APPLIED ECONOMETRICS With Eviews Applications

Book · November 2013

CITATIONS
2

READS
33,977

2 authors, including:

Ali Göksu
University of Eurasia
30 PUBLICATIONS **79** CITATIONS

SEE PROFILE

APPLIED ECONOMETRICS

With Eviews Applications

IBU Publications

APPLIED ECONOMETRICS

With Eviews Applications

Authors:
Uğur ERGÜN
ugur.ergun@ibu.edu.ba

Ali GÖKSU
ali.goksu@ibu.edu.ba

Publisher:
International Burch University

Editor in Chief:
Prof.Dr. Mehmet Uzunoğlu

Reviewed by:
Assoc. Prof. Dr. Ercan BALDEMİR, Muğla Sıtkı Koçman University
Assoc. Prof. Dr. Safet KOZAREVIĆ, Faculty of Economics, University of Tuzla

DTP & Design:
Nihad Obralić

DTP and Prepress:
International Burch University

Printed by:

Circulation: 500 copies
Place of Publication: Sarajevo

International Burch University Publication No: 23

CIP - Katalogizacija u publikaciji
Nacionalna i univerzitetska biblioteka
Bosne i Hercegovine, Sarajevo

330.43(075.8)

ERGUN, Uğur
Applied econometrics : with eviews applications
/ Uğur Ergun, Ali Göksu. - Sarajevo :
International Burch University, 2013. - 272 str. :
graf. prikazi ; 24 cm. - (IBU Publications)

Bibliografija: str. 267-271.

ISBN 978-9958-834-29-5
1. Göksu, Ali
COBISS.BH-ID 20792838

Uğur ERGUN **Ali GÖKSU**

APPLIED ECONOMETRICS

With Eviews Applications

IBU Publications
Sarajevo, 2013

PREFACE

This book is designed as auxiliary source for the students who are taking Applied Econometrics course. It is intended to clarify basic econometrics methods with examples especially for Finance students.

In this book, two main dimensions have been configured. In the first chapters some basic information regarding scientific research is given in order to polish student's ability to understand following chapters better. Following chapters are organized to provide more detailed information about some specific methods and their applications in finance by examples.

15th November 2013

Uğur ERGÜN Ali GÖKSU

LIST OF ABBREVIATIONS

2SLS : Two-Stage Least Squares
ACF : Autocorrelation Function
ADF : Augmented Dickey-Fuller
AIC : Akaike Information Criterion
ANOVA : Analysis of Variance
AR : Autoregressive
ARDL : Autoregressive Distributed Lag
ARIMA : Autoregressive Integrated Moving Average
ARMA : Autoregressive Moving Average
ARCH : Autoregressive Conditional Heteroscedasticity
BLUE : Best Linear Unbiased Estimator
BUE : Best Unbiased Estimator
CAPM : Capital Asset Pricing Model
CLRM : Classical Linear Regression Model
CUSUM : Cumulative Sum of Control Chart
DF : Dickey-Fuller
DW : Durbin Watson
ECM : Error Correction Model
EGARCH : Exponential GARCH
EGLS : Estimated Generalized Least Squares
ESS : Explained Sum of Squares
EU : European Union
FDI : Foreign Direct Investment
FDL : Distributed Lag Model
GARCH : Generalozed Autoregressive Conditional Heteroscedasticty

GDP : Gross Domestic Product
GLS : Generalized Least Squares
GMM : Generalized Method of Moment
GQ : Goldfeld-Quant
IS-LM : Investment Saving–Liquidity Preference Money Supply
ISE : Istanbul Stock Exchange
IV : Instrumental Variables
LM : Langrage Multiplier
MA : Moving Average
NCLRM : Normal Classical Linear Regression Model
OLS : Ordinary least Squares
PRF : Population Regression Function
P-P : Probability-Probablity
PP : Philip-Peron
QQ : Quantile-Quantile
RESET : Regression Specification Errror Test
RSS : Residual Sum of Squares
SEM : Structural Equation Modelling
SER : Standard Errror of the Regression
SIC : Schwarz Information Criterion
SSR : Sum of Squared Residuals
TARCH : Threshold ARCH
TSS : Total Sum of Squares
VAR : Vector Autoregressive
VECM : Vector Error Correction Model
VIF : Variance Inflation Factor
WLS : Weighted Least Squares

TABLE OF CONTENTS

CHAPTER 2

MODELS

CHAPTER 3

DATA

CHAPTER 4

REGRESSION

CHAPTER 5

CLASSICAL LINEAR REGRESSION MODEL (CLRM)

CHAPTER 6

HYPOTHESIS TESTING

CHAPTER 7

SPECIFICATIONS

CHAPTER 8

PARAMETER STABILITY

CHAPTER 9

INTERPRETING THE LINEAR REGRESSION MODELS

CHAPTER 10

ENDOGENEITY

CHAPTER 11

REGRESSIONS WITH DUMMY VARIABLES

CHAPTER 12

TIME SERIES

CHAPTER 14

GRANGER CAUSALITY

CHAPTER 15

VECTOR AUTOREGRESSION MODEL (VAR)

CHAPTER 1

SCIENTIFIC RESEARCH

DEFINITION

Scientific research is a systematic, controlled, empirical, amoral, public and critical investigation of natural phenomena guided by theory and hypotheses about the presumed relations among such phenomena[1].

AIMS OF THE SCIENTIFIC RESEARCH

The main objective of a research is to find out answers to questions through the application of scientific procedures. Furthermore, each research has its own specific purpose.

Research objectives may include but are not limited to following groupings:

a) To review and synthesize available knowledge
b) To test a hypothesis of a causal relationship between variables
c) To investigate some existing situation or problem

[1] Lee, H.B & Kerlinger, F.N(2000)

d) To explore possible solutions to a specific issue
e) To analyze general issues
f) To offer a new procedure
g) To explain a new scientific idea
h) To generate new knowledge, etc.

RESEARCH METHODS & METHODOLOGY

Research Methods

Research methods include all the methods or techniques that are used for performing a research process. Actually, research method is a part of the research methodology that helps in constructing it.

Research methods can be classified as follows:

- Methods related to the data collection;
- Statistical techniques used to analyze research questions;
- Methods used to evaluate the accuracy of the results obtained through statistical techniques.

Research Methodology

Research methodology is a way of solving systematically the research problem. It is actually a science on how a research is done scientifically. Researchers also need to comprehend related theories and to understand the assumptions underlying various techniques which are applicable for the research focus and questions. Certain techniques and procedures are applicable only to specified problems. Thus, a researcher should design his/her research based on the problem which he/she tries to solve by applying appropriate statistical techniques.

A research methodology should answer the following questions;

- Why a research study has been undertaken? What is the motivation of the research?

- What is the problem statement? How is it defined?
- What data have been used and how have the data been collected?
- What statistical or econometric method has been conducted?
- What is the contribution of the research?
- What is the significance of the research?
- What are the implications?

The Scientific Method

The scientific method relies on empirical evidence, utilizes relevant concepts, is committed to only objective considerations, presupposes ethical neutrality, aims at nothing but making only adequate and correct statements about population objects and results into probabilistic predictions. Its methodology is available to all concerned for critical scrutiny and for use in testing the conclusions through replication, aiming at formulating more general axioms or what can be termed as scientific theories.

CRITERIA OF A GOOD RESEARCH

One expects scientific research to satisfy the following criteria[2]:

- The purpose of the research should be clearly defined and common concepts should be used.
- The research procedure used should be described in sufficient detail to permit another researcher to repeat the research for further advancement, keeping the continuity of what has already been attained.
- The procedural design of the research should be carefully planned to yield results that are as the objective as possible.
- The researcher should report with complete frankness, explain all flaws in the procedural design and estimate their effects on the findings.

[2] Fox, J.H.(1958)

- The analysis of data should be sufficiently adequate to reveal its significance and the methods of analysis used should be appropriate. The validity and reliability of the data should be checked carefully.
- Conclusions should be confined to those parts justified by the data of the research and limited to those for which the data provide an adequate basis.
- Greater confidence in research is warranted if the researcher is experienced, has a good reputation in doing research and is a person of integrity.

In a more systematic way, the qualities of a good research can be stated as follows[3]:

i- ***Systematic***: It means that a research is structured through following specified steps that are to be taken in a specified sequence in accordance with the well-defined set of rules. Systematic characteristic of the research does not rule out creative thinking, but it certainly does reject the use of guessing and intuition in arriving at conclusions.

ii- ***Logical***: This implies that research is guided by the rules of logical reasoning and the logical process of induction and deduction are of great value in carrying out a research. Induction is the process of reasoning from a part to the whole whereas deduction is the process of reasoning from some premise to a conclusion which follows from that very premise. In fact, logical reasoning makes research more meaningful in the context of decision making.

iii- ***Empirical***: A research is related basically to one or more aspects of a real situation and deals with concrete data that provides a basis for external validity to the research results.

[3] Bellenger D.N & A. Greenberg, B.A. (1978)

iv- ***Replicable***: This characteristic allows research results to be verified by replicating the study and thereby building a sound basis for making decisions.

POSSIBLE USERS OF RESEARCH OUTCOME

The researches are expected to produce usefull implications for the policy makers, managers, investors, advisors or academicans as follows;

- The Government – for making or/and developing their policies;
- Managers, Investors, Advisors – use for profit maximization;
- Academics – use for conducting further studies;

RESEARCH TYPES[4]

Based on the Research Purpose

Descriptive Research

Descriptive Research includes surveys and fact-finding enquiries of different kinds. The major purpose is generally to describe an event, phenomenon, or a relationship. The term Ex post facto research is used instead of descriptive research in social sciences. The researcher only report what has happened or what is happening and has no control over the variables. The researcher mostly seeks to measure such items as frequency of shopping, preferences of people, or similar variables in this research type. This type of research also includes attempts by researchers to discover causes even if they cannot control the variables. The methods of research utilized in descriptive research are all survey methods, including comparative and correlational methods.

[4] Kothari, C.R. & Garg, G.D. (2004)

Descriptive research is conducted to describe facts as they exist. It is used to identify or obtain information for a specific issue. Descriptive research takes further steps in exploring an issue than exploratory research which is employed to clarify the structural issues. Therefore, even in a descriptive study, you have to spend time in refining your research questions and be specific about the phenomena you are studying. In short, descriptive research may be defined as statistical summarization of the collected data.

Example 1.1

- What are the feelings of Academics faced with redundancy?
- What kind of ethical behaviors do students expect at the universities?

Exploratory Research[5]

The objective of an exploratory research is the development of hypotheses rather than their testing, whereas formalized research studies are those with substantial structure and specific hypotheses to be tested. If there are very scant studies which are used to clarify existing information, exploratory research is conducted. This kind of research is also employed to explore patterns, scientific ideas or hypotheses. In exploratory research, the main idea is to obtain close insights and knowledge regarding the issue which is studied.

The exploratory research techniques include case studies, observations and historical analysis. Quantitative and qualitative data can be used. The researcher assesses the consistency of the results which are obtained empirically with the existing theories and concepts or develop them.

Exploratory research provides informative additional for further research rather than giving answers to problems or issues. Exploratory or qualitative research is used to obtain deep insights

[5] Collis, J & Hussey, R. (2009).

into the behavior of few consumers, or to gain preliminary information about the market.

Common exploratory research methods include depth interviews, projective techniques, and focus groups. Researchers employ exploratory research when little is known about the topic and previous theories or ideas do not apply. For example, if you wanted to study how to get students to use the computer lab in a college environment, you might first have to do exploratory research to figure out which students might need the lab and what appeals to this demographic. Exploratory research clarifies problems, gathers data and creates initial hypothesis and theories about subjects. The primary point of exploratory research is to give researchers pertinent information and help them to form initial hypotheses about the subject.

Both exploratory and descriptive research can be conducted in the same research, but exploratory should be conducted before descriptive research in order to figure out data and hypothesis. Then, in the second part of the research, defined hypotheses are tested.

Exploratory studies are predominantly used:

i- To satisfy the researcher's curiosity and desire for better understanding of a particular topic;
ii- To determine the feasibility of undertaking a more careful study;
iii- To develop research techniques and suggest direction for future research.

Example 1.2

- Why do some cities have higher traffic accident rate than others?
- Why are some children becoming piddling?

Analytical Explanatory Research[6]

Analytical or explanatory research is an extension of the descriptive research. The researcher not only describes the issue, but also analyses and explains the reasons why or how the issue being studied happened. Analytical research goes further by examining and measuring causal relationship among the factors in the phenemenon.

The important elements of Analytical explanatory research are;

- Identification
- Controlling
- Exploring causal links among factors

A variable is a factor or characteristic of an issue which are observable and measurable.

Example 1.3

- Examining the Turkey balance of payment deficit movement during 2008 global financial crisis.
- Examining trends of real and nominal value of New Turkish Lira against Euro after 2000.

Predictive Research

Predictive research aims to generalize findings from the analysis conducted on the basis of hypothesized, general relationships, and provides explanations for interactions or interchanges in a specific and general situations.

If the predictive research provides robust, consistent and clear explanations, its empirical finding can be applicable to the similar cases, helps to understand more complex phenomenons. Predictive research provides 'how', 'why' and 'where' answers to current events and also to similar events that may occur in the future. It is also helpful in situations where 'what if' questions are being asked.

[6] Collis, J & Hussey, R. (2009)

Example 1.4

- What type of marketing strategy can improve the sales?
- How would an increase in interest rates affect inflation rate?
- During financial crises, what would happen to sales of personnel computer?

Based on the Research Process[7]

Looking at the approach adopted by the researcher we can also differentiate among researches. Some researchers prefer to take a quantitative approach in addressing their research question(s) and design a study that involves collecting quantitative data (and/or qualitative data that can be quantified) and analyzing them using statistical methods. Others prefer to take a qualitative approach to address their research question(s) and design a study that involves collecting only qualitative data and analyzing it using interpretative methods. A comprehensive study should incorporate elements of both.

Qualitative and quantitavie research can be used together in the same research such as;

- Collect data through Qualitative techniques and quantify them by counting the frequency of occurrence of specific key words or themes in order to employ statistical methods,
- Collect qualitative data and analyse them by non-numerical methods, or
- Collect numerical data and use statistical methods to analyze them.

Some researchers avoid taking a quantitative approach because they are not confident with statistical methods and think that a qualitative approach is easier. Many researchers consider that it is pretty hard to start and decide about overall design for a quantitative study, but it is easier to conduct the analysis and write up the findings since it is well structured. Qualitative research is realtively more difficult in the stage of analyzing data and interpreting it although it is easier to start.

[7] Collis, J & Hussey, R. (2009).

For instance, quantitative data such as absenteeism rates or productivity levels may be collected in order to analyze the the impact of night shifts on stress rates. Alternatively, same issue may be explored by collecting qualitative data consist of night workers perceptions and analyze it by statistical methods.
In fact, the research scope, philosophical preferences of the researcher and the data vailabilty are the main factors in defining the type of research.

Qualitative Research[8]

The qualitative research is about exploring issues, understanding facts, and investigating the why and how of the decision making. It uses induction scientific method and non-statistical techniques for sample of focus rather than large samples. Qualitative research aims to gather an in-depth understanding of human behavior and the reasons that govern such behavior. The subjective data which is gathered through unstructured or semi-structured techniques (in depth interviews or group discussions) includes the perception and interpretation of people. It produces comprehensive information, and is exploratory or investigative. Findings are not conclusive and cannot be used to make general conclusions about the population of interest. It is necessary, to develop an initial understanding and sound base for further decision making.

This type of research aims at discovering the underlying motives and desires used in depth interviews for this purpose. Other techniques used in such research are word association tests, sentence completion tests, story completion tests and similar projective techniques. Attitude or opinion research i.e., research designed to find out how people feel or what they think about a particular subject or institution is also a type of qualitative research. Qualitative research is especially important in the behavioral sciences where the aim is to discover the underlying motives of particular behavior. Through such research we can analyze the various factors which motivate people to behave in a particular manner or which make

[8] Kothari, C.R. (2004).

people like or dislike something. It may be concluded that, to apply a qualitative research method in practice is relatively huge challenge and therefore, while doing such research, one should seek guidance from experienced psychologists.

Example 1.5

- What is the role of the experience in participation on public health education sessions for smokers and drug users?
- How should the nurses handle patients who refuse to follow instructions?

Quantitative Research

Quantitative research focuses on gathering numerical data and making general conclusions from a group of people. Its objective is to test the hypothesis, investigate cause, effect, relationship and to make sound predictions. Mostly, research samples are consisted of numbers and statistics selected from population randomly. Research questions are determined clearly in the beginning of the research. Confirmatory deduction scientific research method is used to test specific hypothesis and theories with randomly collected data. Conclusion is made based on the statistical finding. In data collection, structured techniques such as online questionnaires, on-street or telephone interviews are used mainly.

Example 1.6

- To examine the interaction between inflation and balance of payment.
- To examine the relationship between usage and need of nursing services in the rural areas compared to urban areas.

Based on the Research Outcome

Applied Research

Applied research is a study that has been conducted in order to apply its findings to solve a specific, existing problem. It is the application of

available knowledge to improve management practices and policies. The research project is likely to be short term (often less than 6 months) and the immediacy of the problem is more important than amply academic theorizing. For example, you might be investigating the reorganization of an office layout, the improvement of safety in the workplace or the reduction of wasting raw materials or energy in a production process. The output from this type of research is likely to be a consultant's report, articles in professional or trade magazines and presentations to practitioners. Another type of applied research that is conducted in academic institutions often goes under the general title of educational scholarship (instructional research or pedagogic research). This type of study is concerned with improving the educational activities within the institution and the output is likely to be case studies, instructional software or textbooks. It focuses on finding an immediate solution to an existing problem.

Example 1.7

- The researcher collects information to test the effectiveness of traffic policies and fines in minimizing traffic causalities.

Basic Research

When the research problem is of a less specific nature and the research is being conducted primarily to improve our understanding of some general issues without emphasis on its immediate application, it is classified as a basic or pure research. For example, you might be interested in whether personal characteristics influence people's professional choices (career, etc.). Basic research is considered as the most academic type of research, since the primary aim is to make a contribution to the existing knowledge, usually for the common good, rather than to solve a specific problem for one organization. Basic research focuses on problem solving theoretically, and provides contribution to the existing scientific knowledge which can be used to find solutions for future problems.

Basic research is conducted;

- to provide new information,
- to make contribution to the knowledge,
- to improve understanding of new phenemenon, and
- to formulate a theory.

Example 1.8

- The researcher collects information to understand further the relationship between socioeconomic status and the intention to obey to traffic rules.

Based on the Research Logic

Deductive Research

Theoretical framework is developed and tested through empirical observations and particular instances are deducted from general inferences. For this reason, the deductive method is referred to as moving from the general to the particular. Deductive research starts with the general and proceeds to the specific and the steps are as follows: develop a model, form a hypothesis, gather data to test the hypothesis, and in the end use the data to conclude whether or not the model describes reality accurate.

Example 1.9

- To test the theories of motivation in different cultures and sectors.

Inductive Research

Inductive research represents type of study in which general inferences are induced from particular instances. Individual observations are used make general statements and it is referred to as moving from the specific to the general. Inductive research starts with the specific and proceeds to the general, thus this type of research starts with data collection, followed by examining the data for patterns, forming a hypothesis and then constructing a theory.

Example 1.10

- To investigate the relationship between production level and shift time in a factory in order to test the theory stating that production levels vary with working time length.

A particular research may be applied study, analytical study with quantitative approach or both qualitative and quantitative approaches together It can employ deductive or inductive methods, exploratory or descriptive research and analytical or predictive research.

CONCEPTUAL RESEARCH

Conceptual research is related to some abstract idea(s) or theory. It is generally used by philosophers and thinkers to develop new concepts or to reinterpret existing ones.

EMPIRICAL RESEARCH

Empirical research relies on experience or observation alone, often without seeking for system and theory. It is data-based research, resulting in conclusions which are capable of being verified by observation or experiment, thus it can be named as experimental type of research.

The researcher starts with a working hypothesis or guesses the probable results. Then he collects the data and sets up experimental designs which according to him will manipulate the persons or the materials concerned so as to bring forth the desired information. Evidence gathered through experiments or empirical studies is today considered to be the most powerful support possible for a given hypothesis.

HISTORICAL RESEARCH[9]

Historical research utilizes historical sources such as documents, remains, etc. in order to study events or ideas from the past, including the philosophes of persons and groups at any remote point of time. The purpose is to make people aware of what has happened in the past offering the opportunity to learn from past failures and successes, and apply them to present problems, make predictions based on that and test hypothesis concerning relationships or trends.

Example 1.11

- To investigate the family trees
- To investigate the role of Ottoman Empire in the First World War 1

[9] Kumar, R. (2008).

CHAPTER 2

MODELS

ECONOMIC MODEL

Economic modeling is one of the most important parts of a research and the economic theory generally. The researcher organizes and designs his/her ideas through economic modeling. The model helps the researcher to rationally and logically support the hypothesis or answer research questions defined in the beginning of the research. The researcher follows certain steps in this model and delivers scientific findings in the end. As a result, he/she produces logical and beneficial results for the society in general.

Example 2.1

- Aggregate Supply – Aggregate Demand (AS/AD) Model,
- Loanable Funds Model
- HMC Macro Sim- simulation model,
- IS/LM Model

TYPES OF ECONOMIC MODELS

Visual Models

Visual models involve graphical representation of different economic issues. It consists of graphs with lines and curves which provides information and ideas in brief about an issue of interest. They are generally used in textbooks for teaching in order to clarify the thoughts or theories. The models help to present the complex relationships between economic variables.

