	Y_1	1	2	3	0	0
1	1	0	Y_2	4	0	0	5	0
0	2	1	Y_3	6	0	0	0	7
Coefficient matrix of endogenous variables				coefficient matrix of exogenous variable				

$$\begin{matrix} & & & & 3 \\ & & & & v & 3 \\ & & & & & & \\ X & & & & & & 4 \\ & & & & & & \\ X & & & & & & \\ & & & & & & 5 \end{matrix}$$

$$\begin{matrix} \lambda_1 & & & & & & U_1 \\ X_2 & & & & & & U_2 \end{matrix}$$

The coefficient matrix of endogenous variables is thus a **triangular one**; hence recursive models are also called as triangular models.

Activity 4.4

a. For the first equation we have

$$G = 10, \quad K = 15, \quad M = 11$$

Order condition:

$$\begin{matrix} (K - M) & (G - 1) \\ (15 - 11) & (10 - 1) \end{matrix} \quad ; \text{ That is, the order condition is not satisfied.}$$

b. For the second equation we have

$$G = 10, \quad K = 15, \quad M = 5$$

Order condition:

$$\begin{matrix} (K - M) & (G - 1) \\ (15 - 5) & (10 - 1) \end{matrix} \quad ; \text{ That is, the order condition is satisfied.}$$

The order condition for identification is necessary for a relation to be identified, but it is not sufficient, that is, it may be fulfilled in any particular equation and yet the relation may not be identified.

Activity 4.5

1. Order condition: $(K - M) \geq (G - 1)$

In our example we have: $K=7$ $M=5$ $G=3$

Therefore, $(K-M) = (G-1)$ or $(7-5) = (3-1) = 2$

Consequently the second equation satisfies the first condition for identification.

2. Rank condition

The table of the coefficients of the structural model is as follows.

Equations	Variables						
	D	P_1	P_2	Y	t	S	C
1 st equation	-1	a_1	a_2	a_3	a_4	0	0
2 nd equation	0	b_1	b_2	0	b_3	-1	b_4
3 rd equation	1	0	0	0	0	1	0

Following the procedure explained earlier we strike out the second row and the second, third, fifth, sixth and seventh columns. Thus we are left with the table of the coefficients of excluded variables:

Complete table of
Structural parameters

-1	a_1	a_2	a_3	a_4	0	0
0	b_1	b_2	0	b_3	-1	b_4
1	0	0	0	0	1	0

Table of parameters
of variables
excluded from the
second equation

-1	a_3
-1	0

From this table we can form only one non-zero determinant of order

$(G-1) = (3-1) = 2$

$$\begin{vmatrix} 1 & a_3 \\ 0 & -1 \end{vmatrix} = (1)(-1) - (a_3)(0) = -1 \neq 0$$

The value of the determinant is non-zero, provided that $a_3 \neq 0$.

We see that both the order and rank conditions are satisfied. Hence the second equation of the model is identified. Furthermore, we see that in the order condition the equality holds: $(7-5) = (3-1) = 2$.

Consequently the second structural equation is **exactly identified**.

Review Exercises

1. Identify the demand function of **activity 4.5** using:
a. Order condition
b. Rank criteria for identification
2. Assume the following simple version of the Keynesian model of income determination.

Consumption function: $C_t = a_0 + a_1 Y_t + a_2 T_t + u$

Investment function: $I_t = b_0 + b_1 Y_{t-1} + u$

Taxation function: $T_t = c_0 + c_1 Y_t + w$

Definition: $Y_t = C_t + I_t + G_t$

This model is mathematically complete in the sense that it contains as many equations as endogenous variables. There are four endogenous variables, C, I, T, Y, and two predetermined variables, lagged income (Y_{t-1}) and government expenditure (G).

- a. Identify consumption function using order and rank conditions
b. Identify investment function using order and rank conditions
3. Detect the identifiability of the tax equation of question 2 using both conditions.
4. The celebrated IS, or goods market equilibrium, model of macroeconomics in its non-stochastic form can be expressed as

Consumption function: $C_t = \beta_0 + \beta_1 Y_{dt}$ $0 < \beta_1 < 1$ (1)

Tax function: $T_t = \alpha_0 + \alpha_1 Y_t$ $0 < \alpha_1 < 1$ (2)

Investment function: $I_t = \gamma_0 + \gamma_1 r_t$ (3)

Definition: $Y_{dt} = Y_t - T_t$ (4)

Government expenditure: $G_t = \bar{G}$ (5)

National income identity: $Y_t = C_t + I_t + G_t$ (6)

where Y = national income
 C = consumption spending
 I = planned or desired net investment
 $\bar{G}$ = given level of government expenditure
 T = taxes
 Y_d = disposable income
 r = interest rate

- a. Obtain the reduced form of IS equation
b. What would happen if we were to estimate, say, the consumption function (1) in isolation?
Could we obtain unbiased and/or consistent estimates of β_0 and β_1 ?

5. The other half of the famous IS-LM paradigm is the LM, or money market equilibrium, relation, which gives the combinations of the interest rate and level of income such that the money market is cleared, that is, the demand for money is equal to its supply.

The IS and LM curves show, respectively, that a whole array of interest rates is consistent with goods market equilibrium and a whole array of interest rates is compatible with equilibrium in the money market. Of course, only one interest rate and one level of income will be consistent simultaneously with the two equilibria.

Algebraically, the model, in the non-stochastic form, may be expressed as:

$$\text{Money demand function: } M_t^d = a + bY_t - cr_t \quad (1)$$

$$\text{Money supply function: } M_t^s = \quad (2)$$

$$\text{Equilibrium condition: } M_t^d = M_t^s \quad (3)$$

where Y = income, r = interest rate, and = assumed level of money supply, say, determined by the Fed.

Derive the reduced form of the system of equations

6. G. Menges developed the following econometric model for the West German economy*:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 I_t + u_{1t}$$

$$I_t = \beta_3 + \beta_4 Y_t + \beta_5 Q_t + u_{2t}$$

$$C_t = \beta_6 + \beta_7 Y_t + \beta_8 C_{t-1} + \beta_9 P_t + u_{3t}$$

$$Q_t = \beta_{10} + \beta_{11} Q_{t-1} + \beta_{12} R_t + u_{4t}$$

Where, Y = national income

I = net capital formation

C = personal consumption

Q = profits

P = cost of living index

R = industrial productivity

t = time

u = stochastic disturbances

- Which of the variables would you regard as endogenous and which as exogenous?
- What is the reason behind including the variable P in the consumption function?
- Determine whether the structural equations of the model are identified.

Answers for Even numbers of the Chapter Review questions

Answers for Review questions of chapter One

2. Ranked the data by X values:

Y	X	Y	X
55	80	55	80
65	100	70	85
70	85	75	90
80	110	65	100
79	120	74	105
84	115	80	110
98	130	84	115
95	140	79	120
90	125	90	125
75	90	98	130
74	105	95	140
110	160	108	145
113	150	113	150
125	165	110	160
108	145	125	165
115	180	115	180
140	225	130	185
120	200	135	190
145	240	120	200
130	185	140	205
152	220	144	210
144	210	152	220
175	245	140	225
180	260	137	230
135	190	145	240
140	205	175	245
178	265	189	250
191	270	180	260
137	230	178	265
189	250	191	270

Answers for Even numbers of the Chapter Review questions

Dropping the middle four observations, the OLS regression based on the first 13 and the last 13 observations and their associated sums of squares are shown next (standard errors in parentheses).

Regression based on the first 13 observations

$$Y_i = 3.4094 + 0.6968X_i + e_i$$

$$(8.7049) \quad (0.0744)$$

$$R^2 = 0.8887$$

$$RSS_1 = 377.17$$

$$df = 11$$

Regression based on the last 13 observations

$$Y_i = 28.0272 + 0.7941X_i + e_i$$

$$(30.6421) \quad (0.1319)$$

$$R^2 = 0.7681$$

$$RSS_2 = 1536.8$$

$$df = 11$$

From these results we obtain:

$$F = \frac{RSS_2 / df}{RSS_1 / df} = \frac{1536.8 / 11}{377.17 / 11}$$

$$F = 4.07$$

The critical F-value for 11 numerator and 11 denominator for df at the 5% level is 2.82. Since the estimated F value exceeds the critical value, we may conclude that there is heteroscedasticity in the error variance. However, if the level of significance is fixed at 1%, we may not reject the assumption of homoscedasticity, because the critical F-value for 11 numerators and 11 denominators for df is 4.46; thus, we accept the assumption of homoscedasticity (there is no difference in the variances of the two groups).

Answers for Even numbers of the Chapter Review questions

4.

	3	0	0		1	$\sqrt{2}$	0		3	$\sqrt{2}$	$\sqrt{4}$
σ	5	0			$\frac{1}{2}$	1	$\frac{1}{2}$		$\frac{1}{2}$	1	$\frac{1}{2}$
	0	0	3		1	$\sqrt{2}$	1		$\frac{1}{4}$	$\sqrt{2}$	2
Heteroscedasticity with no autocorrelation				Homoscedasticity with autocorrelation				Heteroscedasticity with autocorrelation			

6.

Regress Y on X: i.e. $Y_t = X_t \beta + U_t$:

From the above table we can compute the following values.

$$\begin{aligned} \sum xy &= 255, & \bar{Y} &= 7, & (\sum e_i - \sum e_{t1})^2 &= 60.21 \\ \sum x^2 &= 280, & \bar{X} &= 8, & \sum e_t^2 &= 41.767 \\ \sum y^2 &= 274 \end{aligned}$$

$$\begin{aligned} \hat{\rho} &= \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = \frac{255}{\sqrt{280 \cdot 274}} = 0.91 \\ \hat{Y} &= \bar{Y} + \hat{\rho}(\bar{X} - \bar{Y}) = 7 + 0.91(8 - 7) = 7.91 \end{aligned}$$

$$Y = 0.29 + 0.91X + U_i$$

$$\begin{aligned} \hat{Y} &= 0.28 + 0.91X, & R^2 &= 0.85 \\ d &= \frac{(\sum e_t - \sum e_{t1})^2}{\sum e_t^2} = \frac{60.213}{41.767} = 1.442 \end{aligned}$$

Values of d_L and d_U on 5% level of significance, with $n=15$ and one explanatory variable are:

$$d_L = 1.08 \text{ and } d_U = 1.36.$$

Answers for Even numbers of the Chapter Review questions

$$(4 \ d_u) \ 2.64$$

$$d_U \ d \ 4 \ d_U \ (1.364 \ 2.64)$$

$$d^* \ 1.442$$

Since d^* lies between d_U and d_L , we accept H_0 . This implies the data is free from autocorrelation problem.

8. Since the lag value the dependent variable is introduced as explanatory variable, the h-statistic is an appropriate method.

$$h = \frac{1.8}{2} \sqrt{\frac{50}{1 - 50 \times 0.14^2}} = 5$$

Since h-calculated is greater than the tabulated value at 1% & 5% (2.58 and 1.96, respectively), we reject the null hypothesis of no autocorrelation (zero first order autocorrelation) and conclude there is serial correlation among error terms.

Answers for Review questions of chapter Two

2.