Example 2.2

Supply-and-Demand model
General Equilibrium model

Mathematical Models

The mathematical models are systems of simultaneous equations with an equal or greater number of economic variables. They consist of more complex variables and relationships. These models require a good previous knowledge of algebra or calculus for its implementation.

For example, a basic microeconomics model includes a supply function (shows the behavior of producers), a demand curve (shows the behavior of purchasers) and an equilibrium equation. The variables in this model represent a type of economic activity or information that either determines or is determined by that activity and are classified as endogenous or exogenous.

Example 2.3

To examine the demand elasticity for luxury cars in low income countries

Empirical Models

Empirical models are one type of mathematical models designed to be used along with data. The basic model is mathematical and the

data is gathered for the variables of interest; afterward statistical or econometric techniques are applied to estimate the values of the variables which are used previously in the model.

Example 2.4
To investigate the changes in income when investment changes one percent

Simulation models are mainly used and created by using different computer software. The basic features of mathematics are required. The mathematical complexity varies depending on the research scope.

Example 2.5
The equations of the model are programmed in a software programming language

Static Model

Most models which are used in economics are comparative statics models. These models provide information about what happens over time. The model estimate generally starts with predefined equilibrium condition, and then a "shock" to the model is applied (changing the value of one or more of the variables). In the end, the new equilibrium is obtained without an exposition of what happened in the transition from first equilibrium to the second.

Example 2.6
To investigate the impact of class size on the average test score of the students

Dynamic Models

Dynamic models directly incorporate time into the model. This is usually done in economic modeling by using differential equations. Sometimes dynamic models better represent the subtleties of

business cycles, because time lags in behavioral response and timing strongly influence the character of one cycle.

Example 2.7

To examine the role of Interest rate on Inflation rate movements over time

WHICH THEORY IS APPROPRIATE?

While describing the economic activity of consumers in general or a certain type of business decisions the theory of adaptive expectations is probably most suiTable. A rational expectations theory is the most appropriate one when describing an economic activity or en economic environment where the decision-makers are likely to be sophisticated and well-informed. The behavior and actions of traders in the finance markets provide a good example for rational expectations. Portfolio investors, traders of financial assets or company shareholders observe the market and seek for new economic information virtually whole the working day. The market forces adjust themself instantly to the macroeconomic indicators, expecting the impact of such policy on interest rates and consequently share prices. Hence, the choice of the theory of expectations depends on the context. Models incorporating and related to both theories would be developed and used.

ECONOMETRIC MODELS

Econometrics is a science and art of using economic theory and statistical techniques to analyze economic data.[10] An econometric model should be constructed based on economic theory, experience or critical thinking. Econometrics finds and explains the relationship between economic variables based on outcomes of statistical techniques using the available data. It clarifies the interaction and mediation between data, economic theory and statistical techniques.

[10] Stock, J.H & Watson, M.W. (2008).

Econometrics aims to explore relationship between economic variables and interpret the results obtained through statistical techniques.

Econometrics models can be classified as follows:

i- The models developed to find out relationships between past and present.

Example 2.8
The impact of previous days' stock returns on the current stock return

ii- The models that examine relationship between economic variables over time

Example 2.9
To examine the degree of relationship between export and inflation over time

iii- The models which investigate the relationship between different variables measured at a given point of time.

Example 2.10
To investigate the impact of student-teacher ratio on student exam results

iv- The models which consider the relationship between different variables for different units over time

Example 2.11
To examine the relations between inflation and unemployment in different countries

Attributes of a Good Econometric Model

Features of the good econometric model may be summarized as follows;

- Parsimonious, explain a lot with minimum specification, optimum sized model;
- Identifying- unique values exist for the parameter.
- Goodness of fit- it shows the power of the model in explaining changes in regress.
- Theoretical Consistency- consistent with the related economic theories
- Predictive power.

Structural Equation Modeling (SEM)

Structural equations are the equations specific for the economic theory. There are different types of structural equations such as:

i- Behavioral equation, consumption equation

Y= C + I + X – M

ii- Technical relationships, production function

Q = f (L, K)

iii- Identities, Keynesian macro model

$$C_t = \beta_1 + \beta_2 Y_t$$

$$Y_t = C_t + I_t$$

The Reduced Form Model

It includes reduced form of equations. Keynesian macro model is a complete structural form with two equations for the two endogenous variables $C_t \; and \; Y_t$; I_t is supposed to be exogenous. Solving the model for $C_t \; and \; Y_t$ gives two reduced from equations:

$$C_t = \frac{\beta_1}{1-\beta_2} + \frac{\beta_2}{1-\beta_2} I_t$$

$$Y_t = \frac{\beta_1}{1-\beta_2} + \frac{1}{1-\beta_2} I_t$$

In different notation;

$$C_t = \pi_{11} + \pi_{12} I_t,$$
$$Y_t = \pi_{21} + \pi_{22} I_t$$

The parameters $\pi_{11}, \pi_{12}, \pi_{21}\ and\ \pi_{22}$ are called reduced form parameters. Reduced form equations indicate that the endogenous variables are correlated with the exogenous regressors. Thus, each equation can be estimated by 2SLS using all of the exogenous variables in the system.

In the reduced form of equations the endogenous variables are expressed in terms of the exogenous and lagged variables. A special reduced form model is the model with only exogenous explanatory variables. These types of models are called the classical regression model. In this model all of the explanatory variables should be exogenous.

Stochastic Models

Stochastic modeling is a technique of predicting outcomes and takes into account a certain degree of randomness, or unpredictability. Economic relationship is not an exact relationship; a disturbance or error term, u_t should be added to right hand side of the equation. It is also called stochastic term in the non systematic part of the regression equation.

$$y_t = \beta_1 + \beta_2 x_t + u_t$$

Deterministic Models

Deterministic model is a mathematical model in which outcomes are determined through known relationships among variables and events

without random variation. In deterministic models, given input always produces the same output, such as in a known chemical reaction. In comparison, stochastic models use ranges of values for variables in the form of probability distributions.

Analytical Framework

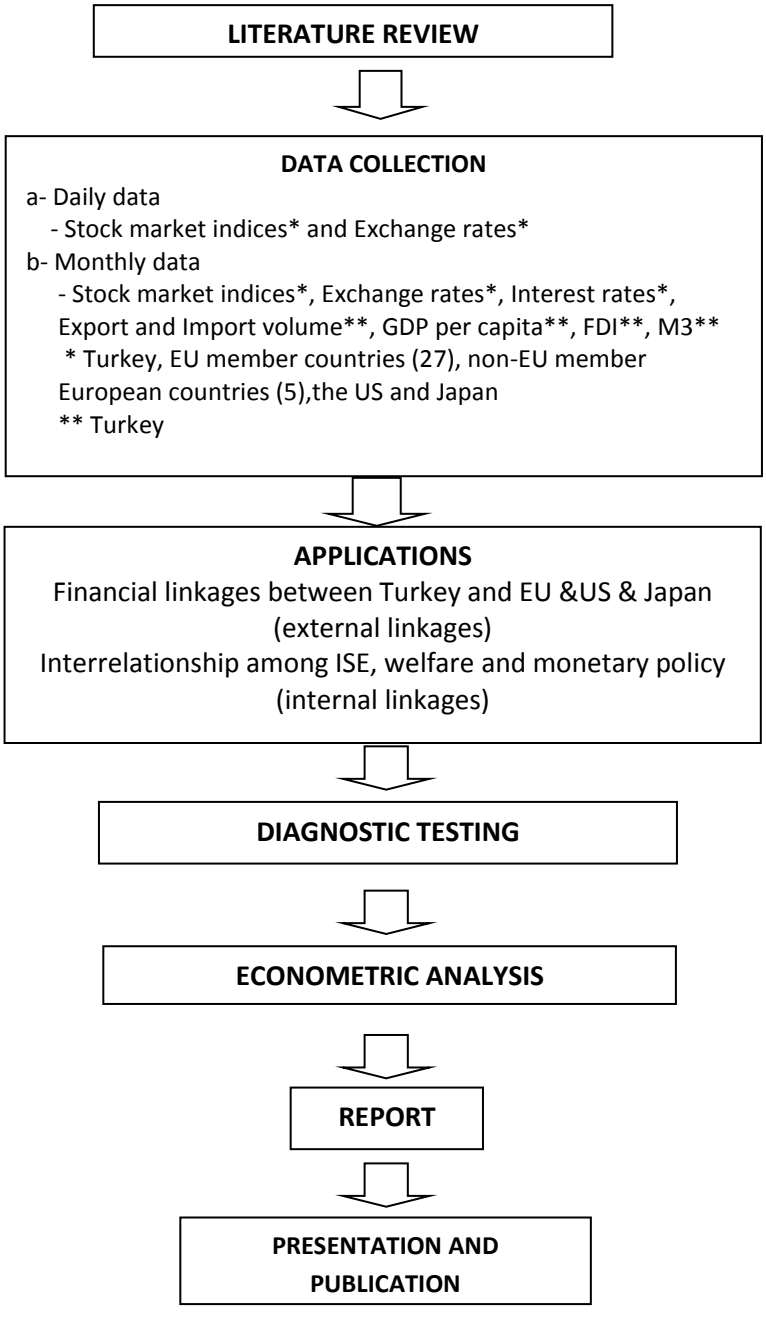


Figure 2.1 Analytic Framework

Methodological Framework

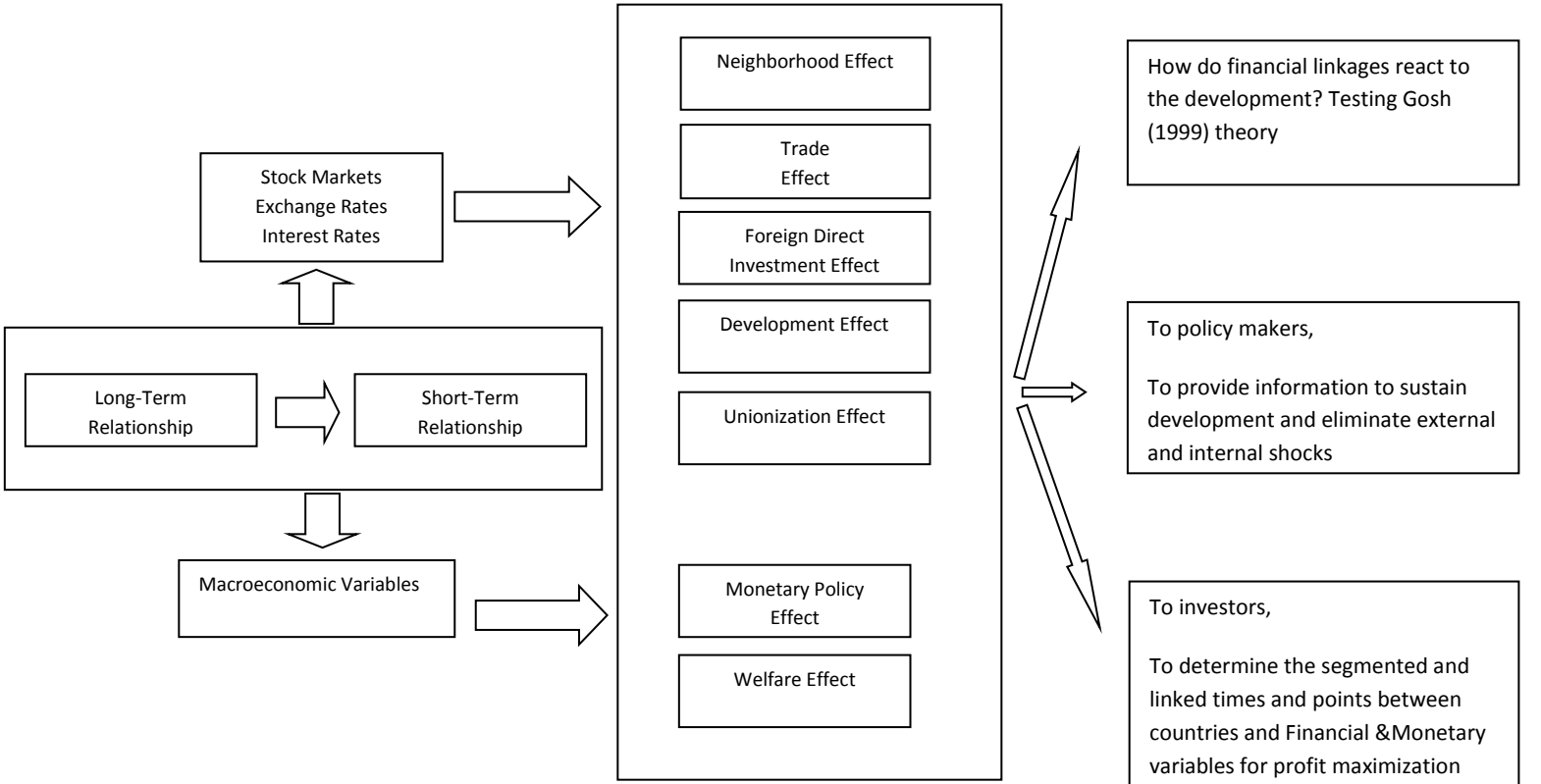


FIGURE 2.2 Methodological Framework

Research Process Flow Chart

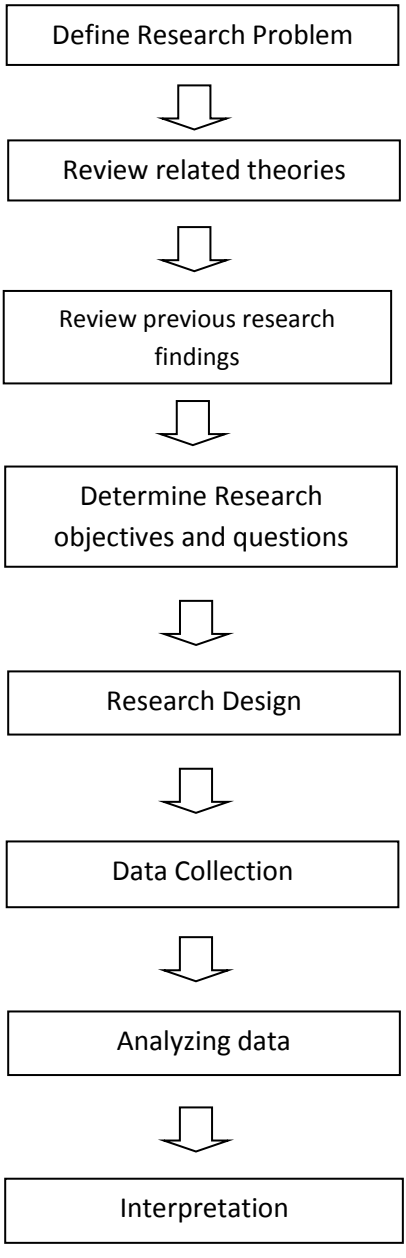
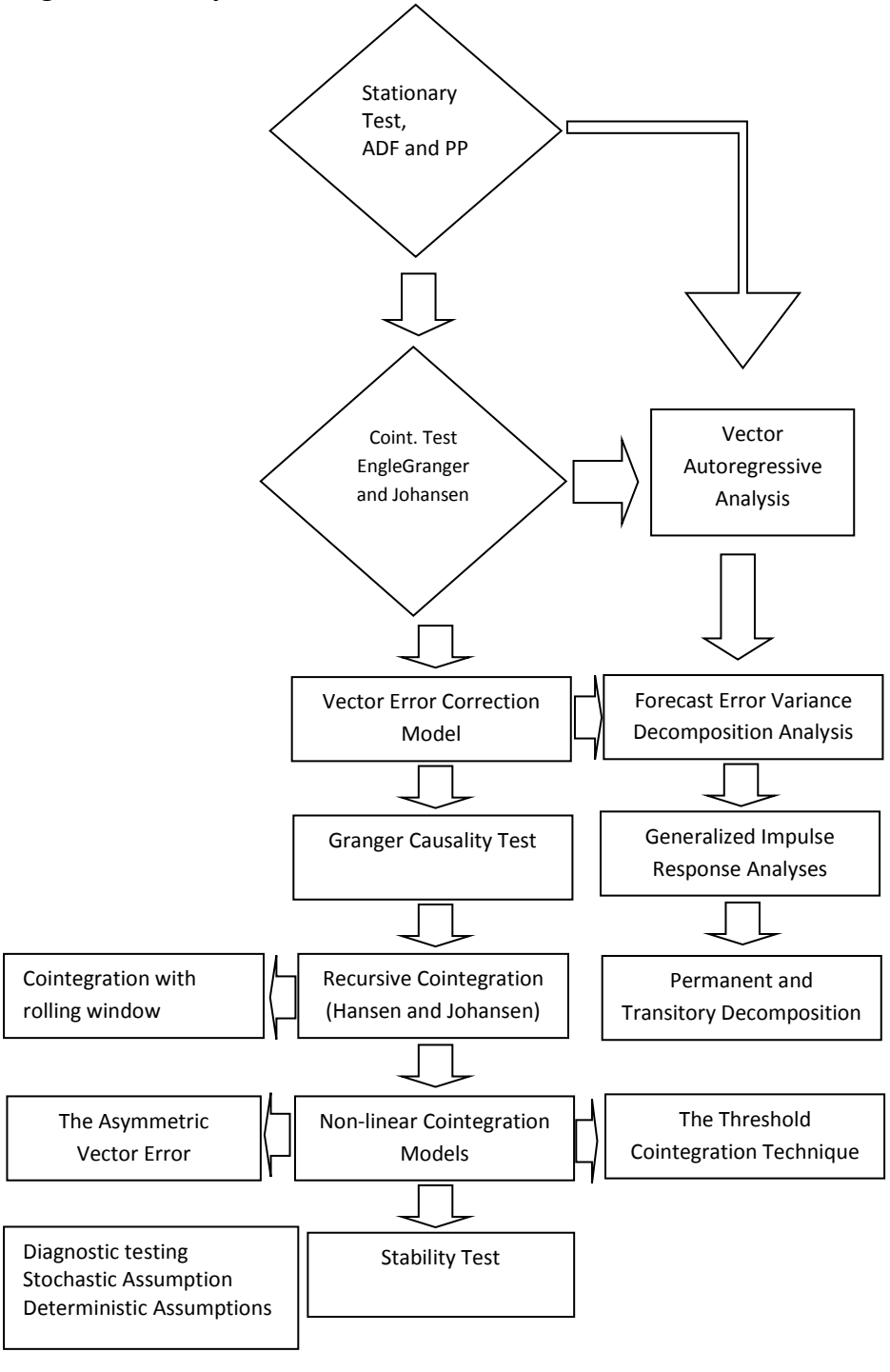


Figure 2.3 Research Process Flow Chart

Regression Analysis Flow Chart



CHAPTER 3

DATA

Data are the piece of information or knowlege which are used for reasoning or calculation as measurement.

QUALITATIVE DATA

Qualitative data is extremely varied in nature. It includes virtually any information that can be captured but is not numerical in nature. Here are some of the major categories or sources of qualitative data:

In-Depth Interviews

In-depth interviews include both individual interviews (one-on-one) as well as "group" interviews (including focus groups). The data can be recorded in a wide variety of ways including audio recording, video recording or written notes. In-depth interviews differ from direct observation primarily in the nature of the interaction. In interviews it is assumed that there is a questioner and one or more interviewees. The purpose of the interview is to probe the ideas of the interviewees about the fact of interest.

Example 3.1

- Which business leaders or role models do you respect?
- How do you handle conflict on a team project?
- How do you see the efficiency of current economic policies?
- What sectors should government support? Why?

Direct Observation

Direct observation differs from interviewing in that the observer does not actively examine the respondent. It includes field research to illustrate some aspect of the phenomenon. The data can be obtained through interviews (stenography, audio, and video) or pictures, photos or drawings (courtroom drawings of witnesses).

Example 3.2

- To analyze the student's reaction to the male and female teachers
- To analyze the impact of family income on their children's behavior in the public schools

Written Documents

Written documents refer to existing and available documents. It includes newspapers, magazines, books, websites, memos, transcripts of conversations, annual reports, and so on.

Example 3.3

To analyze the ethical sensitiveness in the novels published between two world wars

Qualitative Methods

Generally, the qualitative methods are limited by the imagination of the researcher. There are a wide variety of methods that are common in qualitative measurements. Some of them are mentioned below.

Participant Observation

Participant observation is one of the most common and demanding methods for qualitative data collection. The researcher should be a participant in the environment or context being observed. How to enter the context, the role of the researcher as a participant, the collection and storage of field notes, and the analysis of field data are discussed in the literature on participant observation. Participant observation is usually long term, involving several months or years of intensive work because the researcher is required to be natural part of the environment in order to assure that the observations are natural.

Direct Observation

Direct observation is different from participant observation in the following ways:

- A direct observer doesn't necessary become a participant in the environment. Instead, direct observer tries to be as retiring as possible so as not to spoil or bias the observations.
- Direct observation suggests a more detached perspective. The researcher is watching rather than participating. Technology can be a useful part of direct observation. For instance, one can videotape the phenomenon or observe from far or behind one-way mirrors.
- Third, direct observation tends to be more focused than participant observation. The researcher is observing certain sampled situations or people rather than trying to become immersed in the entire context.
- Finally, direct observation tends not to take as long as participant observation. For instance, one might observe child-mother interactions under specific circumstances in a laboratory setting from behind a one-way mirror, looking especially for the nonverbal cues being used.

Structured Interviewing

It is associated with quantitative research and the main goal is to collect data from large samples ensuring consistency of response. In

this interview, the questions are clear and fully understood by the respondents and, ach respondent answer same set of question in order to obtain an objective result minimizing interviwer bias.

The questions are selected based on research objectives. It includes detailed instructions and information regarding interview questions.

Unstructured Interviewing

Unstructured interviewing requires direct interaction between the researcher and a respondent or group. It differs from traditional structured interviewing in the following:

- There is no formal structured instrument, questions or protocol except some limited guiding questions. But, the researcher may have some initial guiding questions or core concepts to ask about.
- The researcher or interviewer can move the conversation in any direction of interest that may come up.
- Unstructured interviewing is particularly useful for exploring a topic broadly.
- Each interview is unique with no predefined questions to all respondents.
- It is more difficult to analyze unstructured interview data.

Case Studies[11]

A case study is an intensive study of a specific individual or specific entity. For instance, Freud developed case studies of several individuals as the basis for the theory of psychoanalysis and Piaget did case studies of children to study developmental phases. There is no single way to conduct a case study, and a combination of methods (e.g., unstructured interviewing, direct observation) can be used.

[11] http://www.socialresearchmethods.net

QUANTITATIVE VERSUS QUALITATIVE DATA

Quantitative methods focus only on numbers and frequencies rather than on meaning and experience; they are associated with the scientific and experimental approach without in depth description.

Qualitative methods are ways of collecting data which are concerned with describing, rather than drawing statistical inferences and provide a more in-depth and rich description analysis.

In modern research, most psychologists tend to adopt a combination of qualitative and quantitative approaches, which allow statistically reliable information obtained from numerical measurement to be backed up and enriched by the information gained from the research participants' explanations.

WHY ARE QUANTITATIVE AND QUALITATIVE DATA IMPORTANT?

Quantitative and qualitative method provide different outcomes, and are often used together to get a full picture of a population. For example, if data are collected on annual income (quantitative), occupation data (qualitative) could also be gathered to get more in detail the information on the average annual income for each type of occupation. Quantitative and qualitative data can be obtained from the same data unit depending on whether the variable of interest is numerical or categorical.

QUANTITATIVE DATA TYPES

Primary Data

Primary data are those that the researcher collects himself. The researcher is expected to have accessed primary data while using

quantitative methods, and primary data collection is necessary when a researcher cannot find the data needed in secondary sources.

Secondary Data

A secondary data research uses existing data. It can be obtained from very wide range of sources such as: literature, industry surveys, reports, databases and information systems.

Experimental Data[12]

Experimental data is obtained through experiments in order to examine economic policy or treatment. It is much more expensive to conduct compared to observational data. It also has administration difficulties and ethical consideration.

In sciences, experimental data is data produced by a measurement, test method, experimental design or quasi-experimental design. In clinical research any data produced as a result of clinical trial is experimental data. It can be qualitative or quantitative, each being appropriate for different investigations.

Observational data

Most of the data in the economic research is obtained through observations. It includes surveys (telephone surveys, on street surveys, mail surveys, and interviews), company records, government records, past financial data records, etc. In economics, the researchers usually do analysis with observational data in different circumstances which are not under the control of researchers.

Researchers who use observational data can obtain data from lab scientists, make reliable conclusions and recognize the limitations of data that are gathered in more natural settings.

[12] http://en.wikipedia.org/wiki/Experimental_data

Time Series[13]

A time series is a collection of observations of variables obtained through repeated measurements over time. A times series allows the researcher to identify presumed changes within a population over time. It can also show the impact of cyclical, seasonal and irregular events on the data item being measured. Time series can be classified into two different types: stock and flow.

A stock series is a measure of certain attributes at a point in time. A flow series is a series which is a measure of activity over a given period. An original time series shows the actual movements in the data over time and includes any movements due to cyclical, seasonal and irregular events.

A cyclical effect is any regular fluctuation or changes in daily, weekly, monthly or annual data. For example, the number of people using public transportation has regular peaks and troughs during each day of the week, depending on the point of time during a day. A seasonal effect is any variation dependent on a particular season of year. For example, fruit and vegeTable prices can vary depending on whether or not they are 'in-season'. An irregular effect is any movement that occurred at a specific point of time, but is unrelated to a season or cycle, for example; a natural disaster, the introduction of legislation, or a one-off major cultural or sporting event.

A seasonally adjusted series involves estimating and removing the cyclical and seasonal effects from the original data. Seasonally adjusting a time series is useful if you wish to understand the underlying patterns of change or movement in a population, without the impact of the seasonal or cyclical effects. For example, employment and unemployment are often seasonally adjusted so that the actual change in employment and unemployment levels can be seen, without the impact of periods of peak employment such as Christmas/New Year when a large number of casual workers are temporarily employed.

[13] http://www.abs.gov.au

A trend series is a seasonally adjusted series that has been further adjusted to remove irregular effects and 'smooth' out the series to show the overall 'trend' of the data over time. For example, the trend is often used when analyzing economic indicators such as employment and unemployment levels.

Example 3.4

Table 3.1 Macroeconomic Indicators

Obs. Number	Year	Inflation	Export	Interest rate
1	1980	2000	1250	20
2	1981	2000	560	25
3	1982	2000	780	10
-	----	---	---	--
-	----	---	---	--
-	----	---	---	--
34	2012	2012	8000	100
35	2013	2012	9000	250

Panel Data[14]

Panel data is consisted of multiple entities or variables observed at multiple time periods. It is also known as longitudinal or cross-sectional time-series data. These entities could be states, companies, individuals, countries, interest rates, exchange rates, etc. Panel analysis uses panel data to examine changes in variables over time but also differences in variables between several subjects simultaneously.

Panel data allows the researcher to control variables which are not possible to be observed or measured such as cultural factors or differences in business practices across companies, variables that change over time but not across entities, and other. New variables can be included at different levels of analysis (i.e. students, schools, districts, states). It is suitable for multilevel or hierarchical modeling.

[14] Baltagi, B.H. (2008).

Example 3.5

Table 3.2 University Graduates by University and Year

Obs. Number	University	Year	Bachelor Grad.	Master Grad.
1	Sivas University	2000	1250	20
2	Tampin University	2000	560	25
3	Bangi University	2000	780	10
-	----	---	---	--
-	----	---	---	--
-	----	---	---	--
34	Sivas University	2012	8000	100
35	Tampin University	2012	9000	250
36	Bangi University	2012	5400	80

Cross Sectional Data

Cross Sectional Data consists of multiple entities or variables which are observed at a point of time. Analysis of cross-sectional data usually consists of comparing the differences among the entities.

For example, the researcher wants to measure current education level in a society. A sample of 2,000 people randomly can be selected from that population (cross section of that population), and examination of the percentage of educated people in the sample can be conducted. This cross-sectional sample provides us with a snapshot of that population, at that one point of time. We can only describe the current proportion of educated people in the society not changes over time, decreasing or increasing, through cross sectional data analysis.

Example 3.6

Table 3.3 Observing Math Test Scores in Bosna Sema Schools

Obs. Number	Test Score	Father Income ($)	Father Education	Gender
1	68	3,000	Primary	Male
2	98	2,000	Primary	Male
3	40	3,500	High School	Female
-	----	---	---	--
-	----	---	---	--
-	----	---	---	--
57	25	10,000	Master	Male

ARTIFICIAL EXPLANATORY VARIABLES

Qualitative explanatory variables are unobserved variables. The constant term, dummy variables and deterministic trends are qualitative explanatory variables. If dependent variable shows a linear trend and not one of the explanatory variables has such pattern, then estimation result will show a lot of residual autocorrelation. A deterministic linear trend variable may be included in the equation to solve this problem.

The Chow Breakpoint Test

If the data exhibits different regimes whether they should have been modeled with different specification or not, it should be tested through the Chow breakpoint test as follows:

$$F = \frac{(S_R - \sum_{i=1}^{m} S_i)/(m-1)K}{(\sum_{i=1}^{m} S_i)/(n-mK)}$$

Where, m : number of sub-samples

S_R : RSS for the whole sample, (restricted sum of squares)

S_i : RSS for the sub- sample,

K: number of estimated parameters

F-distribution is valid if the error terms are independently and identically normally distributed. The null hypothesis of no structural breaks in the sample period is tested with the chow breakpoint test. For a break at one date the test resembles the Chow forecast test, but different null hypothesis is tested with a different test statistic. More dates can be specified for the breakpoint test.[15]

[15] Vogelvang, B. (2005).

Example 3.7

Table 3.4 Eviews Output for Chow Breakpoint Test with One Breakpoint

Chow Breakpoint Test: 1970Q1			
Null Hypothesis: No breaks at specified breakpoints			
Varying regressors: All equation variables			
Equation Sample: 1954Q1 1994Q4			
F-statistic	*74.39175*	*Prob. F(2,160)*	*0.0000*
Log likelihood ratio	*107.8245*	*Prob. Chi-Square(2)*	*0.0000*
Wald Statistic	*148.7835*	*Prob. Chi-Square(2)*	*0.0000*

Table 3.4 Eviews Output for Chow Breakpoint Test with Multiple Breakpoint

Chow Breakpoint Test: 1970Q1 1982Q2 1988Q3			
Null Hypothesis: No breaks at specified breakpoints			
Varying regressors: All equation variables			
Equation Sample: 1954Q1 1994Q4			
F-statistic	*40.60198*	*Prob. F(6,156)*	*0.0000*
Log likelihood ratio	*154.2646*	*Prob. Chi-Square(6)*	*0.0000*
Wald Statistic	*243.6119*	*Prob. Chi-Square(6)*	*0.0000*

DATA TRANSFORMATION

Log Transformation

The log transformation yields appealing interpretation of coefficients and model. The interpretation is good for small changes only. The log transformation makes coefficients invariant to rescaling. For example, lnY may look more normal than Y. But also, lnY has a narrower range than Y.

When or not to Log?

- Do not log zeroes or negative values
- Do not log dummies

- Potentially large monetary variables are often logged: revenue, income, wages;
- Large integer values are often logged: population, number of employees, number of students;
- Small integer values are usually not: age, education, number of children;
- Percentages can be logged or not

For example, an increase in unemployment rate from 4% to 4.5% is half a *percentage point change*. However it implies a 12.5% *percentage change* in the population of unemployed. The difference in logs is log 4.5 − log 4 = 11.8%. This is an approximation of the *percentage change*, good for small changes.

How to Choose?

Sometimes it is unclear which form to choose, and thus we have to:

i- Rely on economic theory and previous studies.
ii- Think about what is implied by particular functional forms for the relevant range of the variables (even if your data does not include the whole relevant range).
iii- Do not compare $R^2 \, or \, AdjR^2$ if the dependent variable is different.
iv- Even in the case of the same dependent variable, e.g. linear and log-linear, beware of selecting the functional form on the sole basis of $R^2 \, or \, AdjR^2$. Selecting the functional form on the basis of fit only gives you an equation that works well for your particular sample.

Differencing

Stationarity is an issue for time series data and is a pre-condition to perform regression analysis. Nonstationary time series data should be converted into stationary data by taking differences. The time series data which is stationary after first differencing is called stationary in order one (I(1)).

Percentage Change

It is calculated by subtracting old one from new one and dividing the difference by old one. It is used for the give emphasis return.

Base Year

A base year is used for comparison of the two or more time series data levels. The arbitrary level of 100 is selected so that percentage changes (either rising or falling) over year can be easily depicted. Any year can be chosen as base year, but generally recent years are chosen and the other observations are adjusted based on the base year.

CHAPTER 4

REGRESSION

DEFINITION

Regression is an econometric technique for estimating the relationships among variables. It helps to analyze how the typical value of the dependent variable changes when any one of the independent variables changes, while the other are held constant (ceteris paribus). Linear regression estimates how much y changes when x changes one unit. The main purpose of linear regression analysis is to assess associations between dependent and independent variables.

SIMPLE REGRESSION MODELS

Linear Regression with one Regressor,

$$y_i = \beta_o + \beta_1 x_i + e_i \quad (4.1)$$

Where;

i denotes observations, i = 1,2,…..,N

$\boldsymbol{y_i}$ is called the dependant variable, regressand, response variable, measured variable, explained variable, outcome variable, experimental variable, or output variable.

$\boldsymbol{x_i}$ is called independent variable, regressor, explanatory variable, predictor variable, controlled variable, input variable or exogenous variable.

$\boldsymbol{\beta_0 + \beta_1 x_i}$ is called population regression line or function

$\boldsymbol{\beta_0}$ is constant term or intercept term
$\boldsymbol{\beta_1}$ is regression coefficient, slope of the regression line

$\boldsymbol{e_i}$ is the deviation of the actual value of an observation from the true regression line which is called the error term, stochastic term, residual term or disturbance term. It consists of omitted independent variables and measurement error.

Regression equation consists of two components:

i- Deterministic component, $\boldsymbol{\beta_0 + \beta_1 x_i}$

ii- Nonsystematic or stochastic components, $\boldsymbol{e_i}$

β_0 is called intercept, $\beta_1, \dots, \beta_K$ are the slope coefficients. They all are called regression coefficients or regression parameters.

Example 4.1

$$\boldsymbol{Income = \beta_0 + \beta_1 Consumption} \qquad (4.2)$$

$\boldsymbol{\beta_1}$ *implies degree and direction of the relationship between Income and Consumption.*

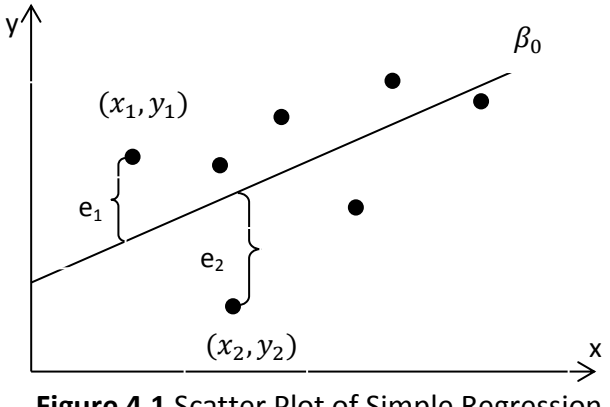


Figure 4.1 Scatter Plot of Simple Regression

$\beta_0 + \beta_1 x_i$ is the population regression line as shown in Figure 4.1
e_1 and e_2 are error terms for the first and second observation

ESTIMATION OF THE LINEAR REGRESSION MODEL

Ordinary Least Squares Estimation Method (OLS)

OLS minimizes the squared difference between observed and predicted parameters from the model. Differences are called residuals or error terms.

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i, \quad (4.3)$$

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_K x_{iK}, \quad (4.4)$$

Predicted values of the regression $\hat{\beta}_1, \hat{\beta}_2, \ldots., \hat{\beta}_K$ are OLS estimators.

And the residuals $\boldsymbol{e_i}$ are computed as follows:

$$e_i = y_i - \hat{y}_i \quad (4.5)$$

the residuals differ from disturbances. The residuals are observed, the disturbances are not. OLS is trying to get a best model in order to obtain residuals as small as possible. Main objective is to minimize the sum of the squares of the residuals (ESS) as follows:

$$f\left(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_K\right) = \sum_{i=1}^{n} u_i^2 \quad (4.6)$$

In order to determine the minimum f with respect to predicted parameters of the regression or to minimize ESS, we can find out the values of regression coefficients which make ESS as small as possible as indicated in Eq.4.5. ESS is a function of regression coefficients.

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{N}(x_i-\bar{x})^2} \quad (4.7)$$

$$and\ \hat{\beta}_1 = \bar{y} - \beta_2\bar{x} \quad (4.8)$$

Example 4.2

Regression equation in Eq.4.2 is estimated using OLS and found following results

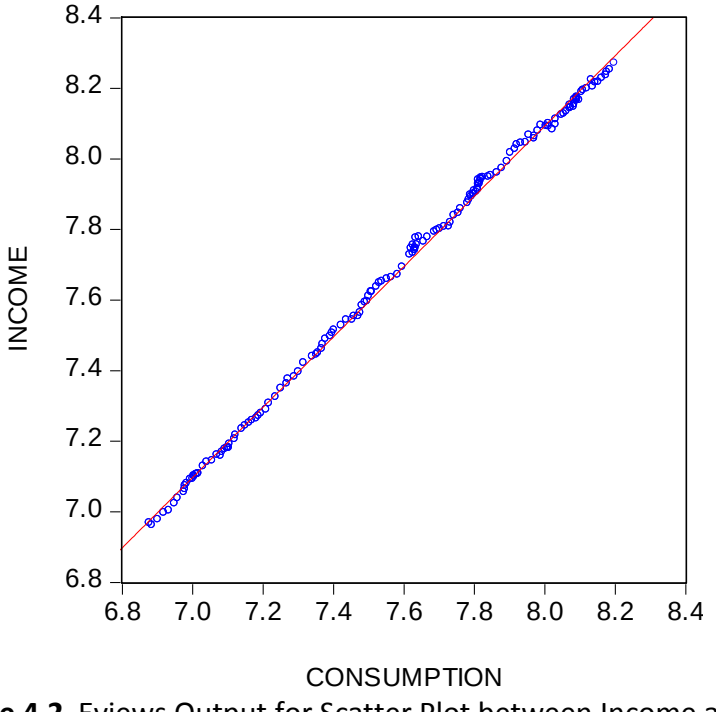


Figure 4.2 Eviews Output for Scatter Plot between Income and Consumption

Using the data income.xls OLS estimation result is presented below:

Table 4.1 Eviews Output for Simple Regression Model Estimates

Dependent Variable: CONSUMPTION
Method: Least Squares
Sample: 1954Q1 1994Q4
Included observations: 164

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
INCOME	*1.001918*	*0.003254*	*307.9252*	*0.0000*
C	*-0.110026*	*0.025047*	*-4.392831*	*0.0000*

R-squared	*0.998294*	*Mean dependent var*	*7.592739*
Adjusted R-squared	*0.998284*	*S.D. dependent var*	*0.388767*
S.E. of regression	*0.016105*	*Akaike info criterion*	*-5.407225*
Sum squared resid	*0.042019*	*Schwarz criterion*	*-5.369422*
Log likelihood	*445.3925*	*Hannan-Quinn criter.*	*-5.391879*
F-statistic	*94817.94*	*Durbin-Watson stat*	*0.284313*
Prob(F-statistic)	*0.000000*		

$$Consumption = -0.110026 + 1.001918\ Income$$
$$\beta_0 = -0.110026,\ and\ \beta_1 = 1.001918$$

Example 4.3:

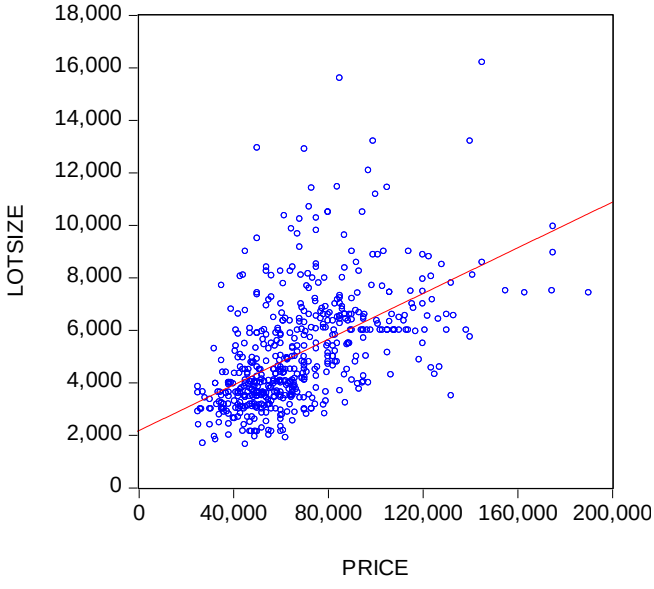


Figure 4.3 Eviews Output for Regression Line

Table 4.2 Eviews Output for Simple Regression Model Estimates

Dependent Variable: PRICE
Method: Least Squares
Sample: 1 546
Included observations: 546

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
LOTSIZE	*6.598768*	*0.445847*	*14.80053*	*0.0000*
C	*34136.19*	*2491.064*	*13.70346*	*0.0000*

R-squared	*0.287077*	*Mean dependent var*	*68121.60*
Adjusted R-squared	*0.285766*	*S.D. dependent var*	*26702.67*
S.E. of regression	*22567.05*	*Akaike info criterion*	*22.89003*
Sum squared resid	*2.77E+11*	*Schwarz criterion*	*22.90579*
Log likelihood	*-6246.977*	*Hannan-Quinn criter.*	*22.89619*
F-statistic	*219.0558*	*Durbin-Watson stat*	*1.089325*
Prob(F-statistic)	*0.000000*		

$$\boldsymbol{House\ Price = 34136.19 + 6.598768\ Lot\ size}$$
$$\beta_0 = 34136.19, and\ \beta_1 = 6.598768$$

Multivariate Regression Model

$$y_i = \beta_o + \beta_1 x_i + \beta_2 z_i + \cdots + \beta_k \theta_i + e_i \qquad (4.9)$$

Multivariate regression model includes more than one regressor.

In regression analysis, main objective is to observe how dependent variable responds to changes in the values of independent variables. The causal relationship which is detected through regression analysis should be based on the relevant economic theory. Each slope coefficient measures the rate of change in the mean value of dependent variable for a unit change in the value of an independent variable, holding the values of the other independent variables constant. How many independent variables should be included in the regression equation depends on the theory, research objective, and assumptions.