(a): Male Professor: $E(Y_i) = (\alpha_1 + \alpha_2) + \beta X_i$

Female Professor: $E(Y_i) = \alpha_1 + \beta X_i$

Holding X constant, the male average salary is different by α_2

(b) Male Professor: $E(Y_i) = (\alpha_1 + 2\alpha_2) + \beta X_i$

Female Professor: $E(Y_i) = (\alpha_1 + \alpha_2) + \beta X_i$

Holding X constant, the male average salary is also different by α_2

(c) Male Professor $E(Y_i) = (\alpha_1 - \alpha_2) + \beta X_i$

Female Professor $E(Y_i) = (\alpha_1 + \alpha_2) + \beta X_i$

Holding X constant, the difference between female and male average salary is $2\alpha_2$.

Since the scale of the dummy variable is arbitrary, there is no particular advantage of one method over the other. For a given data, the answer is invariant to the choice of the dummy scheme.

Answers for Even numbers of the Chapter Review questions

4.

a. If the intercept is present in the model, introduce 11 dummies. If the intercept is suppressed (i.e., regression thorough the origin), introduce 12 dummies.

b. If the intercept is included in the model, introduce 5 dummies, but if the intercept is excluded, introduce 6 dummies.

Answers for Review questions of chapter Three

2.

a. False

The estimators are biased as well as inconsistent

b. True

c. False

The methods produces consistent estimates, although in small samples ($n < 30$) the estimates thus, obtained are biased.

d. True

In such situations, use the Durbin h statistic. However, the Durbin d statistic can be used in the computation of the h-statistic.

e. True

The Granger test is a measure of precedence and information content but does not, by itself, indicate causality in the common use of the term

4.

The P* values are 105, 115, 135, and 160, respectively.

Answers for Even numbers of the Chapter Review questions

6.

- a. Yes, in this case the Koyck model may be estimated with OLS
- b. There will be a finite sample bias due to the lagged regressand, but the estimates are consistent.
- c. Since both ρ and λ are assumed to lie between 0 and 1, the assumption that they both are equal is conceivable.

Answers for Review questions of chapter Four

2A. *The first equation (consumption function) is not identified*

1. Order condition: $(K - M) \geq (G - 1)$

There are six variables in the model ($K=6$) and four equations ($G=4$). The consumption function contains three variables ($M=3$). $(K-M)=3$ and $(G-1)=3$

Thus $(K-M) = (G-1)$, which shows that the order condition for identification is satisfied.

2. Rank condition

The table of structural coefficients is as follows

Equations	Variables					
	C	Y	T	I	Y_{t-1}	G
1 st equation	-1	a_1	$-a_2$	0	0	0
2 nd equation	0	0	0	-1	b_1	0
3 rd equation	0	C_1	-1	0	0	0
4 th equation	1	-1	0	1	0	1

We strike out the first row and the three first columns of the table and thus obtain the table of coefficients of excluded variables.

Answers for Even numbers of the Chapter Review questions

Complete table of
structural parameters

-1	a_1	a_2	0	0	0
0	0	0	-1	b_1	0
0	c_1	-1	0	0	0
1	-1	0	1	0	1

Table of coefficients
of excluded variables

-1	b_1	0
0	0	0
1	0	0

We evaluate the determinant of this table. Clearly the value of this determinant is zero, since the second row contains only zeros. Consequently we cannot form any nonzero determinant of order $3(=G-1)$. The rank condition is violated.

Hence we conclude that the consumption function is not identified, despite the satisfaction of the order criterion.

B. The investment function is overidentified

1. Order condition

The investment function includes two variables.

Hence $K-M = 6-2$

Clearly $(K-M) > (G-1)$, given that $G-1=3$. The order condition is fulfilled.

2. Rank condition

Deleting the second row and the fourth and fifth columns of the structural coefficients table we obtain .

Complete table of structural
Parameters

-1	a_1	a_2	0	0	0
0	0	0	-1	b_1	0
0	c_1	-1	0	0	0
1	-1	0	1	0	1

Table of coefficients of
excluded variables

-1	a_1	-1	0
0	c_1	-1	0
1	-1	0	1

Answers for Even numbers of the Chapter Review questions

The value of the first 3x3 determinant of the parameters of excluded variables is

$$\begin{vmatrix} 1 & c_1 & 1 \\ 1 & 1 & 0 \end{vmatrix} \begin{vmatrix} a & 0 & 1 \\ 1 & 1 & 0 \end{vmatrix} \begin{vmatrix} a & 0 & c_1 \\ 2 & 1 & 1 \end{vmatrix} = 1 \begin{vmatrix} a & a & c \\ 12 & 1 & 1 \end{vmatrix} = 0$$

(Provided $(a_1 \ a_2 c_1) = 1$)

The rank condition is satisfied since we can construct at least one non-zero determinant of order $3 = (G-1)$.

Applying the counting rule $(K - M) \geq (G - 1)$, we see that the inequality sign holds: $4 \geq 3$; hence the investment function is over identified.

4a.

Consumption function: $C_t = \beta_0 + \beta_1 Y_{dt} \quad 0 < \beta_1 < 1$

Substitute equations (4) and (2) in equation (1):

$$C_t = \beta_0 + \beta_1 Y_{dt}$$

$$C_t = \beta_0 + \beta_1 (Y_t - T_t)$$

$$C_t = \beta_0 + \beta_1 (Y_t - (\alpha_0 + \alpha_1 Y_t))$$

$$C_t = \beta_0 + \beta_1 (Y_t - \alpha_0 - \alpha_1 Y_t)$$

$$C_t = \beta_0 + \beta_1 Y_t - \alpha_0 \beta_1 - \alpha_1 \beta_1 Y_t$$

$$C_t = \beta_0 + \beta_1 Y_t (1 - \alpha_1) - \alpha_0 \beta_1 \quad (7)$$

Substitute equations (7), (3) and (5) in equation (6)

$$Y_t = C_t + I_t + G_t$$

$$Y_t = \beta_0 + \beta_1 Y_t (1 - \alpha_1) - \alpha_0 \beta_1 + \gamma_0 + \gamma_1 r_t + \dots$$

$$Y_t - \beta_1 Y_t + \beta_1 Y_t \alpha_1 = \beta_0 - \alpha_0 \beta_1 + \gamma_0 + \gamma_1 r_t + \dots$$

$$Y_t (1 - \beta_1 + \beta_1 \alpha_1) = \beta_0 - \alpha_0 \beta_1 + \gamma_0 + \gamma_1 r_t + \dots$$

Answers for Even numbers of the Chapter Review questions

$$Y_t(1-\beta_1(1-\alpha_1)) = \beta_0 - \alpha_0\beta_1 + \gamma_0 + \gamma_1r_t +$$

$$Y_t = \frac{\beta_0 - \alpha_0\beta_1 + \gamma_0}{1-\beta_1(1-\alpha_1)} + \frac{\gamma_1}{1-\beta_1(1-\alpha_1)}r_t$$

Thus, the reduced form of the IS equation is:

$$Y_t = \frac{\beta_0 - \alpha_0\beta_1 + \gamma_0}{1-\beta_1(1-\alpha_1)} + \frac{\gamma_1}{1-\beta_1(1-\alpha_1)}r_t \quad (8)$$

Where, $\frac{\beta_0 - \alpha_0\beta_1 + \gamma_0}{1-\beta_1(1-\alpha_1)} = \beta_0^*$ and $\frac{\gamma_1}{1-\beta_1(1-\alpha_1)} = \beta_1^*$

Equation (8) is the equation of the IS, or goods market equilibrium, that is, it gives the combinations of the interest rate and level of income such that the goods market clears or is in equilibrium.

b. Such a result is unlikely because consumption depends on disposable income, which depends on national income Y , but the latter depends on r and as well as the other parameters entering in π_0 .

Therefore, unless we take into account all these influences, a simple regression of C on Y_d is bound to give biased and/or inconsistent estimates of β_0 and β_1 .

6.

(a) The endogenous variables are Y , C , Q and I . The predetermined variables are P , R , Y_{t-1} , C_{t-1} and Q_{t-1} , the first two of these being exogenous and the rest being lagged endogenous.

(b) This variable may account for inflation

(c) In this system all the equations are **over identified**.

Statistical Tables

References

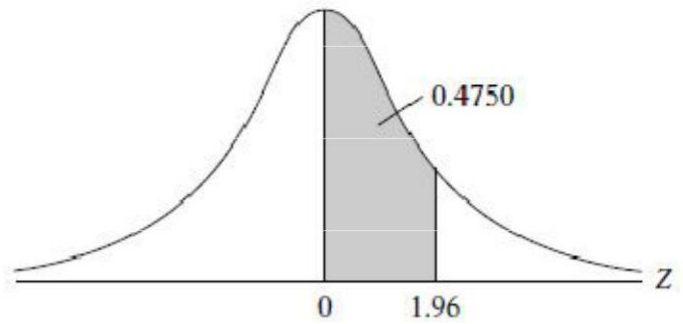
1. A., Koutsoyiannis, 2001. Theory of Econometrics: 2nd edition, Printed by Palgrave Raplika Pvt Ltd
2. Gujarati, D.N, 2004. Basic Econometrics: 4th edition, McGraw Hill Company
3. Thomas Andren, 2007. Econometrics: Thomas Andren and Ventus Publishing APS.
4. Badi H. Baltagi, 2008. Econometrics: 4th Edition, Springer-Verlag Berlin Heidelberg
5. Chang F. Lee et.al, 2000. Statistics for Business and Financial Economics: 2nd edition, World Scientific Publishing co.ptc.Ltd
6. Christopher Dougherty -. Introduction to Econometrics: 3rd edition, Oxford
7. Wooldrige.J, 2000. Introductory econometrics: A modern approach, 2nd edition.
8. Madala.G.S, 1988. Introduction to econometrics: 2nd edition, Macmillan Publishing Company
9. Jack Johnston and Jhon Dinardo, -. Econometric Methods: 4th edition, McGraw Hill
10. A. Colin Cameron and Pravin K. Trivedi, 2005. Microeconometrics: *Methods and Applications*, Cambridge University Press

AREAS UNDER THE STANDARDIZED NORMAL DISTRIBUTION

Example

$$\Pr(0 \leq Z \leq 1.96) = 0.4750$$

$$\Pr(Z \geq 1.96) = 0.5 - 0.4750 = 0.025$$



Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4454	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Note: This table gives the area in the right-hand tail of the distribution (i.e., $Z \geq 0$). But since the normal distribution is symmetrical about $Z = 0$, the area in the left-hand tail is the same as the area in the corresponding right-hand tail. For example, $P(-1.96 \leq Z \leq 0) = 0.4750$. Therefore, $P(-1.96 \leq Z \leq 1.96) = 2(0.4750) = 0.95$.