It is assumed that the independent variables are not random variable. Their values are fixed in repeating samples. In other words, regression analysis is conditional on the given values of independent variables.

Error term, e_i, includes all the independent variables which are not defined in the regression because of the data unavailability or measurement error. We assume that effect of error term on dependent variable is marginal.

Example 4.4

$$H.Price = \beta_0 + \beta_1 Lotsize + \beta_2 NBed + \beta_3 NBath + \beta_4 NSt. + e_i \qquad (4.10)$$

The multivariate regression equation above shows the impact of Lot Size, Number of bedrooms, Number of Bathrooms, and Number of Stories on the determination of House Price

Table 4.3 Eviews Output for Multiple Regression Model Estimates

Dependent Variable: PRICE
Method: Least Squares
Sample: 1 546
Included observations: 546

Variable	Coefficient	Std. Error	t-Statistic	Prob.
LOTSIZE	5.429174	0.369250	14.70325	0.0000
BEDROOMS	2824.614	1214.808	2.325153	0.0204
BATHRMS	17105.17	1734.434	9.862107	0.0000
STORIES	7634.897	1007.974	7.574494	0.0000
C	-4009.550	3603.109	-1.112803	0.2663

R-squared	0.535547	Mean dependent var	68121.60
Adjusted R-squared	0.532113	S.D. dependent var	26702.67
S.E. of regression	18265.23	Akaike info criterion	22.47250
Sum squared resid	1.80E+11	Schwarz criterion	22.51190
Log likelihood	-6129.993	Hannan-Quinn criter.	22.48790
F-statistic	155.9529	Durbin-Watson stat	1.482942
Prob(F-statistic)	0.000000		

$$\boldsymbol{House\ price} = \mathbf{-40009.550} + \mathbf{5.429174}\ \boldsymbol{Lotsize} + \mathbf{2824.614}\boldsymbol{Bedrooms} + \mathbf{1715.17}\ \boldsymbol{Bathrms} + \mathbf{7634.897}\ \boldsymbol{Stories}$$

where, $\beta_0 = -4009.550$, $\beta_1 = 5.429174, \beta_2 = 2824.614, \beta_3 = 17105.17$, *and* β_4= *7634.897*

CLASSICAL LINEAR REGRESSION MODEL (CLRM) ASSUMPTIONS

i- The regression model is **linear in the parameters** (can be nonlinear or linear in the dependent and independent variables). The Regression model should correct variables without omitted variable bias and have correct functional form.

ii- The regressors are assumed to be fixed or non-stochastic in repeated sampling. This assumption may not be appropriate for all economic data; if independent variables and error term are independently distributed, the results based on the classical assumption hold true provided the analysis is conditional on the particular values of independent variables drawn in the sample. However, if independent variable and error term are uncorrelated, the classical results hold true asymptotically (i.e. in large samples).

iii- Given the values of the $\boldsymbol{x}$ (independent) variables, the expected, or mean, value of the error term is zero. $\boldsymbol{E(e_i|x) = 0}$

The conditional expectation of the error term, given the values of the independent variables, is zero. Since the error term represents the influence of other factors that are not included in the regression equation and may be essentially random, it is very logical to assume that their mean or average value is zero.

As a result of this critical assumption, we can write the regression equation as:

$$E(y_i|x) = Bx + E(e_i|x) = Bx \quad (4.11)$$

This can be interpreted as the model for mean or average value of y_i conditional on the x_i values. This is the population (mean) regression function (PRF). In regression analysis, the main objective is to estimate this function. If there is only one independent variable, it can be visualized as the (population) regression line. If there is more than one independent variable, it would be a curve in a multi-dimensional graph. The estimated PRF is denoted by $\hat{y}_i = bx$.

Suppose that $E(e_i|x) = 3$, then $E(e_i - 3|x) = 0$, we subtract 3 from the error term and add 3 to the intercept (constant term) as follows;

$y_i = (\beta_0 + 3) + \beta_1 x_i + (e_i + 3)$, can be written as,

$y_i = \beta_0^* + \beta_1 x_i + e_i^*$, where $\beta_0^* = \beta_0 + 3$ and $e_i^* = e_i - 3$

then $E(e_i^*|x_i) = 0$. Assumption is satisfied.

iv- The variance of each e_i, given the values of x, is constant, or **homoscedastic**

$$var(e_i|x) = \sigma^2 \quad (4.12)$$

v- There is no correlation between two error terms. That is, there is no **autocorrelation.**

$$cov(e_i, e_j|x) = 0 \quad (4.13)$$

If there is autocorrelation, an increase in the error term in one period affect the error term in the next.

vi- There are no perfect linear relationships among the independent variables. This is the assumption of no perfect **multicollinearity**.

vii- Error term is not correlated with the independent variables.

$$E(e_i|x_{1i}, x_{2i}, \dots, x_{Ki}) = 0 \quad (4.14)$$

viii- All independent variables should be **exogenous** which are defined outside of the model.

ix- The regression model should be correctly specified. Alternatively, there is no **specification bias or specification error** in the model used in empirical analysis. It is implicitly assumed that the number of observations, N, is greater than the number of parameters (K) estimated.

x- Although it is not a part of the CLRM, it is assumed that the error term follows the **normal distribution** with zero mean and (constant) variance. It is necessary for the hypothesis testing.

$$\text{Symbolically, } e_i \sim N(0, \sigma^2) \quad (4.15)$$

STATISTICAL PROPERTIES OF OLS

Linearity

The estimators are linear, that is, they are linear functions of the dependent variable, y_i. Linear estimators are easy to understand and deal with compared to nonlinear estimators. Linearity assumption can be expressed as follows:

$$E(y_i/x_i) = \beta_1 + \beta_{2i}x_{2i} + \beta_{3i}x_{3i} + \cdots + \beta_{ki}x_{ki} \quad (4.16)$$

Linear models can be expressed in a form that is linear in the parameters by a transformation of the variables. Nonlinear models, on the other hand, cannot be transformed to the linear form. The non-linearity of interest here is the one which cannot be accommodated into a linear conditional mean after transformation.

Reset-type test (Ramsey, 1969) is the most common test for testing the linearity assumption. This testing procedure involves the estimation of the following (auxiliary) regression:

$$\hat{e}_i = \omega_1 + \omega_{2i}x_{2i} + \cdots + \omega_{ki}x_{ki} + \vartheta\hat{y}_i^2 + \varepsilon_i \tag{4.17}$$

$H_0: \vartheta = 0,$
$H_1: \vartheta \neq 0.$

We can now test the statistical significance of ϑ using t-test, F-statistic or LM test as follows:

$$t - test = \frac{estimate\ of\ \vartheta}{standard\ error\ of\ \hat{\vartheta}} \tag{4.18}$$

$$F - test = \frac{(\sum_i \hat{e}_t^2 - \sum_i \hat{\varepsilon}_i^2)/1}{(\sum_i \hat{\varepsilon}_i^2)/(T-k-1)} \tag{4.19}$$

$$LM - test = TR^2 \sim \chi^2(1) \tag{4.20}$$

when linearity does not hold, the OLS estimators are biased and inconsistent. In other words, estimation and testing results are invalid and the model should be restructured.

Nonlinearity is seen very clear in a plot of the observed versus predicted values or a plot of residuals versus predicted values, which are a part of standard regression output. The points in the first plot should be symmetrically distributed around a diagonal line or a horizontal line in the second plot. The evidence of a "bowed" pattern should be examined carefully. It indicates that the model makes systematic errors whenever it is making unusually large or small predictions.

If the transformation seems to be appropriate, a nonlinear transformation may be applied to the dependent and/or independent variables, as follows:

- A log transformation may be feasible if the data are strictly positive,

- Adding another regressor which is a nonlinear function of one of the other variables. For example, if you have regressed $\mathbf{y}$ on $\mathbf{x}$, and the graph of residuals versus predicted suggests a parabolic curve, then it may make sense to regress $\mathbf{y}$ on both $\mathbf{x}$ and $\mathbf{x^2}$(i.e., x -squared). The latter transformation is possible even when $\mathbf{x}$ and/or $\mathbf{y}$ have negative values, except logging.

Example 4.5

$$House\ price = -40009.550 + 5.429174\ Lotsize + 2824.614 Bedrooms + 1715.17\ Bathrms + 7634.897\ Stories$$

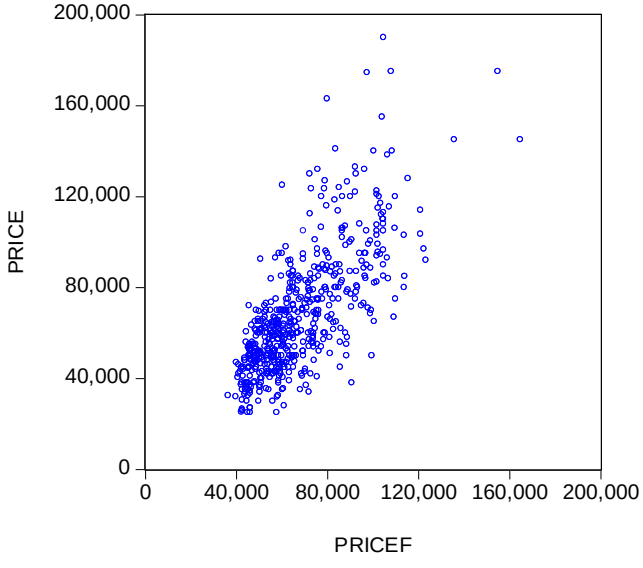


Figure 4.4 Eviews Output for Scatter Graph of Observed and Predicted Values of House Prices Based on The Regression Equation above

Unbiasedness

The estimators $(\hat{\beta}_i)$ are unbiased, that is, in repeated applications of the method, on average, the estimators approach their true values.

$$E(\hat{\beta}_i) = \beta_i \tag{4.21}$$

Efficiency

In the class of linear estimators, OLS estimators have minimum variance. As a result, the true parameter values can be estimated with least possible uncertainty; an unbiased estimator with the least variance is called an efficient estimator.

Normality

The assumption of normality can be expressed as follows:

$$e_i \sim N(0, \sigma^2), \text{or } (y_i/x_i) \sim N(\beta x_i, \sigma^2) \tag{4.22}$$

If the assumption of normality does not hold, then the OLS estimator $(\hat{\beta})$ remains the Best Linear Unbiased Estimator (BLUE), i.e. it has the minimum variance among all linear unbiased estimators. However, without normality one cannot use the standard for the **t** and **F** distributions to perform statistical tests.

The following null hypothesis should be specified before normality test.

The null is that the skewness[16] (α_3) and kurtosis[17] (α_4) coefficients of the conditional distribution of y_i (or, equivalently, of the distribution of e_i) are 0 and 3, respectively:

[16] a measure for the degree of symmetry in the variable distribution
[17] a measure for the degree of peakedness/flatness in the variable distribution.

$H_0: \alpha_3 = 0, (if\ \alpha_3 <$
$0\ then\ f(y_i/x_i)\ is\ skewed\ to\ the\ left\ side$

$H_0: \alpha_3 = 0, (if\ \alpha_3 > 3\ then\ f(y_i/x_i)\ is\ leptokurtic$

The above assumptions can be tested jointly using the Jarque-Bera test (JB, 1981) which follows asymptotically a chi-square distribution:

$$JB\ test = \left[\frac{T}{6}\hat{\alpha}_3^2 + \frac{T}{24}(\hat{\alpha}_4 - 3)^2\right] \sim \chi^2(2) \quad (4.23)$$

$$\hat{\alpha}_3 = \left[(\frac{1}{T}\sum_t \hat{e}_t^3)/(\frac{1}{T}\sum_t \hat{e}_t^2)^{3/2}\right] \quad (4.24)$$

$$\hat{\alpha}_4 = \left[(\frac{1}{T}\sum_t \hat{e}_t^4)/(\frac{1}{T}\sum_t \hat{e}_t^2)^2\right] \quad (4.25)$$

Example 4.6

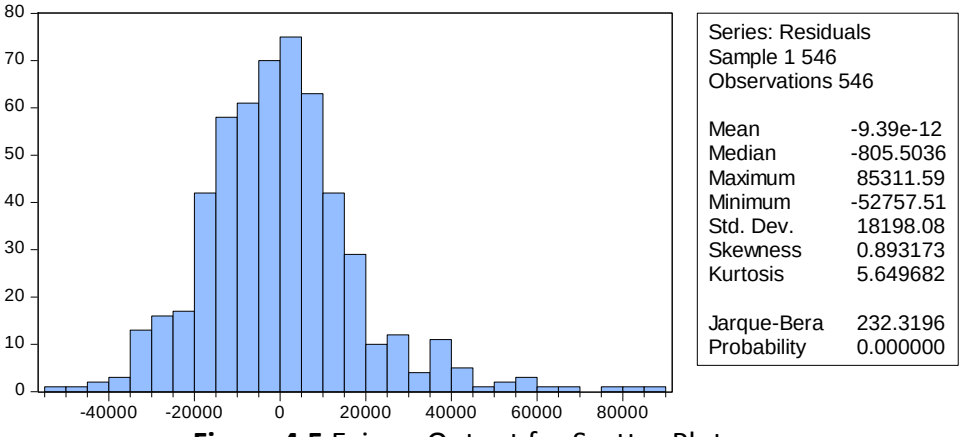


Figure 4.5 Eviews Output for Scatter Plot

Scatter plot shows whether the series are distributed normally or not. It skewed the left side and shows leptokurtic behavior.

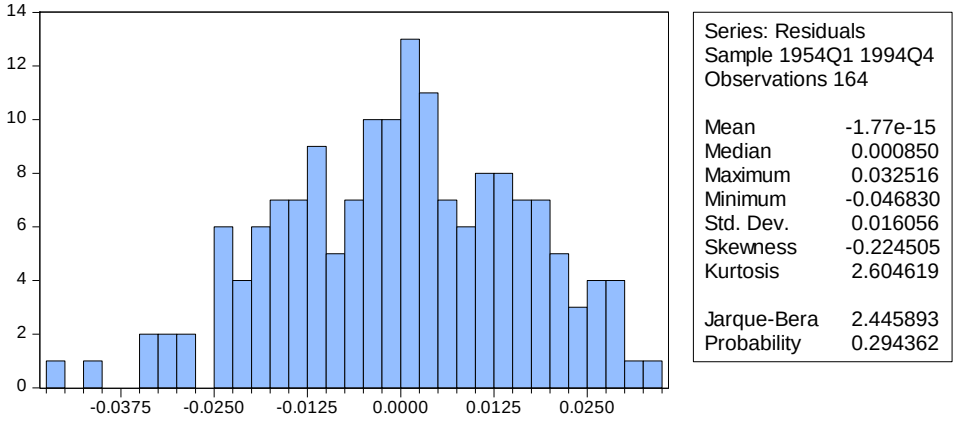


Figure 4.6 Eviews Output for Scatter Plot

Result of skewness and kurtosis indicates that income series are normally distributed around the mean.

If the residuals do not follow a normal pattern, omitted variables, model specification, functional forms and linearity should be checked. Normally, normality does not represent much of a problem in big samples.

GAUSS-MARKOV THEOREM

Under the assumed above mentioned conditions, OLS estimators are BLUE (best linear unbiased estimators). This is the essence of the well-known Gauss–Markov theorem which provides a theoretical justification for the method of least squares. With the added assumption of normality, the OLS estimators are best unbiased estimators (BUE) in the entire class of unbiased estimators, whether linear or not. With normality assumption, CLRM is known as the normal classical linear regression model (NCLRM).

For the linear regression model, an estimate of the **variance of the error term** is:

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{N-K} \qquad (4.26)$$

that is, the residual sum of squares (RSS) divided by (N-K), which is called the degrees of freedom (df), N being the sample size and K being the number of regression parameters estimated, an intercept and (K – 1) slope coefficients. $\hat{\sigma}^2$is called the standard error of the regression (SER) or root mean square.

Example 4.7

Table 4.4 Eviews Output for Simple Regression Model Estimates

Dependent Variable: CONSUMPTION
Method: Least Squares
Included observations: 164

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
INCOME	*1.001918*	*0.003254*	*307.9252*	*0.0000*
C	*-0.110026*	*0.025047*	*-4.392831*	*0.0000*

R-squared	*0.998294*	*Mean dependent var*	*7.592739*
Adjusted R-squared	*0.998284*	*S.D. dependent var*	*0.388767*
S.E. of regression	*0.016105*	*Akaike info criterion*	*-5.407225*
Sum squared resid	*0.042019*	*Schwarz criterion*	*-5.369422*
Log likelihood	*445.3925*	*Hannan-Quinn criter.*	*-5.391879*
F-statistic	*94817.94*	*Durbin-Watson stat*	*0.284313*
Prob(F-statistic)	*0.000000*		

$$SER = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{\sum e_i^2}{N-K}} = \frac{0.042019}{164-2} = 0.016105$$

COEFFICIENT OF DETERMINATION, GOODNESS OF FIT

It measures the proportion of the variance in the dependent variable which is explained by variations in all independent variables.

After estimation of a particular linear model, a question comes up such as: how well does the estimated regression line fit the

observation? A popular measure for the goodness of fit is called R^2 (R squared) and is defined as follows:

Total Sum of Squares (TSS): $\sum y_i^2 = \sum(y_i - \bar{y}^2)$ (4.27)

Explained Sum of Squares (ESS): $\sum(\hat{y}_i - \bar{y})^2$ (4.28)

Residual Sum of Squares (RSS): $\sum e_i^2$ (4.29)

TSS=ESS+RSS

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$R^2 = \frac{1/(N-1)\sum_{i=1}^{N}(\hat{y}_i - \hat{y})^2}{1/(N-1)\sum_{i=1}^{N}(y_i - \hat{y})^2} \quad (4.30)$$

The R^2 indicates which proportion of the sample variation in y_i is explained by the model.

$0 \leq R^2 \leq 1$. The closer to one the better is the fit,
while closer to zero fit is becoming worse.
If $R^2 = 0$, ESS = 0, (RSS = TSS) model does not explain anything.

If $u_i = 0, then\ R^2 = 1$, RSS = 0, (ESS = TSS) meaning that the model fit perfect.

Some source of variation in exploratory variable is much harder to explain and can be incomparable. For example, models which explain consumption, changes in consumption or consumption growth are not comparable in terms of their R^2.

R^2 can be interpreted as a measure of quality of the model and its linear approximation. It measures mainly linear approximation. The value of R^2 never decrease when the number of regressor increase even if the added variables have no explanatory power on regressand. A common way to deal with this problem is to correct the

variance estimates for the degrees of freedom. This gives the value of adjusted R^2, or $\bar{R}^2$ as follows:

$$\bar{R}^2 = \frac{\frac{1}{N-K}\sum_{i=1}^{N}(\hat{y}_i - \hat{y})^2}{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{y})^2} = 1 - \frac{\sum_{i=1}^{N}\frac{u_i^2}{N-K}}{\sum_{i=1}^{N}\frac{(y_i - \hat{y})^2}{N-1}}$$

$$= 1 - [(1 - R^2)(N-1)/(N-K)] \qquad (4.31)$$

This new measure of goodness of fit punishes the inclusion of additional explanatory or independent variables in the model. Even, it may decline when a variable is added. $\bar{R}^2$ is always smaller than R^2.

As the number of independent variable in the regression model increases, the $\bar{R}^2$ become significantly smaller than R^2. The impact of adding more independent variables in the value of R^2 is eliminated through $\bar{R}^2$. The R^2 is always positive, but the values of $\bar{R}^2$ can be negative. Additionaly:

- **i.** An increase in R^2 or $\bar{R}^2$ does not mean that an added variable is statistically significant.
- **ii.** A high R^2 or $\bar{R}^2$ does not mean that the regressors are the true cause of the dependent variable. (still need a causal model.)
- **iii.** A high R^2 or $\bar{R}^2$ does not mean there is no omitted variable bias (nor does low R^2 or $\bar{R}^2$ mean that there is omitted variable bias).
- **iv.** A high R^2 or $\bar{R}^2$ not necessarily mean that the regressors are appropriate nor does a low R^2 or $\bar{R}^2$ mean the regressors are inappropriate.

Example 4.8

$$\bar{R}^2 = 1 - [(1 - 0.998294)(164 - 1)/(164 - 2)] = 0.998284$$

It is possible to test whether the increase in R^2 is statistically significant. It is same as testing whether the coefficient of new added

regressor is equal to zero. The appropriate F-statistics can be written as follows:

$$f = \frac{(R_1^2 - R_0^2)/J}{(1-R_1^2)/(N-K)} \quad (4.32)$$

Where, R_1^2 and R_0^2 denote the R^2 in the model with and without new added regressor. J is the number of newly added regressors. It measures the dispersion of the dependent variables estimates around its mean.

Example 4.9

Table 4.5 Eviews output for multiplee regression model estimates

Dependent Variable: PRICE
Method: Least Squares
Sample: 1 546
Included observations: 546

Variable	Coefficient	Std. Error	t-Statistic	Prob.
LOTSIZE	4.871811	0.355481	13.70484	0.0000
BATHRMS	16285.39	1641.569	9.920629	0.0000
BEDROOMS	2802.653	1147.600	2.442186	0.0149
STORIES	5711.979	981.0879	5.822087	0.0000
AIRCO	13851.00	1702.051	8.137832	0.0000
C	-932.5029	3424.701	-0.272287	0.7855

R-squared	0.586284	Mean dependent var	68121.60
Adjusted R-squared	0.582454	S.D. dependent var	26702.67
S.E. of regression	17254.69	Akaike info criterion	22.36048
Sum squared resid	1.61E+11	Schwarz criterion	22.40776
Log likelihood	-6098.412	Hannan-Quinn criter.	22.37897
F-statistic	153.0489	Durbin-Watson stat	1.546388
Prob(F-statistic)	0.000000		

$$f = \frac{(0.586284 - 0.535547)/1}{(1 - 0.586284^2)/(546 - 6)} = 41.747$$

CHAPTER 5

CLASSICAL LINEAR REGRESSION MODEL (CLRM)

LINEARITY

Linearity is the situation in which the relationship between the dependent variable and the independent variables is linear. It is the main assumption of the OLS estimation method. If linearity exists, the effect of an exogenous variable on the endogenous variable is the same for all values of the other exogenous variables in the model (that is, the slope of the population regression function is constant, so that the effect on **y** of a unit change in **x** does not depend on the value of one or more exogenous variable). In this sense, non-linearity has important consequences for the interpretation of the estimation results.

The regression function can be nonlinear in two different cases:

i- The effect on endogenous variable of a change in exogenous variable might be greater or smaller for different values of the exogenous variable.

ii- The effect on endogenous variable of a change in exogenous variable depends on the value of another exogenous variable. For example, the effect on test scores of reducing the class durations may be greater or lower in schools where the parents of children have a higher level of education.

Nonlinear regression model may be:

- The regression functions of a nonlinear function of the exogenous variables but is a linear function of the unknown parameters.
- The regression functions of a nonlinear function of the unknown parameters and may or may not be a nonlinear function of the exogenous variables (logit, probit, etc.).

Modeling Nonlinearities (with OLS)

- Based on the theory identify a nonlinear relationship
- Specify a nonlinear function (including a quadratic or an interaction term) and estimate by ***OLS***
- Compare two models and investigate whether or not the nonlinear model improves upon a linear model
- Plot the nonlinear regression function
- Estimate the effect on endogenous variable of a unit change in exogenous variable.
- Interpret the estimation results

Testing Linearity

The linearity of the regression can be examined:

i- **Visually**- by plots of the residuals against any of the independent variables, or against the predicted values. The points should be symmetrically distributed around a diagonal line in the former plot or a horizontal line in the latter plot. Look carefully for evidence of a

"bowed" pattern, indicating that the model makes systematic errors whenever it is making unusually large or small predictions.

Example 5.1

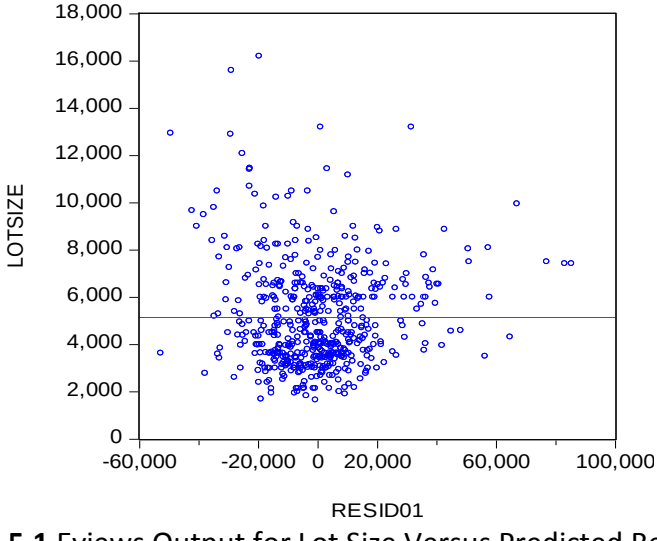


Figure 5.1 Eviews Output for Lot Size Versus Predicted Residual

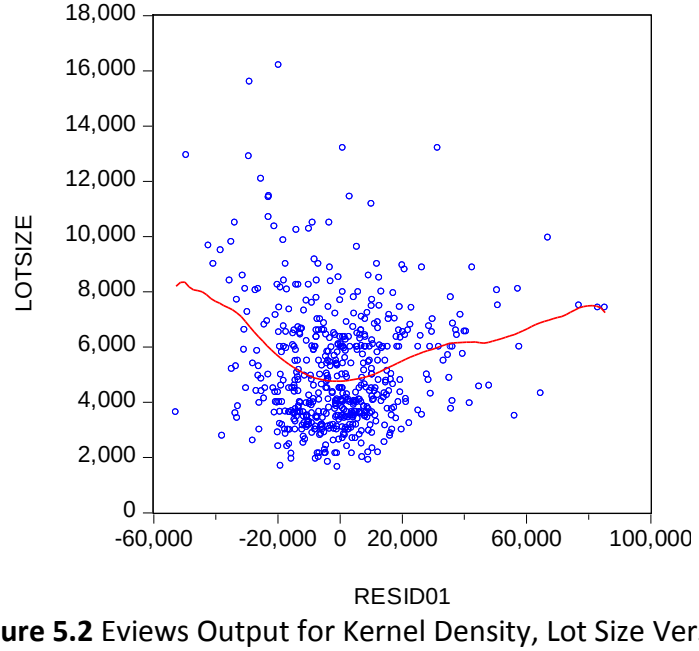


Figure 5.2 Eviews Output for Kernel Density, Lot Size Versus Predicted Residual

Example 5.2

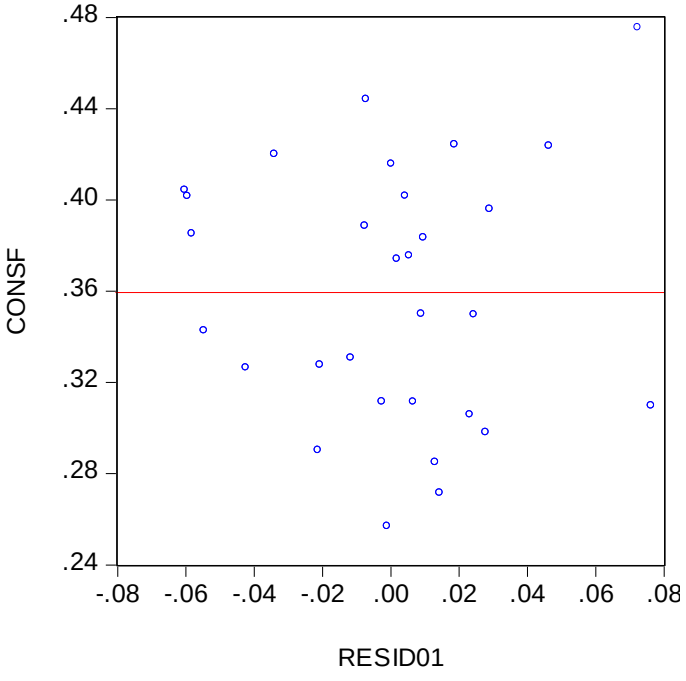


Figure 5.3 Eviews Output for Ice-Cream Consumption Predicted Versus Residual

ii- **Ramsey RESET test** (Ramsey, 1969) may be constructed by adding powers of fitted values to the regression model. It is a test of linear specification against a nonlinear specification. It tests the hypothesis stating that the values of added parameters are zero, such as:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots \hat{\beta}_k x_{ik} + \gamma_1(\hat{y}_i)^2 + \gamma_2(\hat{y}_i)^3 + \gamma_3(\hat{y}_i)^4 + u_i \quad (5.1)$$

includes square fitted values.

The null hypothesis of linearity is

$$H_0: \gamma_1 = \gamma_2 = \gamma_3 = 0 \quad (5.2)$$

which is tested by F-statistics. If ***F-statistics*** is larger than its critical value, conclude that there is enough evidence to reject the null hypothesis of linearity.

$$H_0: the\ specification\ is\ linear$$

$$H_1: the\ specification\ is\ non-linear$$

F-statistics is formed as follows;

$$F_{(M,N-K-1)} = \frac{(SSR_{\hat{y}} - SSR_{\hat{y}^2})/M}{SSR_{\hat{y}^2}/(N-K)} = \frac{(SSR_R - SSR_{UR})/M}{SSR_{UR}/(N-K)} \quad (5.3)$$

SSR : sum of squared residuals
M : the number of restrictions
N : the number of observations
K : the number of explanatory variables

Example 5.3

Table 5.1 Eviews Output for Ramsey Reset Test

Ramsey RESET Test
Equation: UNTITLED
Specification: PRICE LOTSIZE BEDROOMS BATHRMS STORIES C
Omitted Variables: Powers of fitted values from 2 to 4

	Value	*df*	*Probability*
F-statistic	*0.339738*	*(3, 538)*	*0.7966*
Likelihood ratio	*1.033391*	*3*	*0.7932*

F-test summary:

	Sum of Sq.	*df*	*Mean Squares*
Test SSR	*3.41E+08*	*3*	*1.14E+08*
Restricted SSR	*1.80E+11*	*541*	*3.34E+08*
Unrestricted SSR	*1.80E+11*	*538*	*3.35E+08*
Unrestricted SSR	*1.80E+11*	*538*	*3.35E+08*

LR test summary:

	Value	*df*
Restricted LogL	*-6129.993*	*541*
Unrestricted LogL	*-6129.476*	*538*

Unrestricted Test Equation:
Dependent Variable: PRICE
Method: Least Squares
Sample: 1 546
Included observations: 546

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
LOTSIZE	*1.531674*	*19.47670*	*0.078641*	*0.9373*
BEDROOMS	*556.3758*	*10084.57*	*0.055171*	*0.9560*
BATHRMS	*5127.654*	*61702.43*	*0.083103*	*0.9338*
STORIES	*2241.466*	*27288.20*	*0.082140*	*0.9346*
C	*8644.588*	*85554.31*	*0.101042*	*0.9196*
FITTED^2	*1.64E-05*	*6.41E-05*	*0.256362*	*0.7978*
FITTED^3	*-1.42E-10*	*4.81E-10*	*-0.294963*	*0.7681*
FITTED^4	*3.96E-16*	*1.28E-15*	*0.310148*	*0.7566*

R-squared	*0.536426*	*Mean dependent var*	*68121.60*
Adjusted R-squared	*0.530394*	*S.D. dependent var*	*26702.67*
S.E. of regression	*18298.76*	*Akaike info criterion*	*22.48160*
Sum squared resid	*1.80E+11*	*Schwarz criterion*	*22.54464*
Log likelihood	*-6129.476*	*Hannan-Quinn criter.*	*22.50624*
F-statistic	*88.93529*	*Durbin-Watson stat*	*1.484699*
Prob(F-statistic)	*0.000000*		

*The reset **F statistic** is equal to 0.339 and the corresponding p-value is 0.7966. There is no evidence to reject the null hypothesis of linearity at the 5% significance level.*

The Ways to Correct Linearity Problem:

i- Using a nonlinear transformation to the dependent and/or independent variables may be used if the transformation is appropriate. For example, if the data are strictly positive, a log transformation may be feasible.

ii- Add another regressor which is a nonlinear function of one of the other variables. For example, regress $\boldsymbol{y}$ on $\boldsymbol{x}$, and the graph of residuals versus predicted suggests a parabolic curve, then it may make sense to regress $\boldsymbol{y}$ on both $\boldsymbol{x}$ and $\boldsymbol{x^2}$(i.e., $\boldsymbol{x}$-squared). The latter

transformation is possible even when $\boldsymbol{x}$ and/or $\boldsymbol{y}$ have negative values, whereas logging may not be.

HETEROSCEDASTICITY

$\boldsymbol{V(u_t) = \sigma^2}$ for all observations. That is, the variance of the error term is constant (Homoscedasticity) over the sample period. If the error terms do not have constant variance, they are said to be heteroscedastic.

Errors may increase as the value of an independent variable increases. Annual family expenditures for education differ among rich and poor families or educated and uneducated families. A research may include two income group of families may face heteroscedasticity problem. Measurement error can be also occurred if some of respondant gives more accurate responses.

Heteroscedasticity is occurred, If;

- There are subpopulation differences or other interaction effects. For instance, the effect of education in employment differs for villagers and urban parts.
- There are model misspecifications. For instance, instead of using $\boldsymbol{y}$, using the log of $\boldsymbol{y}$ or instead of using $\boldsymbol{x}$, using $\boldsymbol{x^2}$.
- There are omitted variables. Omitting the important variables from the model may cause bias. In the correctly specified model, the patterns of heteroscedasticity are expected disappear.

If the plot of residuals shows some uneven envelope of residuals, so that the width of the envelope is considerably larger for some values of $\boldsymbol{x}$ than for others, a more formal test for heteroscedasticity should be conducted.

H_0 : the variance of the error term is constant (homoscedastic)
H_1 : the variance of the error term is heteroscedastic

Consequences

Heteroscedasticity does not result in unbiased parameter estimates.

OLS estimates are no longer BLUE. That is, among all of the unbiased estimators, OLS does not provide the estimate with the smallest variance. Depending on the nature of the heteroscedasticity, significance tests can be too high or too low. The standard errors are biased when heteroscedasticity is present. This in turn leads to bias in test statistics and confidence interval.

Detection

i- It can be inspected through **Visual Inspection**. Do a visual inspection of residuals plotted against fitted values; or, plot the independent variables suspected to be correlated with the variance of the error term.

Example 5.4

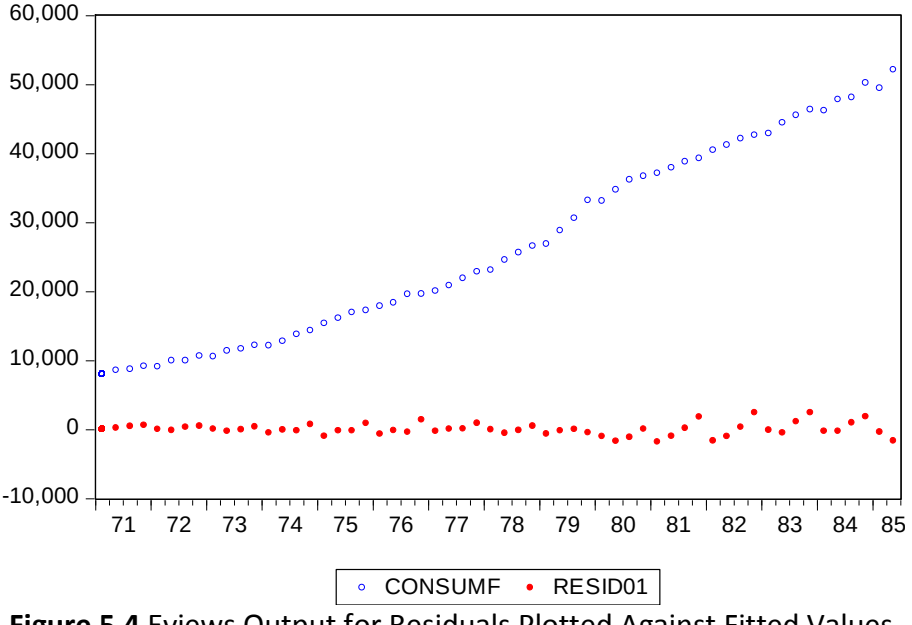


Figure 5.4 Eviews Output for Residuals Plotted Against Fitted Values

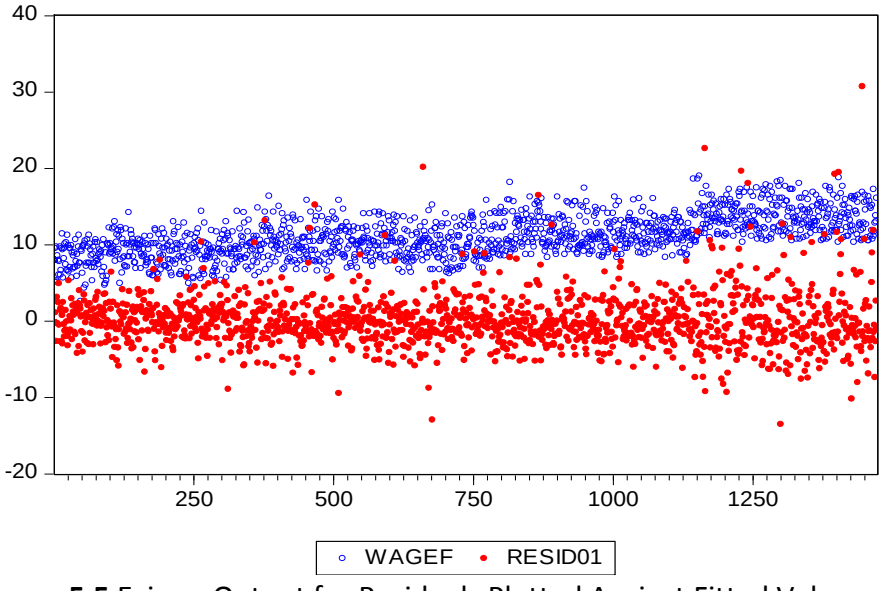


Figure 5.5 Eviews Output for Residuals Plotted Against Fitted Values

ii- White Test for Heteroscedasticity (White, 1980)

It is actually a special case of Breusch-Pagan test. It involves an auxiliary regression of squared residuals, but excludes any higher order terms. It is very general. If the number of the observation is small, power of the white test becomes weak. It can be performed by obtaining least squares residuals and modeling the square residuals as a multiple regression which includes independent variables and their squares and second degree products (interaction term) as follows:

$$e_i^2 = \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 (x_{i1})^2 + \gamma_4 (x_{i2})^2 + \gamma_5 (x_{i1} x_{i2}) + \varepsilon_i \quad (5.4)$$

White test for heteroscedasticity is performed through the null hypothesis of no heteroscedasticity. Instead of ***F*** statistics, the statistics $\boldsymbol{NR^2}$ (sample size multiplied by the coefficient of determination) is in use.

$$H_0: \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5 = 0 \quad (5.5)$$

Example 5.5

Table 5.2 Eviews Output for White Test for Heteroscedasticity

Heteroskedasticity Test: White

F-statistic	*9.564407*	*Prob. F(2,161)*	*0.0001*
*Obs*R-squared*	*17.41601*	*Prob. Chi-Square(2)*	*0.0002*
Scaled explained SS	*13.63430*	*Prob. Chi-Square(2)*	*0.0011*

Dependent Variable: RESID^2
Method: Least Squares
Sample: 1954Q1 1994Q4
Included observations: 164

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
C	*0.009461*	*0.011185*	*0.845830*	*0.3989*
INCOME	*-0.002690*	*0.002937*	*-0.915986*	*0.3610*
INCOME^2	*0.000194*	*0.000192*	*1.006889*	*0.3155*

R-squared	*0.106195*	*Mean dependent var*	*0.000256*
Adjusted R-squared	*0.095092*	*S.D. dependent var*	*0.000326*
S.E. of regression	*0.000310*	*Akaike info criterion*	*-13.30390*
Sum squared resid	*1.54E-05*	*Schwarz criterion*	*-13.24720*
Log likelihood	*1093.920*	*Hannan-Quinn criter.*	*-13.28088*
F-statistic	*9.564407*	*Durbin-Watson stat*	*1.050351*
Prob(F-statistic)	*0.000119*		

iii- The Breusch-Pagan Test

Proposed by Breusch and Pagan (1980). It is a Lagrange multiplier test for heteroscedasticity and designed to detect any linear form of heteroscedasticity.

It tests the null hypothesis that the error variances are all equal versus the alternative that the error variances are a multiplicative function of one or more variables.

Example 5.6

Table 5.3 Eviews Output for the Breush Pagan

Heteroskedasticity Test: Breusch-Pagan-Godfrey			
F-statistic	*29.82872*	*Prob. F(3,1468)*	*0.0000*
*Obs*R-squared*	*84.57450*	*Prob. Chi-Square(3)*	*0.0000*
Scaled explained SS	*450.4796*	*Prob. Chi-Square(3)*	*0.0000*

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Sample: 1 1472
Included observations: 1472

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
C	*-27.03619*	*4.356212*	*-6.206352*	*0.0000*
MALE	*4.711490*	*2.170099*	*2.171094*	*0.0301*
EXPER	*0.716098*	*0.107900*	*6.636675*	*0.0000*
EDUC	*7.224143*	*0.907955*	*7.956499*	*0.0000*

R-squared	*0.057456*	*Mean dependent var*	*12.55743*
Adjusted R-squared	*0.055529*	*S.D. dependent var*	*41.11146*
S.E. of regression	*39.95372*	*Akaike info criterion*	*10.21603*
Sum squared resid	*2343368.*	*Schwarz criterion*	*10.23042*
Log likelihood	*-7515.001*	*Hannan-Quinn criter.*	*10.22140*
F-statistic	*29.82872*	*Durbin-Watson stat*	*1.991852*
Prob(F-statistic)	*0.000000*		

iv- Ramsey's Test for Heteroscedasticity

RESET stands for *Regression Specification Error Test* and was proposed by Ramsey (1969). The classical normal linear regression model is specified as:

$$y = \beta x + e \qquad (5.6)$$

Where disturbance vector $\boldsymbol{e}$ is presumed to follow the multivariate normal distribution $\boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma^2 I})$. Specification error is a comprehensive term which

covers departures the maintained model assumptions. Serial correlation, heteroscedasticity, and non-normality violate error term assumptions which are $\boldsymbol{N(0, \sigma^2 I)}$.

RESET specification test is a general test for the following types of specification errors:

a. Omitted variable bias (if set of independent variables do not include all relevant variables).

b. Wrong functional form (if some or all of the dependent and independent variables should be transformed to logs, powers, reciprocals, or in some other way, but not).

c. Correlation between independent variable and error term (may be caused measurement error in simultaneity).

d. The existence of lagged dependent variable values.

e. Serially correlated error terms.