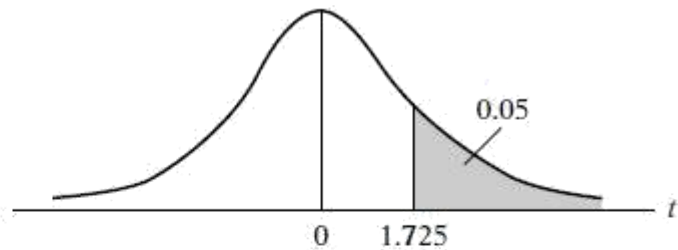
PERCENTAGE POINTS OF THE *t* DISTRIBUTION

Example

$$\Pr(t > 2.086) = 0.025$$

$$\Pr(t > 1.725) = 0.05 \quad \text{for } df = 20$$

$$\Pr(|t| > 1.725) = 0.10$$



Pr df	0.25 0.50	0.10 0.20	0.05 0.10	0.025 0.05	0.01 0.02	0.005 0.010	0.001 0.002
1	1.000	3.078	6.314	12.706	31.821	63.657	318.31
2	0.816	1.886	2.920	4.303	6.965	9.925	22.327
3	0.765	1.638	2.353	3.182	4.541	5.841	10.214
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	0.718	1.440	1.943	2.447	3.143	3.707	5.208
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785
8	0.706	1.397	1.860	2.306	2.896	3.355	4.501
9	0.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.700	1.372	1.812	2.228	2.764	3.169	4.144
11	0.697	1.363	1.796	2.201	2.718	3.106	4.025
12	0.695	1.356	1.782	2.179	2.681	3.055	3.930
13	0.694	1.350	1.771	2.160	2.650	3.012	3.852
14	0.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	0.690	1.337	1.746	2.120	2.583	2.921	3.686
17	0.689	1.333	1.740	2.110	2.567	2.898	3.646
18	0.688	1.330	1.734	2.101	2.552	2.878	3.610
19	0.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	0.686	1.323	1.721	2.080	2.518	2.831	3.527
22	0.686	1.321	1.717	2.074	2.508	2.819	3.505
23	0.685	1.319	1.714	2.069	2.500	2.807	3.485
24	0.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.684	1.316	1.708	2.060	2.485	2.787	3.450
26	0.684	1.315	1.706	2.056	2.479	2.779	3.435
27	0.684	1.314	1.703	2.052	2.473	2.771	3.421
28	0.683	1.313	1.701	2.048	2.467	2.763	3.408
29	0.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	0.681	1.303	1.684	2.021	2.423	2.704	3.307
60	0.679	1.296	1.671	2.000	2.390	2.660	3.232
120	0.677	1.289	1.658	1.980	2.358	2.617	3.160
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.090

Note: The smaller probability shown at the head of each column is the area in one tail; the larger probability is the area in both tails.

TABLE D.3 UPPER PERCENTAGE POINTS OF THE F DISTRIBUTION

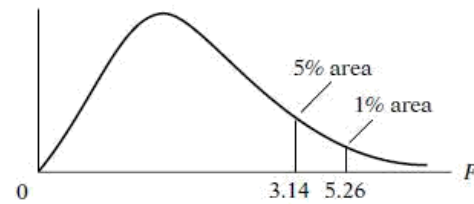
Example

$$\Pr(F > 1.59) = 0.25$$

$$\Pr(F > 2.42) = 0.10 \quad \text{for df } N_1 = 10$$

$$\Pr(F > 3.14) = 0.05 \quad \text{and } N_2 = 9$$

$$\Pr(F > 5.26) = 0.01$$



df for denom- inator N_2	df for numerator N_1												
	Pr	1	2	3	4	5	6	7	8	9	10	11	12
1	.25	5.83	7.50	8.20	8.58	8.82	8.98	9.10	9.19	9.26	9.32	9.36	9.41
	.10	39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2	60.5	60.7
	.05	161	200	216	225	230	234	237	239	241	242	243	244
	.01												
2	.25	2.57	3.00	3.15	3.23	3.28	3.31	3.34	3.35	3.37	3.38	3.39	3.39
	.10	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.40	9.41
	.05	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4
	.01	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4
3	.25	2.02	2.28	2.36	2.39	2.41	2.42	2.43	2.44	2.44	2.44	2.45	2.45
	.10	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.22
	.05	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74
	.01	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	27.1
4	.25	1.81	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08	2.08	2.08
	.10	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.91	3.90
	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91
	.01	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.4
5	.25	1.69	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89
	.10	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.28	3.27
	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.71	4.68
	.01	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.96	9.89
6	.25	1.62	1.76	1.78	1.79	1.79	1.78	1.78	1.78	1.77	1.77	1.77	1.77
	.10	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.92	2.90
	.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00
	.01	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72
7	.25	1.57	1.70	1.72	1.72	1.71	1.71	1.70	1.70	1.69	1.69	1.69	1.68
	.10	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.68	2.67
	.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57
	.01	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47
8	.25	1.54	1.66	1.67	1.66	1.66	1.65	1.64	1.64	1.63	1.63	1.63	1.62
	.10	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.52	2.50
	.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28
	.01	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67
9	.25	1.51	1.62	1.63	1.63	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.58
	.10	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.40	2.38
	.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07
	.01	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11

df for numerator N_1												df for denominator N_2	
15	20	24	30	40	50	60	100	120	200	500	∞	Pr	
9.49	9.58	9.63	9.67	9.71	9.74	9.76	9.78	9.80	9.82	9.84	9.85	.25	1
61.2	61.7	62.0	62.3	62.5	62.7	62.8	63.0	63.1	63.2	63.3	63.3	.10	
246	248	249	250	251	252	252	253	253	254	254	254	.05	
3.41	3.43	3.43	3.44	3.45	3.45	3.46	3.47	3.47	3.48	3.48	3.48	.25	2
0.42	0.44	0.45	0.46	0.47	0.47	0.47	0.48	0.48	0.49	0.49	0.49	.10	
19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	.05	
99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	.01	3
2.46	2.46	2.46	2.47	2.47	2.47	2.47	2.47	2.47	2.47	2.47	2.47	.25	
5.20	5.18	5.18	5.17	5.16	5.15	5.15	5.14	5.14	5.14	5.14	5.13	.10	
8.70	8.66	8.64	8.62	8.59	8.58	8.57	8.55	8.55	8.54	8.53	8.53	.05	4
26.9	26.7	26.6	26.5	26.4	26.4	26.3	26.2	26.2	26.2	26.1	26.1	.01	
2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	.25	
3.87	3.84	3.83	3.82	3.80	3.80	3.79	3.78	3.78	3.77	3.76	3.76	.10	5
5.06	5.00	5.77	5.75	5.72	5.70	5.69	5.68	5.66	5.65	5.64	5.63	.05	
11.2	11.0	13.9	13.8	13.7	13.7	13.7	13.6	13.6	13.5	13.5	13.5	.01	
1.09	1.00	1.00	1.00	1.00	1.00	1.07	1.07	1.07	1.07	1.07	1.07	.25	6
3.24	3.21	3.19	3.17	3.16	3.15	3.14	3.13	3.12	3.12	3.11	3.10	.10	
4.62	4.56	4.53	4.50	4.46	4.44	4.43	4.41	4.40	4.39	4.37	4.36	.05	
9.72	9.55	9.47	9.38	9.29	9.24	9.20	9.13	9.11	9.08	9.04	9.02	.01	7
1.76	1.76	1.75	1.75	1.75	1.75	1.74	1.74	1.74	1.74	1.74	1.74	.25	
2.07	2.04	2.02	2.00	2.70	2.77	2.76	2.75	2.74	2.73	2.70	2.72	.10	
3.94	3.87	3.84	3.81	3.77	3.75	3.74	3.71	3.70	3.69	3.68	3.67	.05	8
7.56	7.40	7.31	7.23	7.14	7.09	7.06	6.99	6.97	6.93	6.90	6.88	.01	
1.68	1.67	1.67	1.66	1.66	1.66	1.65	1.65	1.65	1.65	1.65	1.65	.25	
2.63	2.59	2.58	2.56	2.54	2.52	2.51	2.50	2.49	2.48	2.48	2.47	.10	9
3.51	3.44	3.41	3.38	3.34	3.32	3.30	3.27	3.27	3.25	3.24	3.23	.05	
6.31	6.16	6.07	5.99	5.91	5.86	5.82	5.75	5.74	5.70	5.67	5.65	.01	
1.62	1.61	1.60	1.60	1.59	1.59	1.59	1.53	1.53	1.53	1.53	1.53	.25	10
2.46	2.42	2.40	2.38	2.36	2.35	2.34	2.32	2.32	2.31	2.30	2.29	.10	
3.22	3.15	3.12	3.08	3.04	2.02	3.01	2.97	2.97	2.95	2.94	2.93	.05	
5.52	5.36	5.28	5.20	5.12	5.07	5.03	4.96	4.96	4.91	4.88	4.86	.01	11
1.57	1.56	1.56	1.55	1.55	1.54	1.54	1.53	1.53	1.53	1.53	1.53	.25	
2.34	2.30	2.28	2.25	2.23	2.22	2.21	2.19	2.18	2.17	2.17	2.16	.10	
3.01	2.94	2.90	2.86	2.83	2.80	2.79	2.76	2.75	2.73	2.72	2.71	.05	12
4.96	4.81	4.73	4.65	4.57	4.52	4.48	4.42	4.40	4.36	4.33	4.31	.01	

(Continued)

TABLE D.3 UPPER PERCENTAGE POINTS OF THE *F* DISTRIBUTION (Continued)

df for denom- inator N_2	df for numerator N_1												
	Pr	1	2	3	4	5	6	7	8	9	10	11	12
10	.25	1.49	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55	1.55	1.54
	.10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.30	2.28
	.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91
	.01	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71
11	.25	1.47	1.58	1.58	1.57	1.56	1.55	1.54	1.53	1.53	1.52	1.52	1.51
	.10	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.23	2.21
	.05	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79
	.01	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40
12	.25	1.46	1.56	1.56	1.55	1.54	1.53	1.52	1.51	1.51	1.50	1.50	1.49
	.10	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.17	2.15
	.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69
	.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16
13	.25	1.45	1.55	1.55	1.53	1.52	1.51	1.50	1.49	1.49	1.48	1.47	1.47
	.10	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.12	2.10
	.05	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60
	.01	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96
14	.25	1.44	1.53	1.53	1.52	1.51	1.50	1.49	1.48	1.47	1.46	1.46	1.45
	.10	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.08	2.05
	.05	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53
	.01	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80
15	.25	1.43	1.52	1.52	1.51	1.49	1.48	1.47	1.46	1.46	1.45	1.44	1.44
	.10	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.04	2.02
	.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48
	.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67
16	.25	1.42	1.51	1.51	1.50	1.48	1.47	1.46	1.45	1.44	1.44	1.44	1.43
	.10	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	2.01	1.99
	.05	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42
	.01	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.62	3.55
17	.25	1.42	1.51	1.50	1.49	1.47	1.46	1.45	1.44	1.43	1.43	1.42	1.41
	.10	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.98	1.96
	.05	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38
	.01	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.46
18	.25	1.41	1.50	1.49	1.48	1.46	1.45	1.44	1.43	1.42	1.42	1.41	1.40
	.10	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.96	1.93
	.05	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34
	.01	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43	3.37
19	.25	1.41	1.49	1.49	1.47	1.46	1.44	1.43	1.42	1.41	1.41	1.40	1.40
	.10	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.94	1.91
	.05	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31
	.01	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30
20	.25	1.40	1.49	1.48	1.46	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.39
	.10	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.92	1.89
	.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28
	.01	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23

df for numerator N_1												df for denomi- nator N_2	
15	20	24	30	40	50	60	100	120	200	500	(∞)	Pr	
1.53	1.52	1.52	1.51	1.51	1.50	1.50	1.49	1.49	1.49	1.48	1.48	.25	10
2.24	2.20	2.18	2.16	2.13	2.12	2.11	2.09	2.08	2.07	2.06	2.06	.10	
2.85	2.77	2.74	2.70	2.66	2.64	2.62	2.59	2.58	2.56	2.55	2.54	.05	
4.56	4.41	4.33	4.25	4.17	4.12	4.08	4.01	4.00	3.96	3.93	3.91	.01	
1.50	1.49	1.49	1.48	1.47	1.47	1.47	1.46	1.46	1.46	1.45	1.45	.25	11
2.17	2.12	2.10	2.08	2.05	2.04	2.03	2.00	2.00	1.99	1.98	1.97	.10	
2.72	2.65	2.61	2.57	2.53	2.51	2.49	2.46	2.45	2.43	2.42	2.40	.05	
4.25	4.10	4.02	3.94	3.86	3.81	3.78	3.71	3.69	3.66	3.62	3.60	.01	
1.48	1.47	1.46	1.45	1.45	1.44	1.44	1.43	1.43	1.43	1.42	1.42	.25	12
2.10	2.06	2.04	2.01	1.99	1.97	1.96	1.94	1.93	1.92	1.91	1.90	.10	
2.62	2.54	2.51	2.47	2.43	2.40	2.38	2.35	2.34	2.32	2.31	2.30	.05	
4.01	3.86	3.78	3.70	3.62	3.57	3.54	3.47	3.45	3.41	3.38	3.36	.01	
1.46	1.45	1.44	1.43	1.42	1.42	1.42	1.41	1.41	1.40	1.40	1.40	.25	13
2.05	2.01	1.98	1.96	1.93	1.92	1.90	1.88	1.88	1.86	1.85	1.85	.10	
2.53	2.46	2.42	2.38	2.34	2.31	2.30	2.26	2.25	2.23	2.22	2.21	.05	
3.82	3.66	3.59	3.51	3.43	3.38	3.34	3.27	3.25	3.22	3.19	3.17	.01	
1.44	1.43	1.42	1.41	1.41	1.40	1.40	1.39	1.39	1.39	1.38	1.38	.25	14
2.01	1.96	1.94	1.91	1.89	1.87	1.86	1.83	1.83	1.82	1.80	1.80	.10	
2.46	2.39	2.35	2.31	2.27	2.24	2.22	2.19	2.18	2.16	2.14	2.13	.05	
3.66	3.51	3.43	3.35	3.27	3.22	3.18	3.11	3.09	3.06	3.03	3.00	.01	
1.43	1.41	1.41	1.40	1.39	1.39	1.38	1.38	1.37	1.37	1.36	1.36	.25	15
1.97	1.92	1.90	1.87	1.85	1.83	1.82	1.79	1.79	1.77	1.76	1.76	.10	
2.40	2.33	2.29	2.25	2.20	2.18	2.16	2.12	2.11	2.10	2.08	2.07	.05	
3.52	3.37	3.29	3.21	3.13	3.08	3.05	2.98	2.96	2.92	2.89	2.87	.01	
1.41	1.40	1.39	1.38	1.37	1.37	1.36	1.35	1.35	1.35	1.34	1.34	.25	16
1.94	1.89	1.87	1.84	1.81	1.79	1.78	1.75	1.75	1.74	1.73	1.72	.10	
2.35	2.28	2.24	2.19	2.15	2.12	2.11	2.07	2.06	2.04	2.02	2.01	.05	
3.41	3.26	3.18	3.10	3.02	2.97	2.93	2.86	2.84	2.81	2.78	2.75	.01	
1.40	1.39	1.38	1.37	1.36	1.35	1.35	1.34	1.34	1.34	1.33	1.33	.25	17
1.91	1.86	1.84	1.81	1.78	1.76	1.75	1.73	1.72	1.71	1.69	1.69	.10	
2.31	2.23	2.19	2.15	2.10	2.08	2.06	2.02	2.01	1.99	1.97	1.96	.05	
3.31	3.16	3.08	3.00	2.92	2.87	2.83	2.76	2.75	2.71	2.68	2.65	.01	
1.39	1.38	1.37	1.36	1.35	1.34	1.34	1.33	1.33	1.32	1.32	1.32	.25	18
1.89	1.84	1.81	1.78	1.75	1.74	1.72	1.70	1.69	1.68	1.67	1.66	.10	
2.27	2.19	2.15	2.11	2.06	2.04	2.02	1.98	1.97	1.95	1.93	1.92	.05	
3.23	3.08	3.00	2.92	2.84	2.78	2.75	2.68	2.66	2.62	2.59	2.57	.01	
1.38	1.37	1.36	1.35	1.34	1.33	1.33	1.32	1.32	1.31	1.31	1.30	.25	19
1.86	1.81	1.79	1.76	1.73	1.71	1.70	1.67	1.67	1.65	1.64	1.63	.10	
2.23	2.16	2.11	2.07	2.03	2.00	1.98	1.94	1.93	1.91	1.89	1.88	.05	
3.15	3.00	2.92	2.84	2.76	2.71	2.67	2.60	2.58	2.55	2.51	2.49	.01	
1.37	1.36	1.35	1.34	1.33	1.33	1.32	1.31	1.31	1.30	1.30	1.29	.25	20
1.84	1.79	1.77	1.74	1.71	1.69	1.68	1.65	1.64	1.63	1.62	1.61	.10	
2.20	2.12	2.08	2.04	1.99	1.97	1.95	1.91	1.90	1.88	1.86	1.84	.05	
3.09	2.94	2.86	2.78	2.69	2.64	2.61	2.54	2.52	2.48	2.44	2.42	.01	

(Continued)

TABLE D.3 UPPER PERCENTAGE POINTS OF THE F DISTRIBUTION (Continued)

df for denom- inator N_2	df for numerator N_1												
	Pr	1	2	3	4	5	6	7	8	9	10	11	12
22	.25	1.40	1.48	1.47	1.45	1.44	1.42	1.41	1.40	1.39	1.39	1.38	1.37
	.10	2.95	2.50	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.88	1.86
	.05	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23
	.01	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12
24	.25	1.39	1.47	1.46	1.44	1.43	1.41	1.40	1.39	1.38	1.38	1.37	1.36
	.10	2.93	2.54	2.39	2.19	2.10	2.04	1.99	1.94	1.91	1.88	1.85	1.83
	.05	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.21	2.18
	.01	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.09	3.03
26	.25	1.38	1.46	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.37	1.36	1.35
	.10	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.84	1.81
	.05	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15
	.01	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.19	3.09	3.02	2.96
28	.25	1.38	1.46	1.45	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34
	.10	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.81	1.79
	.05	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12
	.01	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.95	2.90
30	.25	1.38	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.36	1.35	1.35	1.34
	.10	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.79	1.77
	.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09
	.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.91	2.84
40	.25	1.36	1.44	1.42	1.40	1.39	1.37	1.36	1.35	1.34	1.33	1.32	1.31
	.10	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.73	1.71
	.05	4.06	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00
	.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.73	2.66
60	.25	1.35	1.42	1.41	1.38	1.37	1.35	1.33	1.32	1.31	1.30	1.29	1.28
	.10	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.68	1.66
	.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92
	.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50
120	.25	1.34	1.40	1.39	1.37	1.35	1.33	1.31	1.30	1.29	1.28	1.27	1.26
	.10	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.62	1.60
	.05	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.87	1.83
	.01	6.05	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.40	2.34
200	.25	1.33	1.39	1.38	1.36	1.34	1.32	1.31	1.29	1.28	1.27	1.26	1.25
	.10	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66	1.63	1.60	1.57
	.05	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.84	1.80
	.01	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.34	2.27
∞	.25	1.32	1.39	1.37	1.35	1.33	1.31	1.29	1.28	1.27	1.25	1.24	1.24
	.10	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.57	1.55
	.05	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.79	1.75
	.01	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.25	2.18

df for numerator N_1													df for denominator N_2
15	20	24	30	40	50	60	100	120	200	500	∞	Pr	
1.36	1.34	1.33	1.32	1.31	1.31	1.30	1.30	1.30	1.29	1.29	1.28	.25	22
1.81	1.76	1.73	1.70	1.67	1.65	1.64	1.61	1.60	1.59	1.58	1.57	.10	
2.15	2.07	2.03	1.98	1.94	1.91	1.89	1.85	1.84	1.82	1.80	1.78	.05	
2.98	2.83	2.75	2.67	2.58	2.53	2.50	2.42	2.40	2.36	2.33	2.31	.01	
1.35	1.33	1.32	1.31	1.30	1.29	1.29	1.28	1.28	1.27	1.27	1.26	.25	24
1.78	1.73	1.70	1.67	1.64	1.62	1.61	1.58	1.57	1.56	1.54	1.53	.10	
2.11	2.03	1.98	1.94	1.89	1.86	1.84	1.80	1.79	1.77	1.75	1.73	.05	
2.89	2.74	2.66	2.58	2.49	2.44	2.40	2.33	2.31	2.27	2.24	2.21	.01	
1.34	1.32	1.31	1.30	1.29	1.28	1.28	1.26	1.26	1.26	1.25	1.25	.25	26
1.76	1.71	1.68	1.65	1.61	1.59	1.58	1.55	1.54	1.53	1.51	1.50	.10	
2.07	1.99	1.95	1.90	1.85	1.82	1.80	1.76	1.75	1.73	1.71	1.69	.05	
2.81	2.66	2.58	2.50	2.42	2.36	2.33	2.25	2.23	2.19	2.16	2.13	.01	
1.33	1.31	1.30	1.29	1.28	1.27	1.27	1.26	1.25	1.25	1.24	1.24	.25	28
1.74	1.69	1.66	1.63	1.59	1.57	1.56	1.53	1.52	1.50	1.49	1.48	.10	
2.04	1.96	1.91	1.87	1.82	1.79	1.77	1.73	1.71	1.69	1.67	1.65	.05	
2.75	2.60	2.52	2.44	2.35	2.30	2.26	2.19	2.17	2.13	2.09	2.06	.01	
1.32	1.30	1.29	1.28	1.27	1.26	1.26	1.25	1.24	1.24	1.23	1.23	.25	30
1.72	1.67	1.64	1.61	1.57	1.55	1.54	1.51	1.50	1.48	1.47	1.46	.10	
2.01	1.93	1.89	1.84	1.79	1.76	1.74	1.70	1.68	1.66	1.64	1.62	.05	
2.70	2.55	2.47	2.39	2.30	2.25	2.21	2.13	2.11	2.07	2.03	2.01	.01	
1.30	1.28	1.26	1.25	1.24	1.23	1.22	1.21	1.21	1.20	1.19	1.19	.25	40
1.66	1.61	1.57	1.54	1.51	1.48	1.47	1.43	1.42	1.41	1.39	1.38	.10	
1.92	1.84	1.79	1.74	1.69	1.66	1.64	1.59	1.58	1.55	1.53	1.51	.05	
2.52	2.37	2.29	2.20	2.11	2.06	2.02	1.94	1.92	1.87	1.83	1.80	.01	
1.27	1.25	1.24	1.22	1.21	1.20	1.19	1.17	1.17	1.16	1.15	1.15	.25	60
1.60	1.54	1.51	1.48	1.44	1.41	1.40	1.36	1.35	1.33	1.31	1.29	.10	
1.84	1.75	1.70	1.65	1.59	1.56	1.53	1.48	1.47	1.44	1.41	1.39	.05	
2.35	2.20	2.12	2.03	1.94	1.88	1.84	1.75	1.73	1.68	1.63	1.60	.01	
1.24	1.22	1.21	1.19	1.18	1.17	1.16	1.14	1.13	1.12	1.11	1.10	.25	120
1.55	1.48	1.45	1.41	1.37	1.34	1.32	1.27	1.26	1.24	1.21	1.19	.10	
1.75	1.66	1.61	1.55	1.50	1.46	1.43	1.37	1.35	1.32	1.28	1.25	.05	
2.19	2.03	1.95	1.86	1.76	1.70	1.66	1.56	1.53	1.48	1.42	1.38	.01	
1.23	1.21	1.20	1.18	1.16	1.14	1.12	1.11	1.10	1.09	1.08	1.06	.25	200
1.52	1.46	1.42	1.38	1.34	1.31	1.28	1.24	1.22	1.20	1.17	1.14	.10	
1.72	1.62	1.57	1.52	1.46	1.41	1.39	1.32	1.29	1.26	1.22	1.19	.05	
2.13	1.97	1.89	1.79	1.69	1.63	1.58	1.48	1.44	1.39	1.33	1.28	.01	
1.22	1.19	1.18	1.16	1.14	1.13	1.12	1.09	1.08	1.07	1.04	1.00	.25	∞
1.49	1.42	1.38	1.34	1.30	1.26	1.24	1.18	1.17	1.13	1.08	1.00	.10	
1.67	1.57	1.52	1.46	1.39	1.35	1.32	1.24	1.22	1.17	1.11	1.00	.05	
2.04	1.88	1.79	1.70	1.59	1.52	1.47	1.36	1.32	1.25	1.15	1.00	.01	