Under those specification errors above, OLS estimators is biased and inconsistent. Conventional inference procedures are not valid. Ramsey (1969) indicates that any or all of these specification errors above produce a non-zero mean vector.

The null and alternative hypotheses of the RESET test are:

$$H_0: e \sim N(0, \sigma^2 I)$$
$$H_1: e \sim N(\mu, \sigma^2 I) \;\; \mu \neq 0$$

The test is based on following augmented regression equation:

$$y = \beta x + \gamma Z + e \qquad (5.7)$$

Specification error test examines the restriction of $\boldsymbol{\gamma = 0}$. The main question in constructing the RESET test is to identify the variables which should be in the $\boldsymbol{Z}$ matrix. The $\boldsymbol{Z}$ matrix can include the variables which are not in the original specification, so that the test of $\boldsymbol{\gamma = 0}$ is called as the omitted variables test.

Example 5.7

Table 5.4 Eviews Output for the Ramsey´s Test for Heteroscedasticity

Ramsey RESET Test
Equation: UNTITLED
Specification: CONS INCOME PRICE TEMP TIME C
Omitted Variables: Powers of fitted values from 2 to 3

	Value	*df*	*Probability*
F-statistic	*5.331333*	*(2, 23)*	*0.0125*
Likelihood ratio	*11.42686*	*2*	*0.0033*

F-test summary:

	Sum of Sq.	*df*	*Mean Squares*
Test SSR	*0.010957*	*2*	*0.005478*
Restricted SSR	*0.034590*	*25*	*0.001384*
Unrestricted SSR	*0.023634*	*23*	*0.001028*
Unrestricted SSR	*0.023634*	*23*	*0.001028*

LR test summary:

	Value	*df*
Restricted LogL	*58.91245*	*25*
Unrestricted LogL	*64.62588*	*23*

Unrestricted Test Equation:
Dependent Variable: CONS
Method: Least Squares
Sample: 1 30
Included observations: 30

Variable	Coefficient	Std. Error	t-Statistic	Prob.
INCOME	0.036697	0.021613	1.697907	0.1030
PRICE	-19.02496	12.82023	-1.483980	0.1514
TEMP	0.060590	0.038732	1.564338	0.1314
TIME	0.017399	0.012889	1.349847	0.1902
C	3.614740	2.410354	1.499672	0.1473
FITTED^2	-53.50632	32.01791	-1.671137	0.1082
FITTED^3	54.47286	29.16295	1.867879	0.0746

R-squared	0.811717	Mean dependent var	0.359433
Adjusted R-squared	0.762599	S.D. dependent var	0.065791
S.E. of regression	0.032056	Akaike info criterion	-3.841725
Sum squared resid	0.023634	Schwarz criterion	-3.514779
Log likelihood	64.62588	Hannan-Quinn criter.	-3.737132
F-statistic	16.52607	Durbin-Watson stat	1.063651
Prob(F-statistic)	0.000000		

v- Test for ARCH Effect

It is common for financial variables. Engel (1982) detected that large and small forecast errors tend to occur in clusters so that the conditional variance of error term is the autoregressive function of the past errors. Ignoring ARCH effect can result in inefficiency of the estimation. ARCH(q) effect can be written as follows:

$$\sigma_t^2 = \gamma_0 + \gamma_1 e_{t-1}^2 + \gamma_2 e_{t-2}^2 + \cdots + \gamma_q e_{t-q}^2 + \varepsilon_t \qquad (5.8)$$

This is a test for ARCH (q) versus ARCH (0).

Example 5.8

Table 5.5 Eviews Output for the ARCH Effect

Heteroskedasticity Test: ARCH			
F-statistic	*0.373376*	*Prob. F(1,27)*	*0.5463*
*Obs*R-squared*	*0.395563*	*Prob. Chi-Square(1)*	*0.5294*

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Sample (adjusted): 2 30
Included observations: 29 after adjustments

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
C	*0.000879*	*0.000326*	*2.700516*	*0.0118*
RESID^2(-1)	*0.111842*	*0.183033*	*0.611045*	*0.5463*

R-squared	*0.013640*	*Mean dependent var*	*0.000993*
Adjusted R-squared	*-0.022892*	*S.D. dependent var*	*0.001426*
S.E. of regression	*0.001442*	*Akaike info criterion*	*-10.17897*
Sum squared resid	*5.62E-05*	*Schwarz criterion*	*-10.08467*
Log likelihood	*149.5950*	*Hannan-Quinn criter.*	*-10.14943*
F-statistic	*0.373376*	*Durbin-Watson stat*	*1.456000*
Prob(F-statistic)	*0.546280*		

vi- Goldfeld-Quant (GQ) Test

The GQ test (GQ, 1965) can be used when it is believed that the variance of the error term increases consistently or decreases consistently as $\boldsymbol{x}$ increases. This test is commonly used because it is easy to apply when one of the regressors (or another r.v.) is considered to possess proportionality factor of heteroscedasticity. The test has two limits: its difficult to reject the null hypothesis of homoscedasticity and the fact that it do not allow to verify other forms of heteroscedasticity. This test is based on the hypothesis that the error variance is related to a regressor $\boldsymbol{x}$.

The test procedure is the following:

a) The observations on $\boldsymbol{y}$ and $\boldsymbol{x}$ are sorted following the ascending order of the regressor $\boldsymbol{x}$.

b) The sample observations are divided into three subsamples omitting the central one.

c) The regression models are estimated through OLS on the first and third subsample (then on (n-c)/2 observations; the number of observations considered has to be sufficiently large.

d) The relative RSS is calculated, denoted as RSS_1 and RSS_2 and derived the Goldfeld-Quandt test as follows:

$$GQ = R = RSS_2/RSS_1 \qquad (5.9)$$

If the sample value of the ***F*** statistics is greater than the critical value at the chosen significance level, the null hypothesis of homoscedasticity is rejected. If ***R*** is large, then $\boldsymbol{RSS_2}$ is greater that $\boldsymbol{RSS_1}$, which means that residuals increase with the regressor.

The power of this test depends on the number of omitted observations (usually n/3 observations have to be omitted). If too many observations are excluded, the $\mathbf{RSS_2}$ and $\mathbf{RSS_1}$ will take too low degrees of freedom, if too few observations are excluded, the test power is low because the comparison between $\mathbf{RSS_2}$ and $\mathbf{RSS_1}$ becomes less effective.

vii- Park Test

This test procedure involves three steps:

a) Modeling OLS estimation to derive the OLS residuals,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, \quad (5.10)$$

OLS estimates produce following OLS residuals;

$$\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}) \quad (5.11)$$

b) The derivation of the $\ln(e_i^2)$ which are considered as dependent variable in the regression;

$$\ln(e_i^2) = \alpha_0 + \alpha_1 lnz_i + u_i \quad (5.12)$$

where the only regressor is the log of the real value considered as proportionality factor.

c) The estimation results of the model are used to verify the presence of heteroscedastic errors.

viii- Glesjer LM Test

The Glejser (1969) test is similar to the Breusch-Pagan-Godfrey test. This test tests against an alternative hypothesis of heteroscedasticity. The auxiliary regression that Glejser proposes regresses the absolute value of the residuals from the original equation. As with the previous tests, this statistic is distributed from a chi-squared distribution with degrees of freedom equal to the number of variables.

ix- Harvey-Godfrey LM Test

The Harvey (1976) test for heteroscedasticity is similar to the Breusch-Pagan-Godfrey test. Harvey tests a null hypothesis of no heteroscedasticity against heteroscedasticity of the form of

$$\sigma_t^2 = \exp(z_t'\alpha), \tag{5.13}$$

where, again z_t is a vector of independent variables.

Example 5.9

Table 5.6 Eviews Output for the Glesjer LM Test

Heteroskedasticity Test: Glejser			
F-statistic	*1.415749*	*Prob. F(4,25)*	*0.2578*
*Obs*R-squared*	*5.540550*	*Prob. Chi-Square(4)*	*0.2362*
Scaled explained SS	*5.568536*	*Prob. Chi-Square(4)*	*0.2338*

Test Equation:
Dependent Variable: ARESID
Method: Least Squares
Sample: 1 30
Included observations: 30

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
C	*0.237778*	*0.194819*	*1.220507*	*0.2337*
INCOME	*-0.000224*	*0.001399*	*-0.160103*	*0.8741*
PRICE	*-0.737679*	*0.506247*	*-1.457151*	*0.1575*
TEMP	*0.000317*	*0.000287*	*1.105262*	*0.2796*
TIME	*-0.000379*	*0.000936*	*-0.404613*	*0.6892*

R-squared	*0.184685*	*Mean dependent var*	*0.025450*
Adjusted R-squared	*0.054235*	*S.D. dependent var*	*0.022864*
S.E. of regression	*0.022235*	*Akaike info criterion*	*-4.623290*
Sum squared resid	*0.012360*	*Schwarz criterion*	*-4.389757*
Log likelihood	*74.34934*	*Hannan-Quinn criter.*	*-4.548580*
F-statistic	*1.415749*	*Durbin-Watson stat*	*1.380946*
Prob(F-statistic)	*0.257755*		

Example 5.10

Table 5.7 Eviews Output for the Harvey-Godfrey Test

Heteroskedasticity Test: Harvey			
F-statistic	*1.026380*	*Prob. F(4,25)*	*0.4130*
*Obs*R-squared*	*4.231692*	*Prob. Chi-Square(4)*	*0.3756*
Scaled explained SS	*6.189628*	*Prob. Chi-Square(4)*	*0.1854*

Test Equation:
Dependent Variable: LRESID2
Method: Least Squares
Sample: 1 30
Included observations: 30

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
C	*0.490030*	*23.89891*	*0.020504*	*0.9838*
INCOME	*-0.103229*	*0.171591*	*-0.601596*	*0.5529*
PRICE	*-4.340469*	*62.10253*	*-0.069892*	*0.9448*
TEMP	*0.025815*	*0.035195*	*0.733499*	*0.4701*
TIME	*-0.023174*	*0.114794*	*-0.201871*	*0.8417*

R-squared	*0.141056*	*Mean dependent var*	*-8.529700*
Adjusted R-squared	*0.003625*	*S.D. dependent var*	*2.732573*
S.E. of regression	*2.727615*	*Akaike info criterion*	*4.995744*
Sum squared resid	*185.9971*	*Schwarz criterion*	*5.229277*
Log likelihood	*-69.93616*	*Hannan-Quinn criter.*	*5.070453*
F-statistic	*1.026380*	*Durbin-Watson stat*	*1.543329*
Prob(F-statistic)	*0.413020*		

Which Test to Apply?

To choose the wrong test may not detect presence of heteroscedasticty. The most general test is the White test. But, it has limited power against large number of alternatives. In most cases visual inspections of the residuals can be the best method to detect heteroscedasticty correctly.

The Ways to Correct Heteroscedasticity Problem

A common way to alleviate this problem is to use logarithms of all variables rather than their levels. Consequently, our first step in handling the heteroscedasticity problem is to consider a log linear model.

As noted before, sometimes heteroscedasticity results from improper model specification. There may be subgroup differences. Effects of variables may not be linear. Perhaps some important variables have been left out of the model. If these represent a problem then deal with them first.

Using Weighted Least Squares is a more difficult option but superior when you can make it work right. It uses weighted least squares. Generalized Least Squares (GLS) is a technique that will always yield estimators that are BLUE when either heteroscedasticity or serial correlation is present.

AUTOCORRELATION

Autocorrelation can only occur in the models that include time series data and it means that either the model is specified with an insufficient number of lagged variables or not all the relevant explanatory variables are specified in the model.

The error term catch the influence of the not included variables affecting dependent variable. Persistence effect of the excluded variables causes positive autocorrelation. If those excluded variables are observable and includable in the model, autocorrelation test result is an indication of a misspecification model.

Autocorrelation test is also regarded as misspecification test. Incorrect functional forms, omitted variables and an inadequate dynamic specification of the model can cause autocorrelation.

Consequences of Autocorrelation in the Residuals

i- The standard errors are underestimated, so ***t-values*** are overestimated.

ii- High values for the ***t-statistics*** and $\mathbf{R^2}$ are observed in the estimation output. It means that the result is false if the output is not correctly interpreted

iii- OLS estimates remain unbiased, but it becomes inefficient.

First Order Autocorrelation

The most popular form of the autocorrelation is the first-order autoregressive process;

$$y_t = \grave{x}_t\beta + u_t \tag{5.14}$$

Error term $\boldsymbol{u_t}$ is assumed to depend on its previous value as follows;

$$u_t = \rho u_{t-1} + v_t \tag{5.15}$$

where, $\boldsymbol{v_t}$ is an error term with mean zero and constant variance $\boldsymbol{\sigma_v^2}$, which exhibits no serial correlation. It means that the value of the error term in any observation is equal to $\boldsymbol{\rho}$ times its value in the previous observation plus a new error term. If $\boldsymbol{\rho = 0, u_t = v_t}$ then there is no residual autocorrelation and autocorrelation condition of OLS (Gauss-Markov) is satisfied.

If $|\boldsymbol{\rho}| < 1$, then the first autoregressive process is stationary. It means that the mean, variance and covariance of $\boldsymbol{u_t}$ do not change over time.

$\boldsymbol{v_t}$satisfies the Gauss-Markov conditions. The transformation like $\boldsymbol{u_t - \rho u_{t-1}}$ will generate homoscedastic non autocorrelated errors.

That is all observations should be transformed as $\boldsymbol{y_t - \rho y_{t-1}}$ and $\boldsymbol{x_t - \rho x_{t-1}}$.[18]

If $\boldsymbol{\rho = 0}$, no autocorrelation is present and OLS is BLUE. If $\boldsymbol{\rho \neq 0}$, OLS estimator will be misleading because standard errors will be based on the wrong formula.

There are various autocorrelation tests. All tests have same null hypothesis of absence of autocorrelation in the disturbance term. The tests have different alternative hypothesis because of differences of the order of the autocorrelation.

The existence of autocorrelation may be an indication of misspecification. A possible way to eliminate autocorrelation problem is to change model specification.

For example;

H_0: no autocorrelation in the disturbance term

$$H_1: u_t = \varphi_1 u_{t-1} + \varphi_2 u_{t-2} + \cdots \varphi_p u_{t-p} + \nu_t \qquad (5.16)$$

(the disturbance term has autocorrelation of the pth order)

With $\upsilon_t \sim NID(0, \sigma_\upsilon^2)$

Detection

i- Graphical method

Plotting error term to detect autocorrelation

[18] Verbeek, M. (2004).

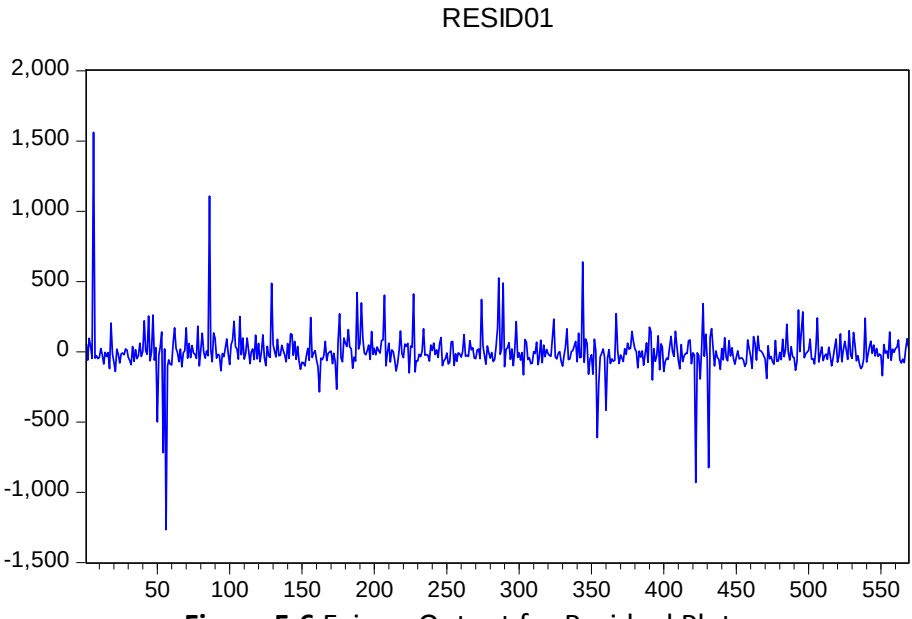


Figure 5.6 Eviews Output for Residual Plot

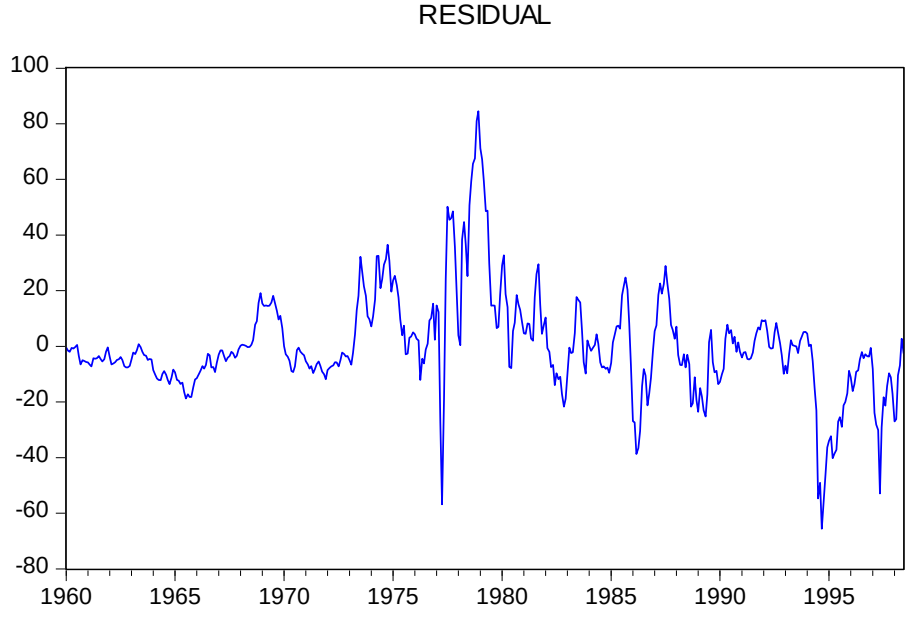


Figure 5.7 Eviews Output for Residual Plot

ii- The Breusch-Godfrey LM-test

Breusch(1978) and Godfrey (1978) developed this autocorrelation test. The idea behind the Breusch-Godfrey test is as follows:

Suppose that the following linear model is estimated with ***OLS.***

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 Z_t + u_t \qquad (5.17)$$

Next residuals are computed and following equation is estimated with ***OLS***.

$$u_t = \varphi_1 u_{t-1} + \ldots + \varphi_p u_{t-p} + \beta_1^* + \beta_2^* X_t + \beta_3^* Z_t + \varepsilon_t \qquad (5.18)$$

Null hypothesis $\boldsymbol{H_0: \varphi_1 = \varphi_2 = \cdots = \varphi_p = 0}$ (no autocorrelation)

In Eviews ***BG*** test is called the serial correlation ***LM*** test. Eviews computes and ***F-statistics*** to test that all the $\boldsymbol{\varphi}$ are zero.

Example 5.11

Table 5.8 *Eviews Output the Breusch-Godfrey LM-Test*

Breusch-Godfrey Serial Correlation LM Test:			
F-statistic	*1.814271*	*Prob. F(3,22)*	*0.1740*
*Obs*R-squared*	*5.949988*	*Prob. Chi-Square(3)*	*0.1141*

Test Equation:
Dependent Variable: RESID
Method: Least Squares
Included observations: 30
Presample missing value lagged residuals set to zero.

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
PRICE	*-0.328497*	*0.848957*	*-0.386942*	*0.7025*
TEMP	*-0.000742*	*0.000829*	*-0.894623*	*0.3807*
TIME	*0.001529*	*0.001950*	*0.784184*	*0.4413*
INCOME	*-0.002193*	*0.002959*	*-0.741184*	*0.4664*
C	*0.290480*	*0.381957*	*0.760505*	*0.4550*
RESID(-1)	*0.516004*	*0.239559*	*2.153970*	*0.0425*
RESID(-2)	*0.064238*	*0.325189*	*0.197540*	*0.8452*
RESID(-3)	*0.176768*	*0.310595*	*0.569127*	*0.5750*

R-squared	*0.198333*	*Mean dependent var*	*4.67E-17*
Adjusted R-squared	*-0.056743*	*S.D. dependent var*	*0.034537*
S.E. of regression	*0.035503*	*Akaike info criterion*	*-3.615225*
Sum squared resid	*0.027730*	*Schwarz criterion*	*-3.241572*
Log likelihood	*62.22838*	*Hannan-Quinn criter.*	*-3.495690*
F-statistic	*0.777545*	*Durbin-Watson stat*	*1.571966*
Prob(F-statistic)	*0.612620*		

iii- The Box-Pierce and the Ljung-Box tests (Q test) (1970)

The ***Box-Pierce*** and the ***Ljung-Box*** test have asymptotic χ^2 distribution, with p degrees of freedom under the null hypothesis of no autocorrelation. This test uses autocorrelation of the residuals. The estimated autocorrelation coefficients are defined as $\hat{\rho}_i$.

$$\hat{\rho}_i = \frac{cov(u_t, u_{t-i})}{\sqrt{var(u_t)}.\sqrt{var(u_{t-i})}} \qquad (5.19)$$

The theoretical autocorrelation coefficients $\widehat{\boldsymbol{\rho}}_{\boldsymbol{i}}$ are zero under the null hypothesis. The ***Q-test*** does not look at the individual autocorrelation coefficients. It considers the sum of a number of squared autocorrelation coefficients as follows:

$$Q = n\sum_{i=1}^{p}\hat{\rho}_i^2 \qquad (5.20)$$

but this test has low power of detecting autocorrelation. The main difference between ***Q-test*** and **BG test** is that one specific order of the autoregressive process specified should be chosen under the alternative hypothesis.