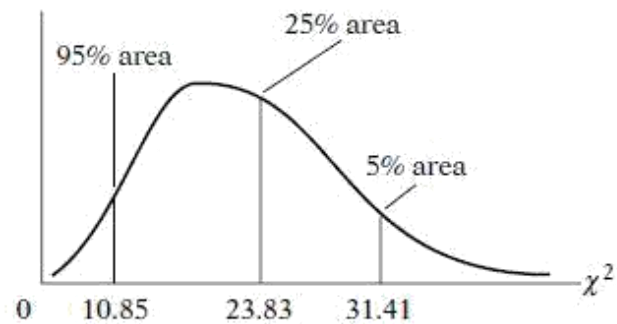
UPPER PERCENTAGE POINTS OF THE χ^2 DISTRIBUTION

Example

$$\Pr(\chi^2 > 10.85) = 0.95$$

$$\Pr(\chi^2 > 23.83) = 0.25 \quad \text{for df} = 20$$

$$\Pr(\chi^2 > 31.41) = 0.05$$



Degrees of freedom \ Pr	.995	.990	.975	.950	.900
1	392704×10^{-10}	157088×10^{-9}	982069×10^{-9}	393214×10^{-8}	.0157908
2	.0100251	.0201007	.0506356	.102587	.210720
3	.0717212	.114832	.215795	.351846	.584375
4	.206990	.297110	.484419	.710721	1.063623
5	.411740	.554300	.831211	1.145476	1.61031
6	.675727	.872085	1.237347	1.63539	2.20413
7	.989265	1.239043	1.68987	2.16735	2.83311
8	1.344419	1.646482	2.17973	2.73264	3.48954
9	1.734926	2.087912	2.70039	3.32511	4.16816
10	2.15585	2.55821	3.24697	3.94030	4.86518
11	2.60321	3.05347	3.81575	4.57481	5.57779
12	3.07382	3.57056	4.40379	5.22603	6.30380
13	3.56503	4.10691	5.00874	5.89186	7.04150
14	4.07468	4.66043	5.62872	6.57063	7.78953
15	4.60094	5.22935	6.26214	7.26094	8.54675
16	5.14224	5.81221	6.90766	7.96164	9.31223
17	5.69724	6.40776	7.56418	8.67176	10.0852
18	6.26481	7.01491	8.23075	9.39046	10.8649
19	6.84398	7.63273	8.90655	10.1170	11.6509
20	7.43386	8.26040	9.59083	10.8508	12.4426
21	8.03366	8.89720	10.28293	11.5913	13.2396
22	8.64272	9.54249	10.9823	12.3380	14.0415
23	9.26042	10.19567	11.6885	13.0905	14.8479
24	9.88623	10.8564	12.4011	13.8484	15.6587
25	10.5197	11.5240	13.1197	14.6114	16.4734
26	11.1603	12.1981	13.8439	15.3791	17.2919
27	11.8076	12.8786	14.5733	16.1513	18.1138
28	12.4613	13.5648	15.3079	16.9279	18.9392
29	13.1211	14.2565	16.0471	17.7083	19.7677
30	13.7867	14.9535	16.7908	18.4926	20.5992
40	20.7065	22.1643	24.4331	26.5093	29.0505
50	27.9907	29.7067	32.3574	34.7642	37.6886
60	35.5346	37.4848	40.4817	43.1879	46.4589
70	43.2752	45.4418	48.7576	51.7393	55.3290
80	51.1720	53.5400	57.1532	60.3915	64.2778
90	59.1963	61.7541	65.6466	69.1260	73.2912
100*	67.3276	70.0648	74.2219	77.9295	82.3581

*For df greater than 100 the expression $\sqrt{2\chi^2} - \sqrt{(2k-1)} = Z$ follows the standardized normal distribution, where k represents the degrees of freedom.

.750	.500	.250	.100	.050	.025	.010	.005
.1015308	.454937	1.32330	2.70554	3.84146	5.02389	6.63490	7.87944
.575364	1.38629	2.77259	4.60517	5.99147	7.37776	9.21034	10.5966
1.212534	2.36597	4.10835	6.25139	7.81473	9.34840	11.3449	12.8381
1.92255	3.35670	5.38527	7.77944	9.48773	11.1433	13.2767	14.8602
2.67460	4.35146	6.62568	9.23635	11.0705	12.8325	15.0863	16.7496
3.45460	5.34812	7.84080	10.6446	12.5916	14.4494	16.8119	18.5476
4.25485	6.34581	9.03715	12.0170	14.0671	16.0128	18.4753	20.2777
5.07064	7.34412	10.2188	13.3616	15.5073	17.5346	20.0902	21.9550
5.89983	8.34283	11.3887	14.6837	16.9190	19.0228	21.6660	23.5893
6.73720	9.34182	12.5489	15.9871	18.3070	20.4831	23.2093	25.1882
7.58412	10.3410	13.7007	17.2750	19.6751	21.9200	24.7250	26.7569
8.43842	11.3403	14.8454	18.5494	21.0261	23.3367	26.2170	28.2995
9.29906	12.3398	15.9839	19.8119	22.3621	24.7356	27.6883	29.8194
10.1653	13.3393	17.1170	21.0642	23.6848	26.1190	29.1413	31.3193
11.0365	14.3389	18.2451	22.3072	24.9958	27.4884	30.5779	32.8013
11.9122	15.3385	19.3688	23.5418	26.2962	28.8454	31.9999	34.2672
12.7919	16.3381	20.4887	24.7690	27.5871	30.1910	33.4087	35.7185
13.6753	17.3379	21.6049	25.9894	28.8683	31.5264	34.8053	37.1564
14.5620	18.3376	22.7178	27.2036	30.1435	32.8523	36.1908	38.5822
15.4518	19.3374	23.8277	28.4120	31.4104	34.1696	37.5662	39.9968
16.3444	20.3372	24.9348	29.6151	32.6705	35.4789	38.9321	41.4010
17.2396	21.3370	26.0393	30.8133	33.9244	36.7807	40.2894	42.7956
18.1373	22.3369	27.1413	32.0069	35.1725	38.0757	41.6384	44.1813
19.0372	23.3367	28.2412	33.1963	36.4151	39.3641	42.9798	45.5585
19.9393	24.3366	29.3389	34.3816	37.6525	40.6465	44.3141	46.9278
20.8434	25.3364	30.4345	35.5631	38.8852	41.9232	45.6417	48.2899
21.7494	26.3363	31.5284	36.7412	40.1133	43.1944	46.9630	49.6449
22.6572	27.3363	32.6205	37.9159	41.3372	44.4607	48.2782	50.9933
23.5666	28.3362	33.7109	39.0875	42.5569	45.7222	49.5879	52.3356
24.4776	29.3360	34.7998	40.2560	43.7729	46.9792	50.8922	53.6720
33.6603	39.3354	45.6160	51.8050	55.7585	59.3417	63.6907	66.7659
42.9421	49.3349	56.3336	63.1671	67.5048	71.4202	76.1539	79.4900
52.2938	59.3347	66.9814	74.3970	79.0819	83.2976	88.3794	91.9517
61.6983	69.3344	77.5766	85.5271	90.5312	95.0231	100.425	104.215
71.1445	79.3343	88.1303	96.5782	101.879	106.629	112.329	116.321
80.6247	89.3342	98.6499	107.565	113.145	118.136	124.116	128.299
90.1332	99.3341	109.141	118.498	124.342	129.561	135.807	140.169

Source: "The Standard Normal Distribution Table," *Journal of the Royal Statistical Society*, 1935, 97, 172-176.