Example 5.12[19]

Table 5.9 *Eviews Output for the Box-Pierce and the Ljung-Box Tests (Q test)*

Included observations: 30

Autocorrelation	*Partial Correlation*		*AC*	*PAC*	*Q-Stat*	*Prob*
. |*** |	. |*** |	1	0.367	0.367	4.4564	0.035
. |* . |	. | . |	2	0.104	-0.036	4.8264	0.090
. | . |	. | . |	3	0.013	-0.015	4.8326	0.184
. *| . |	. *| . |	4	-0.134	-0.150	5.4918	0.240
***| . |	.**| . |	5	-0.370	-0.318	10.759	0.056
.**| . |	. | . |	6	-0.241	-0.003	13.074	0.042
.**| . |	.**| . |	7	-0.281	-0.226	16.366	0.022
. *| . |	. | . |	8	-0.171	-0.021	17.635	0.024
. *| . |	. *| . |	9	-0.094	-0.115	18.043	0.035
. *| . |	.**| . |	10	-0.171	-0.331	19.452	0.035
. | . |	. | . |	11	-0.051	-0.038	19.585	0.051
. |* . |	. | . |	12	0.206	0.059	21.845	0.039
. |*** |	. |**. |	13	0.365	0.255	29.365	0.006
. |* . |	. *| . |	14	0.199	-0.133	31.745	0.004
. |* . |	. *| . |	15	0.105	-0.204	32.455	0.006
. |* . |	. |* . |	16	0.129	0.093	33.601	0.006

[19] Vogelvang,B. (2005)

iv- The Durbin Watson test (1950)

This is the most known autocorrelation test. This test is used for first order autocorrelation.

DW test assumes that error term is stationary and normally distributed with zero mean. It tests the null hypothesis ***H_0*** that the errors are uncorrelated.

Test the H_0: no autocorrelation versus H_1:first order residual autocorrelation.

This test can be used if the explanatory variables are exogenous and a constant term has been included in the model. ***DW –statistics*** is not used, if lagged dependent variables are present as explanatory variables in the model. It can be employed if the explanatory variables are exogenous and the model includes intercept. ***DW-statistics*** should be used if all the conditions are satisfied. Otherwise it is more informative to use **Breusch-Godfrey test** in the research paper.

DW –statistic has been defined as:

$$DW = \frac{\sum(u_t - u_{t-1})^2}{\sum u_t^2} = 2(1 - \varphi_1) \qquad (5.21)$$

Properties of the DW-statistics;

$$\varphi_1 = 0\,, DW \approx 2 \;\; No\ residual\ autocorrelation$$
$$\varphi_1 > 0\,, DW < 2 \;\; Positive\ residual\ autocorrelation$$
$$\varphi_1 < 0\,, DW > 2 \;\; Negative\ residual\ autocorrelation$$

DW-statistics cannot be tabulated. But it is possible to drive distributions of a lower (d_L) and an upper (d_U) bound. These two distributions depend on the number of the observations ***(n)*** and the number of the explanatory variables ***(K)***. Therefore ***DW–statistics*** is tabulated as follows:

$$If\ DW \geq d_U: do\ not\ reject\ H_0$$
$$If\ d_L < DW < d_U: the\ test\ is\ inconclusive$$
$$If\ DW \leq d_L: reject\ H_0\ for\ the\ favor\ of\ first\ order\ residual\ autocorrelation$$

Table 5.10 Lower and Upper Bounds for 5% Critical Values of the DW Test[20]

	Number of regressors including intercept							
Number of Observations	K=3		K=5		K=7		K=9	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
25	1.206	1.550	1.038	1.767	0.868	2.012	0.702	2.280
50	1.462	1.628	1.378	1.721	1.291	1.822	1.201	1.930
75	1.571	1.680	1.515	1.739	1.458	1.801	1.399	1.867
100	1.634	1.715	1.592	1.758	1.550	1.803	1.506	1.850
200	1.748	1.789	1.728	1.810	1.707	1.831	1.686	1.852

If there is autocorrelation. ***OLS*** is no longer ***BLUE***, then **EGLS** (Cochrane-Orcutt) method can be used.

DW test for autocorrelation has some shortcomings such as;

a) The form of model (explanatory variables) should be known

b) The test result is sometimes inconclusive

Example 5.13 Ice-cream

Table 5.11 Eviews Output for the Durbin Watson Test

Dependent Variable: CONS
Method: Least Squares
Sample: 1 30
Included observations: 30

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
INCOME	*0.001890*	*0.002340*	*0.807659*	*0.4269*
PRICE	*-1.104229*	*0.846905*	*-1.303841*	*0.2042*
TEMP	*0.003341*	*0.000480*	*6.961297*	*0.0000*
TIME	*0.001099*	*0.001565*	*0.702250*	*0.4890*
C	*0.322449*	*0.325914*	*0.989367*	*0.3320*

R-squared	*0.724430*	*Mean dependent var*	*0.359433*
Adjusted R-squared	*0.680339*	*S.D. dependent var*	*0.065791*
S.E. of regression	*0.037197*	*Akaike info criterion*	*-3.594163*
Sum squared resid	*0.034590*	*Schwarz criterion*	*-3.360630*
Log likelihood	*58.91245*	*Hannan-Quinn criter.*	*-3.519454*
F-statistic	*16.43025*	*Durbin-Watson stat*	*0.947014*
Prob(F-statistic)	*0.000001*		

[20] Savin, N.E. & White, K.J. (1977)

The Ways to Correct Autocorrelation Problem

An autocorrelation is an indicator of model misspecification. If the model misspecification occurs then change the model not the estimation method (from OLS to EGLS). There are three types of model misspecifications such as:

- Omitted variable bias- excluding relevant independent variables can cause autocorrelation. Find out the relevant omitted variable and include in the model
- Functional form; to use log transformation can eliminate autocorrelation problem
- Dynamic misspecification; whether the model is static or dynamic should be decided. Inclusion of the lagged endogenous and exogenous variables eliminate the autocorrelation problem.

MULTICOLLINEARITY

Multicollinearity is a data problem. Collinearity between variables is always present. It becomes a problem and violation of the classical assumptions if the correlations among the independent variables are very strong. It can affect accuracy of the parameter estimates. Multicollinearity misleadingly increases the standard errors. Thus, it makes some variables statistically insignificant while they should be otherwise significant. It is like two or more people singing loudly at the same time. One cannot discern who is singing how. The influence of the independents offset each other. In multicollinearity, the parameter estimates become inaccurate.

To use too many dummy variables may also cause multicollinearity.

Consequences

The precision of estimation falls because of very big errors that may be highly correlated errors and very large sampling variances of the coefficients. Large standard errors can be caused by things besides multicollinearity. If two independent variables are highly and

positively correlated, then their slope coefficient estimators tend to be highly and negatively correlated. If the effect of one parameter is overestimated, then the effect of the other parameter is underestimated. Hence, coefficient estimates tend to vary from one sample to the next.

When high multicollinearity is present, confidence intervals for coefficients tend to be very wide and ***t-statistics*** tend to be very small. Coefficients have to be larger in order to be statistically significant. It is very difficult to reject the null hypothesis when multicollinearity is present.

If the objective is only to predict endogenous variable from a set of exogenous variables, then multicollinearity is not a problem. The predictions can be accurate, and the overall $\mathbf{R^2}$ (or adjusted $\mathbf{R^2}$) quantifies how well the model predicts the ***y*** values.

If the objective is to understand how the various ***x*** variables impact *y* over observations, then multicollinearity can be a big problem. One problem is that the individual ***p*** values can be misleading (a ***p*** value can be high, even though the variable is important). The second problem is that the confidence intervals on the regression coefficients can be very wide. The confidence intervals may even include zero, which means you can't even be confident whether an increase in the ***x*** value is associated with an increase, or a decrease, in ***y***. Because the confidence intervals are so wide, excluding a subject (or adding a new one) can change the coefficients and may even change their signs.

Causes of Multicollinearity

i- Improper use of dummy variables

ii- Including a variable computed from other variables in the equation (e.g. family income = husband's income + wife's income, and the regression includes all three income measures)

iii- Including the same or almost the same variable twice (height in feet and height in inches; or, more commonly, two different operationalization of the same identical concept)

Detection

- Variance inflation factors (VIF) is used to measure how much the variance of the estimated coefficients is increased over the case of no correlation among the independent variables.

    - If VIF = 0, No multicollinearity
    - If VIF ≥ 0, There is multicollinearity

- Regression coefficients change drastically while adding or deleting an ***x*** variable.
- A regression coefficient is negative when theoretically endogenous variable increase with increasing values of that exogenous variable, or the regression coefficient is positive when theoretically endogenous variable decrease with increasing values of that exogenous variable.
- Collinearity diagnostics less than 0.2 indicates high collinearity.
- None of the individual coefficients has a significant ***t-statistic***, but the overall ***F-test*** is significant.
- A regression coefficient has a non-significant t-statistic, even though on theoretical grounds that exogenous variable should provide substantial information about endogenous variable.
- High R^2 (larger than .75) but only a few significant ***t-values***.
- High correlation coefficients between estimated values in the correlation matrix indicate possible multicollinnearity (≥ .80).
- Large standard errors and low ***t-statistic*** values
- Wrong signs of parameter estimates
- Whether the coefficients are sTable should be checked when different samples are used. Sample period may be divided into subsamples randomly. If the coefficients differ dramatically, multicollinearity may be a problem.
- Using the same data do a slightly different specification of a model. Observe the changes.

- In particular, as variables are added, changes in the signs of effects should be considered (e.g. switches from positive to negative) which seems theoretically questionable. If there are suppressor effects, such changes may be logical, otherwise they may indicate multicollinearity.

The Ways to Handle Multicollinearity

- The sample size can be increased since when sample size is increased, standard error decreases (when all other things are equal). This mitigates the high multicollinearity problem which causes high standard errors of coefficients.
- The highly inter-correlated variable(s) may be omitted from analysis (generally if larger than 0.80). This method is misguided if the variables were there due to the theory of the model, and they should have been. Remove each variable individually and observe R^2. The model with higher R^2 should be chosen.
- Combine variables into a composite variable through building indexes (creating an index theoretical and empirical reason to justify this action).
- Transform the offending independent variables by subtracting the mean from each case. The resulting centered data may well display considerably lower multicollinearity. Interpretations of ***b*** and ***beta*** must be changed accordingly.
- Information from prior researches can be used
- The model may be re specified. Intercorrelated variables may be omitted from analysis but substitute their cross product as an interaction term or in some other way combine the intercorrelated variables. If a correlated variable is a dummy variable, other dummies in that set should also be included in the combined variable in order to keep the set of dummies conceptually together. To do this, compute the mean of each independent variable, and then replace each value with the difference between it and the mean. For example, if the variable is weight and the mean is

72, then enter "6" for a weight of 78 and "-3" for a weight of 69.

- Assign the common variance to each of the covariates by some probably arbitrary procedure. Treat the common variance as a separate variable and decontaminate each covariate by regressing them on the others and using the residuals. That is, analyze the common variance as a separate variable.

Example 5.14

Table 5.12 SPSS Outputs for Variance Inflation Factor

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	Nbedroom	.768	1.301
	Nbath	.825	1.212
	Nstories	.799	1.251

a. Dependent Variable: lot size

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	Nbedroom	.854	1.171
	Nbath	.841	1.189
	lot size	.955	1.047

a. Dependent Variable: Nstories

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	Nbedroom	.820	1.220
	lot size	.976	1.024
	Nstories	.833	1.200

a. Dependent Variable: Nbath

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	lot size	.962	1.040
	Nstories	.895	1.118
	Nbath	.867	1.153

a. Dependent Variable: Nbedroom

EXOGENEITY

There are two types of variables in macro-econometric models: endogenous and exogenous. Endogenous variables are explained by the equations, either the stochastic or the identities. Exogenous variables are determined outside of the model not explained within the model. They are taken as given.

The concept of exogeneity has been analyzed and elaborated by Engle, Hendry and Richard (1983). All explanatory variables should be uncorrelated with the error term. It means that explanatory variables have to be determined outside of the model (they are exogenous).

$$E(u_i|x_{i1}, x_{i2}, \ldots., x_{ik}) = 0 \qquad (5.22)$$

Suppose we have the following model, where $\boldsymbol{x_i}$ *and* $\boldsymbol{u_i}$ are positively correlated.

$$y_i = \beta_0 + \beta_1 x_i + u_i \qquad (5.23)$$

If we are trying to find out the relationship between the price of a sish kebab and the quantity sold across a wide variety of Turkish restaurants.

$$Quantity = \beta_0 + \beta_1 Price + u_i\ , \qquad (5.24)$$

Together with the price, it is obvious that the quality has also significant impact on sales quantity. Also, there is very strong positive relationship between price of sish kebab and the quality. Therefore $\boldsymbol{x_i}$ *and* $\boldsymbol{u_i}$ are highly correlated.

NORMALITY

Normality test was introduced by Fisher (1948), but, a standard Jarque-Bera (JB, 1980) test is being used widely. It is expressed in terms of the third and fourth moments of the disturbances, as follows:

The null hypothesis is H_0 and the residuals are normally distributed.

$$JB = (n-K)\left(\frac{S^2}{6} + \frac{(K-3)^2}{24}\right) \quad (5.25)$$

S: measure of skewness

$$S = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{(\mu_2)^{3/2}} \quad (5.26)$$

K: measure of Kurtosis

$$K = \frac{\mu_4}{\sigma^4} = \frac{\mu_4}{(\mu_2)^2} \quad (5.27)$$

The values for ***K*** and ***S*** when the variable is normally distributed: K=3 and S=0

Example 5.15

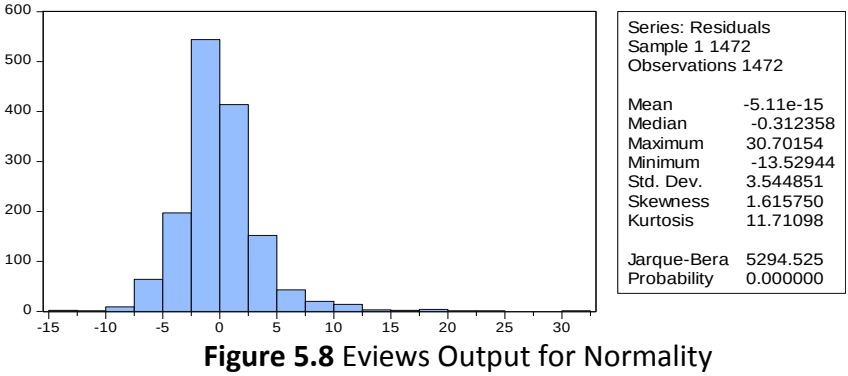


Figure 5.8 Eviews Output for Normality

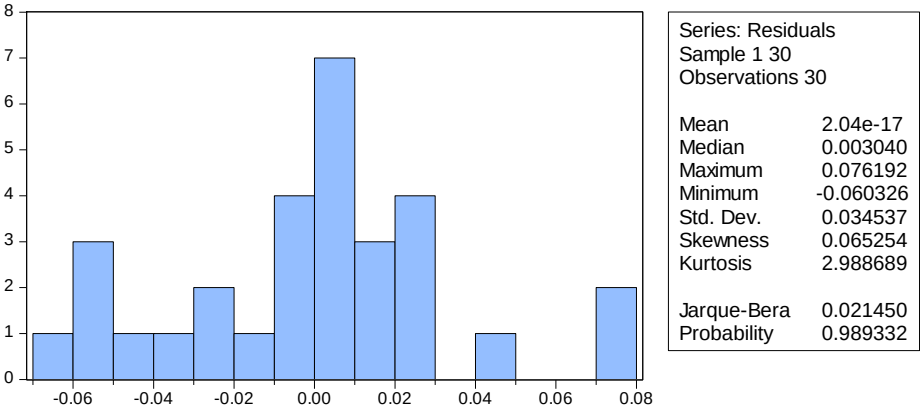


Figure 5.9 Eviews Output for Normality

ii- QQ and PP plots

Plotting the empirical distribution of residuals against the normal counterparts can reveal normality problem.

Example 5.16

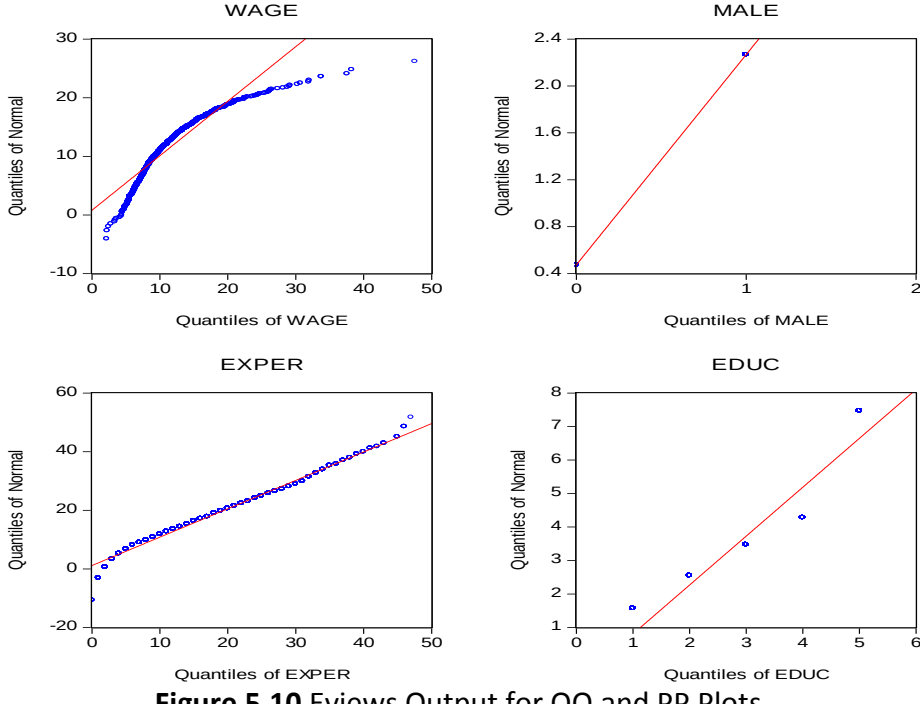


Figure 5.10 Eviews Output for QQ and PP Plots

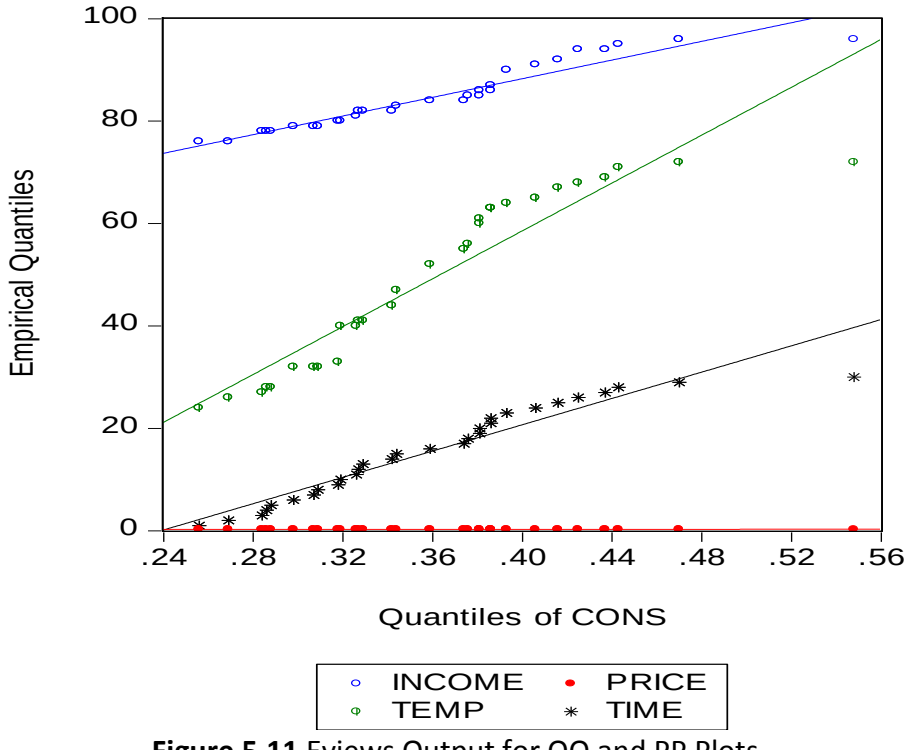


Figure 5.11 Eviews Output for QQ and PP Plots

If normality is rejected, the test gives nothing about how to correct this deviation from normality. The best way is to look at the plot of the residuals. It can be caused by outliers in the residuals. These outliers can be eliminated through using a dummy variable in the regression equation.

What to do if data is not normal:

i- Increase the sample size

ii- Transform data (log values). Misspecification and functional form can spoil normality assumptions.