DURBIN-WATSON d STATISTIC: SIGNIFICANCE POINTS OF d_L AND d_U AT 0.05 LEVEL OF SIGNIFICANCE

n	$k' = 1$		$k' = 2$		$k' = 3$		$k' = 4$		$k' = 5$		$k' = 6$		$k' = 7$		$k' = 8$		$k' = 9$		$k' = 10$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
6	0.610	1.400	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
7	0.700	1.356	0.467	1.896	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
8	0.763	1.332	0.559	1.777	0.368	2.287	—	—	—	—	—	—	—	—	—	—	—	—	—	—
9	0.824	1.320	0.629	1.699	0.455	2.128	0.296	2.588	—	—	—	—	—	—	—	—	—	—	—	—
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822	—	—	—	—	—	—	—	—	—	—
11	0.927	1.324	0.659	1.604	0.595	1.928	0.444	2.283	0.316	2.645	0.203	3.005	—	—	—	—	—	—	—	—
12	0.971	1.331	0.682	1.579	0.658	1.864	0.512	2.177	0.379	2.506	0.268	2.932	0.171	3.149	—	—	—	—	—	—
13	1.010	1.340	0.661	1.562	0.715	1.816	0.574	2.094	0.445	2.390	0.328	2.692	0.230	2.985	0.147	3.266	—	—	—	—
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.030	0.505	2.296	0.389	2.572	0.286	2.848	0.200	3.111	0.127	3.360	—	—
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220	0.447	2.472	0.343	2.727	0.251	2.979	0.175	3.216	0.111	3.438
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157	0.502	2.388	0.398	2.624	0.304	2.860	0.222	3.090	0.155	3.304
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104	0.554	2.318	0.451	2.537	0.356	2.757	0.272	2.975	0.198	3.184
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060	0.603	2.257	0.502	2.461	0.407	2.667	0.321	2.873	0.244	3.073
19	1.180	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023	0.649	2.206	0.549	2.396	0.456	2.589	0.369	2.783	0.290	2.974
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991	0.692	2.162	0.595	2.339	0.502	2.521	0.416	2.704	0.336	2.885
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.829	1.964	0.732	2.124	0.637	2.290	0.547	2.460	0.461	2.633	0.380	2.808
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940	0.769	2.090	0.677	2.246	0.588	2.407	0.504	2.571	0.424	2.734
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.785	0.895	1.920	0.804	2.061	0.715	2.208	0.628	2.360	0.545	2.514	0.465	2.670
24	1.273	1.446	1.188	1.546	1.101	1.656	1.013	1.775	0.925	1.902	0.837	2.035	0.751	2.174	0.666	2.318	0.584	2.464	0.506	2.613
25	1.289	1.454	1.206	1.550	1.123	1.654	1.039	1.767	0.952	1.886	0.869	2.012	0.784	2.144	0.702	2.290	0.621	2.419	0.544	2.560
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873	0.897	1.992	0.816	2.117	0.735	2.246	0.657	2.379	0.581	2.513
27	1.316	1.469	1.240	1.556	1.162	1.651	1.084	1.753	1.004	1.861	0.925	1.974	0.845	2.093	0.767	2.216	0.691	2.342	0.616	2.470
28	1.328	1.476	1.255	1.560	1.181	1.650	1.104	1.747	1.028	1.850	0.951	1.958	0.874	2.071	0.798	2.188	0.723	2.309	0.650	2.431
29	1.341	1.483	1.270	1.563	1.198	1.650	1.124	1.743	1.050	1.841	0.975	1.944	0.900	2.052	0.826	2.164	0.753	2.278	0.682	2.396
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.071	1.833	0.998	1.931	0.926	2.034	0.854	2.141	0.782	2.251	0.712	2.363
31	1.363	1.496	1.297	1.570	1.229	1.650	1.160	1.735	1.090	1.825	1.020	1.920	0.950	2.018	0.879	2.120	0.810	2.226	0.741	2.333
32	1.373	1.502	1.309	1.574	1.244	1.650	1.177	1.732	1.109	1.819	1.041	1.909	0.972	2.004	0.904	2.102	0.836	2.203	0.769	2.306
33	1.383	1.508	1.321	1.577	1.258	1.651	1.193	1.730	1.127	1.813	1.061	1.900	0.994	1.991	0.927	2.085	0.861	2.181	0.795	2.281
34	1.393	1.514	1.333	1.580	1.271	1.652	1.208	1.728	1.144	1.808	1.080	1.891	1.015	1.979	0.950	2.069	0.885	2.162	0.821	2.257
35	1.402	1.519	1.343	1.584	1.283	1.653	1.222	1.726	1.160	1.803	1.097	1.884	1.034	1.967	0.971	2.054	0.908	2.144	0.845	2.236
36	1.411	1.525	1.354	1.587	1.295	1.654	1.236	1.724	1.175	1.799	1.114	1.877	1.053	1.957	0.991	2.041	0.930	2.127	0.868	2.216
37	1.419	1.530	1.364	1.590	1.307	1.655	1.249	1.723	1.190	1.795	1.131	1.870	1.071	1.948	1.011	2.029	0.951	2.112	0.891	2.198
38	1.427	1.535	1.373	1.594	1.318	1.656	1.261	1.722	1.204	1.792	1.146	1.864	1.088	1.939	1.029	2.017	0.970	2.098	0.912	2.180
39	1.435	1.540	1.382	1.597	1.328	1.658	1.273	1.722	1.218	1.789	1.161	1.859	1.104	1.932	1.047	2.007	0.990	2.085	0.932	2.164
40	1.442	1.544	1.391	1.600	1.338	1.659	1.285	1.721	1.230	1.786	1.175	1.854	1.120	1.924	1.064	1.997	1.008	2.072	0.952	2.149
45	1.475	1.566	1.430	1.615	1.383	1.666	1.336	1.720	1.287	1.776	1.238	1.835	1.180	1.895	1.139	1.958	1.080	2.022	1.038	2.088
50	1.503	1.585	1.462	1.628	1.421	1.674	1.378	1.721	1.335	1.771	1.291	1.822	1.246	1.875	1.201	1.930	1.156	1.986	1.110	2.044
55	1.528	1.601	1.490	1.641	1.452	1.681	1.414	1.724	1.374	1.768	1.334	1.814	1.294	1.861	1.253	1.909	1.212	1.959	1.170	2.010
60	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767	1.372	1.808	1.335	1.850	1.298	1.894	1.260	1.939	1.222	1.984
65	1.567	1.629	1.536	1.662	1.503	1.696	1.471	1.731	1.438	1.767	1.404	1.805	1.370	1.843	1.336	1.882	1.301	1.923	1.266	1.964
70	1.583	1.641	1.554	1.672	1.525	1.703	1.494	1.735	1.464	1.768	1.433	1.802	1.401	1.837	1.369	1.873	1.337	1.910	1.305	1.948
75	1.598	1.652	1.571	1.680	1.543	1.709	1.515	1.739	1.487	1.770	1.458	1.801	1.428	1.834	1.399	1.867	1.369	1.901	1.339	1.935
80	1.611	1.662	1.586	1.688	1.560	1.715	1.534	1.743	1.507	1.772	1.480	1.801	1.453	1.831	1.425	1.861	1.397	1.893	1.369	1.925
85	1.624	1.671	1.600	1.696	1.575	1.721	1.550	1.747	1.525	1.774	1.500	1.801	1.474	1.829	1.448	1.857	1.422	1.886	1.396	1.916
90	1.635	1.679	1.612	1.703	1.589	1.726	1.566	1.751	1.542	1.776	1.518	1.801	1.494	1.827	1.469	1.854	1.445	1.881	1.420	1.909
95	1.645	1.687	1.623	1.709	1.602	1.732	1.579	1.755	1.557	1.778	1.535	1.802	1.512	1.827	1.489	1.852	1.465	1.877	1.442	1.903
100	1.654	1.694	1.634	1.715	1.613	1.736	1.592	1.758	1.571	1.780	1.550	1.803	1.528	1.826	1.506	1.850	1.484	1.874	1.462	1.898
150	1.720	1.746	1.706	1.760	1.693	1.774	1.679	1.788	1.665	1.802	1.651	1.817	1.637	1.832	1.622	1.847	1.608	1.862	1.594	1.877
200	1.758	1.778	1.748	1.789	1.738	1.799	1.728	1.810	1.718	1.820	1.707	1.831	1.697	1.841	1.686	1.852	1.675	1.863	1.665	1.874

	$k' = 11$		$k' = 12$		$k' = 13$		$k' = 14$		$k' = 15$		$k' = 16$		$k' = 17$		$k' = 18$		$k' = 19$		$k' = 20$	
n	d_i	d_{ij}	d_i	d_{ij}	d_i	d_{ij}	d_i	d_{ij}	d_i	d_{ij}	d_i	d_{ij}	d_i	d_{ij}	d_i	d_{ij}	d_i	d_{ij}	d_i	d_{ij}
10	0.098	3.503	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
17	0.138	3.578	0.087	3.557	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
18	0.177	3.265	0.123	3.441	0.070	3.603	—	—	—	—	—	—	—	—	—	—	—	—	—	—
19	0.220	3.159	0.180	3.335	0.111	3.498	0.070	3.642	—	—	—	—	—	—	—	—	—	—	—	—
20	0.263	3.063	0.200	3.234	0.146	3.305	0.100	3.512	0.083	3.673	—	—	—	—	—	—	—	—	—	—
21	0.307	2.970	0.240	3.141	0.182	3.300	0.132	3.448	0.091	3.583	0.058	3.700	—	—	—	—	—	—	—	—
22	0.340	2.897	0.291	3.057	0.220	3.211	0.186	3.358	0.120	3.405	0.083	3.610	0.052	3.731	—	—	—	—	—	—
23	0.391	2.820	0.322	2.979	0.259	3.129	0.232	3.272	0.153	3.409	0.110	3.555	0.070	3.650	0.048	3.753	—	—	—	—
24	0.431	2.761	0.362	2.908	0.294	3.053	0.239	3.193	0.186	3.327	0.141	3.454	0.101	3.572	0.070	3.678	0.044	3.773	—	—
25	0.470	2.702	0.400	2.844	0.335	2.903	0.275	3.119	0.221	3.251	0.172	3.376	0.130	3.494	0.094	3.604	0.035	3.702	0.041	3.790
26	0.508	2.649	0.438	2.784	0.373	2.819	0.312	3.051	0.256	3.179	0.205	3.303	0.160	3.420	0.120	3.531	0.087	3.632	0.060	3.724
27	0.544	2.600	0.475	2.730	0.400	2.850	0.348	2.987	0.291	3.112	0.238	3.233	0.191	3.310	0.140	3.460	0.112	3.563	0.081	3.659
28	0.578	2.555	0.510	2.680	0.445	2.805	0.383	2.926	0.325	3.050	0.271	3.168	0.222	3.283	0.178	3.392	0.138	3.495	0.104	3.592
29	0.612	2.515	0.544	2.634	0.470	2.755	0.418	2.874	0.350	2.992	0.305	3.107	0.254	3.219	0.208	3.327	0.158	3.431	0.120	3.529
30	0.643	2.477	0.577	2.592	0.512	2.708	0.451	2.823	0.382	2.937	0.337	3.050	0.288	3.180	0.238	3.266	0.195	3.368	0.156	3.465
31	0.674	2.443	0.614	2.553	0.545	2.665	0.484	2.776	0.425	2.887	0.370	2.986	0.317	3.103	0.269	3.204	0.224	3.309	0.183	3.406
32	0.705	2.411	0.633	2.517	0.570	2.625	0.515	2.733	0.457	2.840	0.401	2.940	0.349	3.050	0.299	3.153	0.253	3.252	0.211	3.348
33	0.731	2.382	0.663	2.484	0.606	2.588	0.546	2.692	0.488	2.795	0.432	2.899	0.379	3.000	0.329	3.100	0.283	3.198	0.236	3.293
34	0.750	2.355	0.695	2.454	0.634	2.554	0.575	2.654	0.510	2.754	0.462	2.854	0.409	2.954	0.359	3.051	0.312	3.147	0.267	3.240
35	0.783	2.330	0.722	2.425	0.662	2.521	0.604	2.619	0.547	2.718	0.492	2.813	0.439	2.910	0.388	3.005	0.340	3.099	0.295	3.190
36	0.808	2.306	0.743	2.398	0.680	2.492	0.631	2.586	0.575	2.690	0.520	2.774	0.467	2.868	0.417	2.961	0.360	3.053	0.323	3.142
37	0.831	2.285	0.772	2.374	0.714	2.464	0.657	2.555	0.602	2.643	0.548	2.738	0.495	2.829	0.445	2.920	0.397	3.009	0.351	3.097
38	0.854	2.265	0.798	2.351	0.730	2.438	0.683	2.526	0.628	2.614	0.575	2.703	0.522	2.792	0.472	2.890	0.424	2.968	0.379	3.054
39	0.875	2.240	0.819	2.329	0.763	2.413	0.707	2.499	0.653	2.585	0.600	2.671	0.549	2.757	0.499	2.843	0.451	2.929	0.404	3.013
40	0.896	2.228	0.840	2.309	0.785	2.391	0.731	2.473	0.678	2.557	0.626	2.641	0.575	2.724	0.525	2.809	0.477	2.892	0.430	2.974
45	0.980	2.158	0.903	2.225	0.887	2.298	0.838	2.367	0.700	2.403	0.740	2.512	0.692	2.506	0.644	2.659	0.593	2.793	0.550	2.907
50	1.064	2.103	1.019	2.163	0.973	2.225	0.927	2.287	0.682	2.350	0.836	2.414	0.792	2.479	0.747	2.544	0.703	2.610	0.660	2.675
55	1.120	2.062	1.087	2.116	1.045	2.170	1.003	2.226	0.661	2.281	0.819	2.338	0.777	2.396	0.836	2.454	0.795	2.512	0.754	2.571
60	1.184	2.031	1.145	2.079	1.106	2.127	1.058	2.177	1.029	2.227	0.990	2.278	0.951	2.330	0.913	2.382	0.874	2.434	0.836	2.487
65	1.231	2.006	1.195	2.049	1.160	2.093	1.124	2.138	1.088	2.183	1.052	2.220	1.018	2.276	0.980	2.322	0.944	2.371	0.906	2.419
70	1.272	1.986	1.239	2.026	1.206	2.068	1.172	2.106	1.139	2.143	1.105	2.189	1.072	2.232	1.038	2.275	1.005	2.318	0.971	2.362
75	1.308	1.970	1.277	2.006	1.247	2.043	1.215	2.080	1.184	2.118	1.153	2.156	1.121	2.195	1.090	2.235	1.058	2.275	1.027	2.315
80	1.340	1.957	1.311	1.991	1.280	2.024	1.253	2.059	1.224	2.093	1.195	2.129	1.165	2.185	1.136	2.201	1.109	2.236	1.076	2.275
85	1.369	1.946	1.342	1.977	1.315	2.009	1.287	2.040	1.260	2.073	1.232	2.105	1.205	2.139	1.177	2.172	1.143	2.206	1.121	2.241
90	1.396	1.937	1.369	1.966	1.344	1.995	1.318	2.026	1.292	2.065	1.266	2.095	1.240	2.116	1.213	2.149	1.187	2.179	1.160	2.211
95	1.418	1.929	1.394	1.956	1.370	1.984	1.345	2.012	1.321	2.040	1.296	2.068	1.271	2.097	1.247	2.126	1.222	2.156	1.197	2.186
100	1.430	1.923	1.418	1.948	1.393	1.974	1.371	2.000	1.347	2.028	1.324	2.053	1.301	2.080	1.277	2.109	1.253	2.135	1.220	2.164
150	1.579	1.892	1.564	1.908	1.550	1.924	1.535	1.940	1.519	1.955	1.504	1.972	1.489	1.989	1.474	2.006	1.458	2.023	1.443	2.040
200	1.654	1.885	1.643	1.896	1.632	1.908	1.621	1.919	1.610	1.931	1.599	1.943	1.589	1.955	1.576	1.967	1.565	1.979	1.554	1.991

DURBIN-WATSON d STATISTIC: SIGNIFICANCE POINTS OF d_L AND d_U AT 0.01 LEVEL OF SIGNIFICANCE

n	$k' = 1$		$k' = 2$		$k' = 3$		$k' = 4$		$k' = 5$		$k' = 6$		$k' = 7$		$k' = 8$		$k' = 9$		$k' = 10$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
6	0.390	1.142	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
7	0.435	1.036	0.294	1.676	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
8	0.497	1.003	0.345	1.489	0.229	2.102	—	—	—	—	—	—	—	—	—	—	—	—	—	—
9	0.554	0.998	0.408	1.389	0.279	1.875	0.183	2.433	—	—	—	—	—	—	—	—	—	—	—	—
10	0.604	1.001	0.468	1.333	0.340	1.733	0.230	2.193	0.150	2.690	—	—	—	—	—	—	—	—	—	—
11	0.653	1.010	0.519	1.297	0.396	1.640	0.286	2.030	0.193	2.453	0.124	2.892	—	—	—	—	—	—	—	—
12	0.697	1.023	0.569	1.274	0.449	1.575	0.339	1.913	0.244	2.280	0.164	2.865	0.105	3.053	—	—	—	—	—	—
13	0.738	1.038	0.616	1.261	0.499	1.526	0.391	1.826	0.294	2.150	0.211	2.490	0.140	2.838	0.090	3.182	—	—	—	—
14	0.776	1.054	0.660	1.254	0.547	1.490	0.441	1.757	0.343	2.049	0.257	2.354	0.183	2.667	0.122	2.981	0.078	3.287	—	—
15	0.811	1.070	0.700	1.252	0.591	1.464	0.488	1.704	0.391	1.967	0.303	2.244	0.226	2.530	0.161	2.817	0.107	3.101	0.068	3.374
16	0.844	1.086	0.737	1.252	0.633	1.446	0.532	1.663	0.437	1.900	0.349	2.153	0.269	2.416	0.200	2.681	0.142	2.944	0.094	3.201
17	0.874	1.102	0.772	1.255	0.672	1.432	0.574	1.630	0.480	1.847	0.393	2.078	0.313	2.319	0.241	2.566	0.179	2.811	0.127	3.053
18	0.902	1.118	0.805	1.259	0.708	1.422	0.613	1.604	0.522	1.803	0.435	2.015	0.355	2.238	0.282	2.467	0.216	2.697	0.160	2.925
19	0.928	1.132	0.835	1.265	0.742	1.415	0.650	1.584	0.561	1.767	0.476	1.963	0.396	2.169	0.322	2.381	0.255	2.597	0.196	2.813
20	0.952	1.147	0.863	1.271	0.773	1.411	0.685	1.567	0.598	1.737	0.515	1.918	0.436	2.110	0.362	2.308	0.294	2.510	0.232	2.714
21	0.975	1.161	0.890	1.277	0.803	1.408	0.719	1.554	0.633	1.712	0.552	1.881	0.474	2.059	0.400	2.244	0.331	2.434	0.268	2.625
22	0.997	1.174	0.914	1.284	0.831	1.407	0.749	1.543	0.667	1.691	0.587	1.849	0.510	2.015	0.437	2.189	0.368	2.367	0.304	2.548
23	1.018	1.187	0.938	1.291	0.858	1.407	0.777	1.534	0.698	1.673	0.620	1.821	0.545	1.977	0.473	2.140	0.404	2.308	0.340	2.479
24	1.037	1.199	0.960	1.298	0.882	1.407	0.805	1.528	0.728	1.658	0.652	1.797	0.578	1.944	0.507	2.097	0.439	2.255	0.375	2.417
25	1.055	1.211	0.981	1.305	0.906	1.409	0.831	1.523	0.756	1.645	0.682	1.776	0.610	1.915	0.540	2.059	0.473	2.209	0.409	2.362
26	1.072	1.222	1.001	1.312	0.928	1.411	0.855	1.519	0.783	1.635	0.711	1.759	0.640	1.889	0.572	2.026	0.505	2.168	0.441	2.313
27	1.089	1.233	1.019	1.319	0.949	1.413	0.878	1.515	0.808	1.626	0.738	1.743	0.669	1.867	0.602	1.997	0.536	2.131	0.473	2.269
28	1.104	1.244	1.037	1.325	0.969	1.415	0.900	1.513	0.832	1.618	0.764	1.729	0.696	1.847	0.630	1.970	0.566	2.098	0.504	2.229
29	1.119	1.254	1.054	1.332	0.988	1.418	0.921	1.512	0.855	1.611	0.788	1.718	0.723	1.830	0.658	1.947	0.595	2.068	0.533	2.193
30	1.133	1.263	1.070	1.339	1.006	1.421	0.941	1.511	0.877	1.606	0.812	1.707	0.748	1.814	0.684	1.925	0.622	2.041	0.562	2.160
31	1.147	1.273	1.085	1.345	1.023	1.425	0.960	1.510	0.897	1.601	0.834	1.698	0.772	1.800	0.710	1.906	0.649	2.017	0.589	2.131
32	1.160	1.282	1.100	1.352	1.040	1.428	0.979	1.510	0.917	1.597	0.856	1.690	0.794	1.788	0.734	1.889	0.674	1.995	0.615	2.104
33	1.172	1.291	1.114	1.358	1.055	1.432	0.996	1.510	0.936	1.594	0.876	1.683	0.816	1.776	0.757	1.874	0.698	1.975	0.641	2.080
34	1.184	1.299	1.128	1.364	1.070	1.435	1.012	1.511	0.954	1.591	0.896	1.677	0.837	1.766	0.779	1.860	0.722	1.957	0.665	2.057
35	1.195	1.307	1.140	1.370	1.085	1.439	1.029	1.512	0.971	1.589	0.914	1.671	0.857	1.757	0.800	1.847	0.744	1.940	0.689	2.037
36	1.206	1.315	1.153	1.376	1.098	1.442	1.043	1.513	0.988	1.588	0.932	1.666	0.877	1.749	0.821	1.836	0.766	1.925	0.711	2.018
37	1.217	1.323	1.165	1.382	1.112	1.446	1.058	1.514	1.004	1.586	0.950	1.662	0.895	1.742	0.841	1.825	0.787	1.911	0.733	2.001
38	1.227	1.330	1.176	1.388	1.124	1.449	1.072	1.515	1.019	1.585	0.966	1.658	0.913	1.735	0.860	1.816	0.807	1.899	0.754	1.985
39	1.237	1.337	1.187	1.393	1.137	1.453	1.085	1.517	1.034	1.584	0.982	1.655	0.930	1.729	0.878	1.807	0.826	1.887	0.774	1.970
40	1.246	1.344	1.198	1.399	1.148	1.457	1.098	1.518	1.048	1.584	0.997	1.652	0.946	1.724	0.895	1.799	0.844	1.876	0.749	1.956
45	1.288	1.376	1.245	1.423	1.201	1.474	1.156	1.528	1.111	1.584	1.065	1.643	1.019	1.704	0.974	1.768	0.927	1.834	0.881	1.902
50	1.324	1.403	1.285	1.446	1.245	1.491	1.205	1.538	1.164	1.587	1.123	1.639	1.081	1.692	1.039	1.748	0.997	1.895	0.955	1.864
55	1.356	1.427	1.320	1.466	1.284	1.506	1.247	1.548	1.209	1.592	1.172	1.638	1.134	1.685	1.095	1.734	1.057	1.785	1.018	1.837
60	1.383	1.449	1.350	1.484	1.317	1.520	1.283	1.558	1.249	1.598	1.214	1.639	1.179	1.682	1.144	1.726	1.109	1.771	1.072	1.817
65	1.407	1.468	1.377	1.500	1.346	1.534	1.315	1.568	1.283	1.604	1.251	1.642	1.218	1.690	1.196	1.720	1.153	1.761	1.120	1.802
70	1.429	1.485	1.400	1.515	1.372	1.546	1.343	1.578	1.313	1.611	1.283	1.645	1.253	1.680	1.223	1.716	1.192	1.754	1.162	1.792
75	1.448	1.501	1.422	1.529	1.395	1.557	1.368	1.587	1.340	1.617	1.313	1.649	1.284	1.682	1.256	1.714	1.227	1.748	1.199	1.783
80	1.466	1.515	1.441	1.541	1.416	1.568	1.390	1.595	1.364	1.624	1.338	1.653	1.312	1.683	1.285	1.714	1.259	1.745	1.232	1.777
85	1.482	1.528	1.458	1.553	1.435	1.578	1.411	1.603	1.386	1.630	1.362	1.657	1.337	1.685	1.312	1.714	1.287	1.743	1.262	1.773
90	1.496	1.540	1.474	1.563	1.452	1.587	1.429	1.611	1.406	1.636	1.383	1.661	1.360	1.687	1.336	1.714	1.312	1.741	1.288	1.769
95	1.510	1.552	1.489	1.573	1.468	1.596	1.446	1.618	1.425	1.642	1.403	1.666	1.381	1.690	1.358	1.715	1.336	1.741	1.313	1.767
100	1.522	1.562	1.503	1.583	1.482	1.604	1.462	1.625	1.441	1.647	1.421	1.670	1.400	1.693	1.378	1.717	1.357	1.741	1.335	1.765
150	1.611	1.637	1.598	1.651	1.584	1.665	1.571	1.679	1.557	1.693	1.543	1.708	1.530	1.722	1.515	1.737	1.501	1.752	1.486	1.767
200	1.664	1.684	1.653	1.693	1.643	1.704	1.633	1.715	1.623	1.725	1.613	1.735	1.603	1.746	1.592	1.757	1.582	1.768	1.571	1.779

<i>n</i>	<i>k'</i> = 11		<i>k'</i> = 12		<i>k'</i> = 13		<i>k'</i> = 14		<i>k'</i> = 15		<i>k'</i> = 16		<i>k'</i> = 17		<i>k'</i> = 18		<i>k'</i> = 19		<i>k'</i> = 20	
	<i>d_L</i>	<i>d_U</i>	<i>d_L</i>	<i>d_U</i>	<i>d_L</i>	<i>d_U</i>	<i>d_L</i>	<i>d_U</i>	<i>d_L</i>	<i>d_U</i>	<i>d_L</i>	<i>d_U</i>	<i>d_L</i>	<i>d_U</i>	<i>d_L</i>	<i>d_U</i>	<i>d_L</i>	<i>d_U</i>	<i>d_L</i>	<i>d_U</i>
16	0.060	3.446	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
17	0.084	3.286	0.053	3.506	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
18	0.113	3.146	0.075	3.358	0.047	3.357	—	—	—	—	—	—	—	—	—	—	—	—	—	—
19	0.145	3.023	0.102	3.227	0.067	3.420	0.043	3.601	—	—	—	—	—	—	—	—	—	—	—	—
20	0.178	2.914	0.131	3.109	0.092	3.297	0.061	3.474	0.038	3.639	—	—	—	—	—	—	—	—	—	—
21	0.212	2.817	0.162	3.004	0.119	3.185	0.084	3.358	0.055	3.521	0.035	3.671	—	—	—	—	—	—	—	—
22	0.246	2.729	0.194	2.909	0.148	3.084	0.109	3.252	0.077	3.412	0.050	3.562	0.032	3.700	—	—	—	—	—	—
23	0.281	2.651	0.227	2.822	0.178	2.991	0.136	3.155	0.100	3.311	0.070	3.459	0.046	3.597	0.029	3.725	—	—	—	—
24	0.315	2.580	0.260	2.744	0.209	2.906	0.165	3.065	0.125	3.218	0.092	3.363	0.065	3.501	0.043	3.629	0.027	3.747	—	—
25	0.348	2.517	0.292	2.674	0.240	2.829	0.194	2.982	0.152	3.131	0.116	3.274	0.085	3.410	0.060	3.538	0.039	3.657	0.025	3.766
26	0.381	2.460	0.324	2.610	0.272	2.758	0.224	2.906	0.180	3.050	0.141	3.191	0.107	3.325	0.079	3.452	0.055	3.572	0.036	3.682
27	0.413	2.409	0.356	2.552	0.303	2.694	0.253	2.836	0.208	2.976	0.167	3.113	0.131	3.245	0.100	3.371	0.073	3.490	0.051	3.602
28	0.444	2.363	0.387	2.499	0.333	2.635	0.283	2.772	0.237	2.907	0.194	3.040	0.156	3.169	0.122	3.294	0.093	3.412	0.068	3.524
29	0.474	2.321	0.417	2.451	0.363	2.582	0.313	2.713	0.266	2.843	0.222	2.972	0.182	3.098	0.146	3.220	0.114	3.338	0.087	3.450
30	0.503	2.283	0.447	2.407	0.393	2.533	0.342	2.659	0.294	2.785	0.249	2.909	0.208	3.032	0.171	3.152	0.137	3.267	0.107	3.379
31	0.531	2.248	0.475	2.367	0.422	2.487	0.371	2.609	0.322	2.730	0.277	2.851	0.234	2.970	0.196	3.087	0.160	3.201	0.128	3.311
32	0.558	2.216	0.503	2.330	0.450	2.446	0.399	2.563	0.350	2.680	0.304	2.797	0.261	2.912	0.221	3.026	0.184	3.137	0.151	3.246
33	0.585	2.187	0.530	2.296	0.477	2.408	0.426	2.520	0.377	2.633	0.331	2.746	0.287	2.858	0.246	2.969	0.209	3.078	0.174	3.184
34	0.610	2.160	0.556	2.266	0.503	2.373	0.452	2.481	0.404	2.590	0.357	2.699	0.313	2.808	0.272	2.915	0.233	3.022	0.197	3.126
35	0.634	2.136	0.581	2.237	0.529	2.340	0.478	2.444	0.430	2.550	0.383	2.655	0.339	2.761	0.297	2.865	0.257	2.969	0.221	3.071
36	0.658	2.113	0.605	2.210	0.554	2.310	0.504	2.410	0.455	2.512	0.409	2.614	0.364	2.717	0.322	2.818	0.282	2.919	0.244	3.019
37	0.680	2.092	0.628	2.186	0.578	2.282	0.528	2.379	0.480	2.477	0.434	2.576	0.389	2.675	0.347	2.774	0.306	2.872	0.268	2.969
38	0.702	2.073	0.651	2.164	0.601	2.256	0.552	2.350	0.504	2.445	0.458	2.540	0.414	2.637	0.371	2.733	0.330	2.828	0.291	2.923
39	0.723	2.055	0.673	2.143	0.623	2.232	0.575	2.323	0.528	2.414	0.482	2.507	0.438	2.600	0.395	2.694	0.354	2.787	0.315	2.879
40	0.744	2.039	0.694	2.123	0.645	2.210	0.597	2.297	0.551	2.386	0.505	2.476	0.461	2.566	0.418	2.657	0.377	2.748	0.338	2.838
45	0.835	1.972	0.790	2.044	0.744	2.118	0.700	2.193	0.655	2.269	0.612	2.346	0.570	2.424	0.528	2.503	0.488	2.582	0.448	2.661
50	0.913	1.925	0.871	1.987	0.829	2.051	0.787	2.116	0.746	2.182	0.705	2.250	0.665	2.318	0.625	2.387	0.586	2.456	0.548	2.526
55	0.979	1.891	0.940	1.945	0.902	2.002	0.863	2.059	0.825	2.117	0.786	2.176	0.748	2.237	0.711	2.298	0.674	2.359	0.637	2.421
60	1.037	1.865	1.001	1.914	0.965	1.964	0.929	2.015	0.893	2.067	0.857	2.120	0.822	2.173	0.786	2.227	0.751	2.283	0.716	2.338
65	1.087	1.845	1.053	1.889	1.020	1.934	0.986	1.980	0.953	2.027	0.919	2.075	0.886	2.123	0.852	2.172	0.819	2.221	0.786	2.272
70	1.131	1.831	1.099	1.870	1.068	1.911	1.037	1.953	1.005	1.995	0.974	2.038	0.943	2.082	0.911	2.127	0.880	2.172	0.849	2.217
75	1.170	1.819	1.141	1.856	1.111	1.893	1.082	1.931	1.052	1.970	1.023	2.009	0.993	2.049	0.964	2.090	0.934	2.131	0.905	2.172
80	1.205	1.810	1.177	1.844	1.150	1.878	1.122	1.913	1.094	1.949	1.066	1.984	1.039	2.022	1.011	2.059	0.983	2.097	0.955	2.135
85	1.236	1.803	1.210	1.834	1.184	1.866	1.158	1.898	1.132	1.931	1.106	1.965	1.080	1.999	1.053	2.033	1.027	2.068	1.000	2.104
90	1.264	1.798	1.240	1.827	1.215	1.856	1.191	1.886	1.166	1.917	1.141	1.948	1.116	1.979	1.091	2.012	1.066	2.044	1.041	2.077
95	1.290	1.793	1.267	1.821	1.244	1.848	1.221	1.876	1.197	1.905	1.174	1.934	1.150	1.963	1.126	1.993	1.102	2.023	1.079	2.054
100	1.314	1.790	1.292	1.816	1.270	1.841	1.248	1.868	1.225	1.895	1.203	1.922	1.181	1.949	1.158	1.977	1.136	2.006	1.113	2.034
150	1.473	1.783	1.458	1.799	1.444	1.814	1.429	1.830	1.414	1.847	1.400	1.863	1.385	1.880	1.370	1.897	1.355	1.913	1.340	1.931
200	1.561	1.791	1.550	1.801	1.539	1.813	1.528	1.824	1.518	1.836	1.507	1.847	1.495	1.860	1.484	1.871	1.474	1.883	1.462	1.896

Note: *n* = number of observations

k' = number of explanatory variables excluding the constant term.



ARBAMINCH UNIVERSITY
COLLEGE OF BUSINESS AND ECONOMICS
DEPARTMENT OF ECONOMICS
ASSIGNMENT ON ECONOMETRICS II

Maximum marks: 30%

Course Code: Econ 421

Credit hours: 3

General Instructions

- Make sure that your assignment paper contains **Two** parts and **Three** pages including the cover page.
- Answer all the questions on the separate answer sheet
- Credit will be given for neat and appropriate presentation
- Be tangent to the point

NAME-----
ID.NO-----
TERM-----
CENTER-----

Part I: Discussion Questions

1. Why the assumption of constant variance of the random variable does not hold in many econometric applications? (2 pts)
2. Discuss the consequences of autocorrelation (3 pts).
3. Differentiate distributed lag and autoregressive models. Give examples for each (2 pts).
4. Discuss the consequences of simultaneity bias (3 pts).
5. Why recursive systems are also called triangular systems (2 pts).

Part II: Work-out Questions

1. A financial analyst is interested in the relationship between dividend per share (DPS) and earnings per share (EPS). He collects information on these two variables and estimates the regression

$$DPS_i = \quad + \quad EPS_i + e_i$$

From this regression he computes the error from the regression for each company.

Company	EPS	e
1	1.00	.05
2	2.10	.35
3	1.20	.10
4	3.00	.72
5	1.25	.25
6	5.20	1.21
7	8.00	2.12

Does heteroscedasticity appear to be a problem in this regression? (5 pts)

2. Given the regression result: $Y = 35.25 + 5.45D + 2.25X$

$$t \quad (23.65) \quad (8.64) \quad (9.33)$$

Where, Y is income in thousands, D is a dummy variable taking on the value of 1 if male 0 for female and X = level of education.

- a. Calculate the mean salaries of male and female assuming both are illiterate (**2 pts**).
 - b. Test the significance of coefficient of educational level and Calculate the salary addition for male or female employee after one additional year of schooling (**2 pts**).
 - c. Holding educational achievement constant, check whether the mean salaries of male and female accountants are different or not (**2 pts**).
3. You are given the following error terms, which are the result of a regression of sales versus advertising expenditures for the Huessey Corporation over a 20-year period.

Year	e_t	Year	e_t
1	.075	11	.065
2	.080	12	.180
3	-.100	13	-.120
4	-.070	14	-.070
5	.500	15	.050
6	-.230	16	-.230
7	-.007	17	-.107
8	.088	18	.288
9	-.101	19	-.131
10	-.007	20	-.007

- a. Compute the Durbin-Watson statistic (**5 pts**).
- b. Does autocorrelation exist? (**2 pts**)

GOOD LUCK!!!