Example 5.17

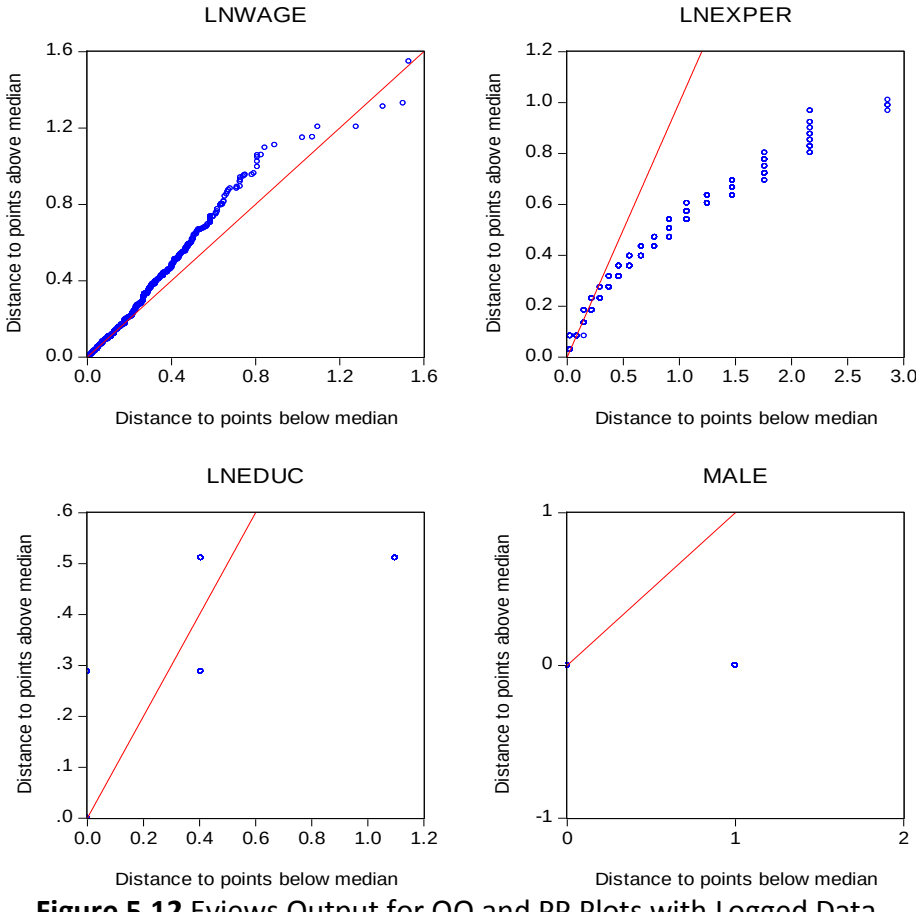


Figure 5.12 Eviews Output for QQ and PP Plots with Logged Data

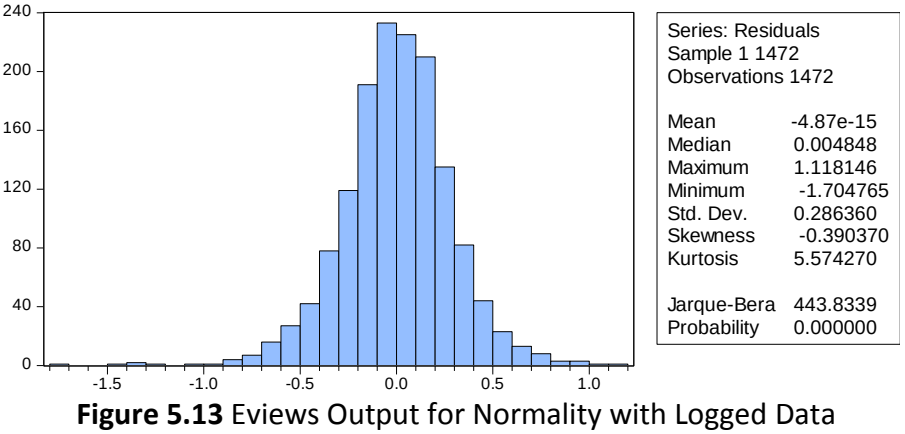


Figure 5.13 Eviews Output for Normality with Logged Data

CONSISTENCY

Consistency is a large sample property. If there are more observations, the probability that the estimator is some positive number far from the true value becomes smaller. A minimum requirement for an estimator to be useful appears to be that it is consistent. OLS estimator is consistent if it holds that:

$$\frac{1}{N}\sum_{i=1}^{N} x_i \dot{x}_i \tag{5.28}$$

The variance of b decreases as the sample size increases.

$$E\{x_i \varepsilon_i\} = 0 \tag{5.29}$$

It says that the error term is mean zero and uncorrelated with any of the explanatory, independent variables. The consistency property implies that the estimator $\widehat{\boldsymbol{\beta}}$ itself does not have an asymptotic distribution. Sufficient condition for a consistent estimator is that the bias and the variance should both tend to be zero as the sample size increases. The specification of the linear regression model is as follows:

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + u_i \tag{5.29}$$

Deterministic assumptions:

- **i-** $\boldsymbol{y_i}$ is endogenous, dependent variable
- **ii-** $\boldsymbol{y_i}$ is explained by the ***K*** linearly independent exogenous variables
- **iii-** The first exogenous variable equals to one. This is constant term included in the model.
- **iv-** The difference between the number of observations (***N***) and the number of explanatory variables (***K***) gives degrees of freedom (***N-K***) of the model.
- **v-** Exogenous variables (explanatory) are linearly independent and form the systematic part of the model.

CHAPTER 6

HYPOTHESIS TESTING

Its objective is to determine whether or not sample data support a belief or hypothesis about the population(s) from the sample(s) which is drawn. For example, one may want to know if a proposed policy is equally supported by men and women.

DEFINITIONS

- **Statistical significance**: Is a result that is not likely to occur randomly, but rather is likely to be attribuTable to a specific cause.

- **Null hypothesis:** denoted by H_0 and always asserts that no difference exists between a population parameter and a single value.

- **Alternative hypothesis:** is denoted by H_1 and assert that there is difference between a population parameter and a single value.

- **Test Statistic:** a tool employed by the researcher in order to make decision (reject or not reject) on the null hypothesis.

- **Degrees of freedom:** refer to how free one is to set the value of a variable once an observation has been drawn. Drawing one observation removes one degree of freedom (to set a value, ***N - 1***, where ***N*** is the sample size). ***N – 1*** is the number of degrees of freedom for the ***t-test***.[21]

- **Confidence level**: the percentage of times one would expect that the sample represents the population; it is commonly set at 0.90, 0.95, or 0.99.

- **Significance level**: the probability of rejecting a null hypothesis when it is true. It is denoted by the ″α″ and is commonly set at 0.10, 0.05, or 0.01.[22] Before running any statistical test, significance level should be defined first.

- **Confidence interval:** refers to the interval of values around the test statistic in which the null hypothesis is not rejected. It is the probability of making wrong decision. Confidence intervals are generally notated as follows:

$[b_k - 1.96se(b_k),\ \ b_k + 1.96se(b_k)]$, 95% confidence interval (6.1)

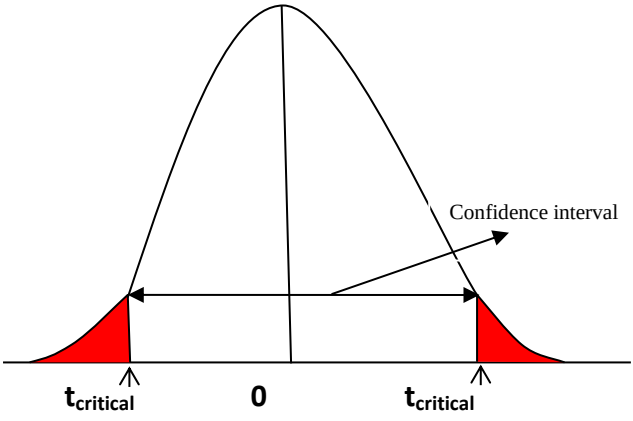


Figure 6.1 Confidence Interval two Sided

[21] the ***z-test*** does not involve degrees of freedom since it assumes that the population variance is known

[22] significance level + confidence level = 1

One-sided hypothesis test puts all of the significance level "at one end of a distribution.

One-sided test: $H_0: b_k \leq \beta_k^0, \quad H_1: b_k > \beta_k^0$

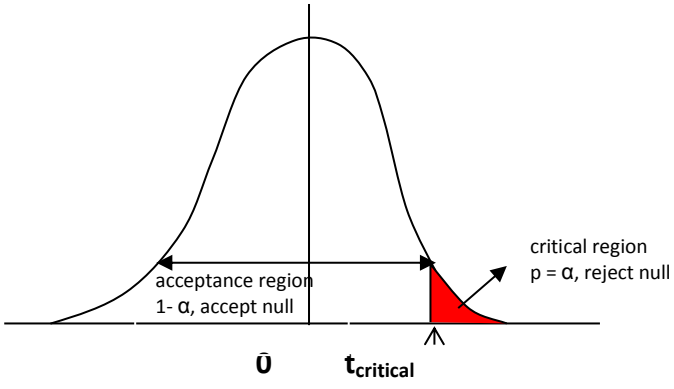


Figure 6.2 Confidence Interval one Sided

Two-sided hypothesis test: Suppose that we have null hypothesis, $\boldsymbol{H_0: b_k = \beta_k^0}$, where b_k is the estimated value and $\boldsymbol{\beta_k^0}$ is the value[i] chosen by researcher . If null hypothesis is not true following alternative hypothesis holds:

$H_1: b_k \neq \beta_k^0$

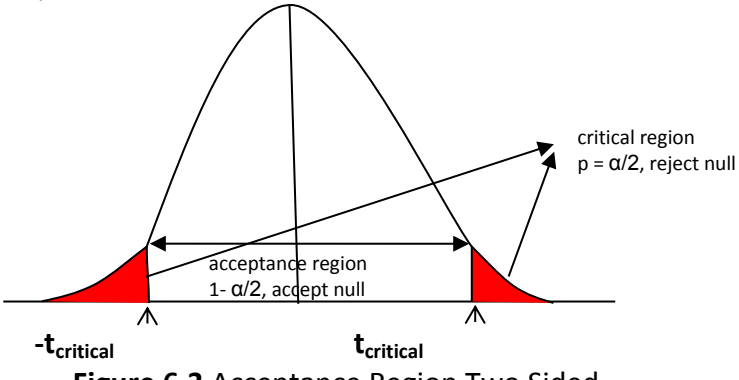


Figure 6.3 Acceptance Region Two Sided

The **p-value** is the smallest value that would lead to rejection of the null hypothesis. It is also called the significance probability. Generally, the smaller the ***p-value***, the more likely is it that one can reject the null hypothesis. A ***p–value*** of 0.003 means that there is only a 3 in a thousand chance of reaching the wrong conclusion. The

smaller the ***p- value***, the stronger the evidence against the null hypothesis.

$$p - value = 2\Phi(-|t_k|) \quad (6.2)$$

Generally, following conventions are applied:

If p-value > .10 → the observed difference is "***not significant***"
if p-value ≤ .10 → the observed difference is "***marginally significant***"
if p-value ≤ .05 → the observed difference is "***significant***"
if p-value ≤ .01 → the observed difference is "***highly significant***"

However, for hypothesis testing to work properly, data should be relatively normally distributed. That is, the distribution of data should be:

- mound-shaped
- fairly symmetrical
- free of significant skew and kurtosis

The less these conditions hold, the less reliable are tests of hypothesis and, thereby, the conclusions become less reliable. The ordered steps of hypothesis testing for one sample are as follow:

i- Calculate the mean and standard error (***S_e***) of the sample

ii- Define the hypothesized value to be tested

iii- Define the significance level

iv- Construct the null and alternative hypothesis

v- Calculate the standard error of the mean

$$S_M = \frac{S}{\sqrt{N}} \quad (6.3)$$

vi- Calculate the t statistics

$$t = \frac{M - \mu_0}{S_M} \quad (6.4)$$

vii- With ***N-1*** degree of freedom, specify the critical value form the t distribution Table

viii- Reject null hypothesis
if $|t| > critical\ value$ or
if probability value, p < significance level

ix- Interpret the result based on the research questions

Ultimately, only two possible conclusions can be drawn for a hypothesis test: Reject the null hypothesis, in which case one concludes that there is sufficient evidence to conclude that the alternative hypothesis is true.

Do not reject the null hypothesis, in which case one concludes that there is insufficient evidence to indicate that the alternative hypothesis is true.

Test statistic is used to determine whether the researcher should reject or not reject the null hypothesis.[23]

$$t_k = \frac{b_k - \beta_k^0}{se(b_k)} \qquad (6.5)$$

Where $\boldsymbol{se(b_k)}$ is the standard error of $\boldsymbol{b_k}$.

It follows t distribution with ***N-K*** degrees of freedom. To be precise we can reject the null hypothesis if ***t-test*** result is larger than critical values which are summarized in Table 6.1.

Example 6.1

We aim to measure relative impact of a diploma on the beginning salary. Our sample includes graduates from the different universities with a bachelor degree having a mean annual starting salary of $10,450. Faculty of Economics graduates from eighty universities, are randomly selected. Their starting salaries have a mean of $12,050. If

[23] The term accept should not be used in the context of hypothesis testing.

the standard deviation is $950, use a 0.01 level of significance to test the claim that University graduates with an economics degree have a mean starting salary that is greater than the mean for graduates with a bachelor degree.

Information given in the question is following:
μ=10,450
n=80
$\overline{x}$=12,050
S=950
A=0.01

Step one: H_0: μ=10,450
H_1: μ>10,450

Step two: parametric →one group of samples → σ unknown but s is known

Therefore, we use t test with the formula

$$t = (\bar{x} - \mu)/(s/\sqrt{n})$$

Calculate the test statistics t value:

$$t = (\bar{x} - \mu)/(s/\sqrt{n}) = (12{,}050 - 10{,}450)/(950/\sqrt{80}) = 15.06$$

Step three: Identify the critical value or ***p-value.***
We find the critical t value 2.66 by
$df = n - 1 = 80 - 1 = 79, \alpha = 0.01$
$in\ t\ distribution\ table$
Or, we find p-value is less than 0.0001 by using the test statistics t=15.068,
one-tailed, in Standard Normal Distribution Table.

Step four: Draw a graph included the test statistics value, the critical value and the critical region(s) or compare the P-value with the significance level α. And then make a conclusion of the hypothesis.

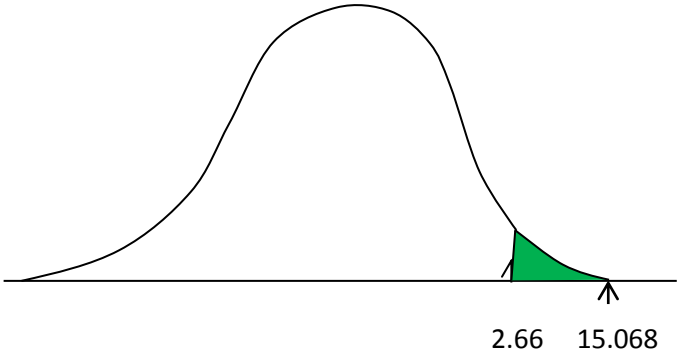


Figure 6.4 Critical Region

Step five: the value of t-statistics is highly significant. It means that there is enough evidence to reject the null hypothesis and the mean of our sample equals to the population mean.

If probability of obtaining ***t value*** is less than chosen probability value the significance levels (10%, 5%, or 1%) we can reject null hypothesis. We conclude that estimated ***t value*** is statistically significant or significantly different from zero.

Table 6.1 Critical Values for One Sided and Two Sided t-Test (for infinite sample)

Significance Levels (%)	Critical values	
	One-sided	two-sided
1	2.58	$\pm$2.58
5	1.96	$\pm$1.96
10	1.64	$\pm$1.64

Example 6.2

If the impact of income level on the demand for ice-cream is examined:

First step: define a regression equation based on the related economic theories, such as:

$$Cons_i = \beta_0 + \beta_1 Price_i + \beta_2 Income_i + \beta_3 Temp_i + e_i$$

Second step: regress consumption on Price, Income and Temperature using OLS, WLS or other estimation methods

$$Cons_i = 0{,}197 + (-1.044) Price_i + 0.0033 Income_i + 0.0034 Temp_i$$

Third step: Define the significance level (10%, 5% or 1%)

Let's define the significance level as 5%.

Fourth step: compute t-statistics

$$Cons_i = 0{,}197 + (-1.044) Price_i + 0.0033 Income_i + 0.0034 Temp_i$$

	(0.8343)	(0.0011)	(0.0004)
t-values	-1.25	2.82	7.76

$$t = (-1.044 - 0) \div 0.8343 = -1.25$$
$$t = (0.0033 - 0) \div 0.0011 = 2.82$$
$$t = (0.0034 - 0) \div 0.0004 = 7.76$$

Fifth step: Interpret the results

Reported t-values above indicate that there is enough evidence to conclude that income and temperature have positive significant impact on the ice-cream consumption at the 5% significance level. Low t- value of the estimates of the price coefficient (-1.25) indicates that there is not enough evidence to reject null hypothesis that no relationship between price and consumption exists.

F-Test: A JOINT TEST OF SIGNIFICANCE

It is possible to test more than one restriction simultaneously.

$$H_0: \beta_{K-J+1} = \cdots = \beta_K = 0 \tag{6.6}$$

Alternative hypothesis is that H_0 is not true (at least one of coefficients is not equal to zero).

$$H_0: \beta_2 = \beta_3 = \beta_4 = \cdots = \beta_K = 0 \tag{6.7}$$

It implies that there is no linear relationship between dependent and independent (explanatory) variables.

$$f = \frac{(S_0 - S_1)/J}{S_1/(N-K)} \tag{6.8}$$

S_1: residual sum of squares of the full model,
S_0: residual sum of squares of the restricted model
N: Sample size,
K: number of variables

Large values for the test statistics imply rejection of null hypothesis. Critical values for the F-test are attached.

If we use the definition of R^2, f statistic can be written as:

$$f = \frac{(R_1^2 - R_0^2)/J}{(1-R_1^2)/(N-K)} \tag{6.9}$$

Where, R_1^2 *and* R_0^2 are the goodness-of-fit measures for the unrestricted and the restricted model respectively.

Example 6.3

At first we regress price of coffee on price on cocoa and report the results in the Table 6.2

Table 6.2 Eviews Output

Dependent Variable: PCOFFEE
Included observations: 462

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
PCOCOA	*1.183739*	*0.030164*	*39.24323*	*0.0000*
C	*12.33664*	*2.307020*	*5.347436*	*0.0000*

R-squared	*0.770004*	*Mean dependent var*	*89.77775*
Adjusted R-squared	*0.769504*	*S.D. dependent var*	*53.50328*
S.E. of regression	*25.68695*	*Akaike info criterion*	*9.334163*
Sum squared resid	*303516.8*	*Schwarz criterion*	*9.352065*
Log likelihood	*-2154.192*	*Hannan-Quinn criter.*	*9.341211*
F-statistic	*1540.031*	*Durbin-Watson stat*	*0.166605*
Prob(F-statistic)	*0.000000*		

In the second step we add price of tea as a regressor and report the results in Table 6.3

Table 6.3 Eviews Output

Dependent Variable: PCOFFEE
Included observations: 462

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
PCOCOA	*1.022619*	*0.039617*	*25.81265*	*0.0000*
PTEA	*0.016525*	*0.002759*	*5.989058*	*0.0000*
C	*-5.431432*	*3.707959*	*-1.464803*	*0.1437*

R-squared	*0.786674*	*Mean dependent var*	*89.77775*
Adjusted R-squared	*0.785745*	*S.D. dependent var*	*53.50328*
S.E. of regression	*24.76546*	*Akaike info criterion*	*9.263249*
Sum squared resid	*281517.5*	*Schwarz criterion*	*9.290103*
Log likelihood	*-2136.811*	*Hannan-Quinn criter.*	*9.273822*
F-statistic	*846.3184*	*Durbin-Watson stat*	*0.161762*
Prob(F-statistic)	*0.000000*		

In the third step calculate ***F-value*** as follows:

$$F = \frac{(0.786 - 0.770)/1}{(1 - 0.786)/(462 - 3)} = 34.320$$

In the fourth step we interpret the results.

The significant F-value reported above indicates that adding the regressors has significant contribution to the model.

Restricted model has R_0^2 of zero by construction; its test statistic can also be written using the relationship between F and R^2:

$$F = \frac{(R^2/(K-1)}{(1-R^2)/(N-K)} \tag{6.10}$$

If the test statistics based on F does not reject the null hypothesis, we can conclude that the model performs rather poorly, a model with only one intercept term would not be statistically worse. If the test statistics reject the null hypothesis, we cannot conclude that the model is good.

If the test statistic is enough to reject null hypothesis, we can conclude that the model is good. It is possible to observe contradiction between t-test and F-test results. While t-test do not reject null hypothesis, F-test can reject or vice versa.

Example 6.4[24]:

$$F = \frac{0.786/(3-1)}{(1-0.786)/(462-3)} = 842.98$$

The result indicates that reported R^2 is significantly different than zero.

[24] Verbeek, M. (2004)

CHAPTER 7

SPECIFICATIONS

Before estimating a regression model, model specification should be checked. Specification errors in the regression model estimation may result series problem with the robustness and reliability of the estimation. Specification checking is constructed under three subtitles which are choosing independent variables, functional form, and stochastic term.

CHOOSING INDEPENDENT VARIABLES

The researchers define the independent variables based on the economic theory, experience, common sense, critical thinking and research objectives. The independent variables which are chosen should satisfy the estimation assumptions. Any departure from those assumptions may cause unreliable and wrong findings.

If the economic theory requires including one or more independent variables in the regression equation, then they should be definitely included in the model. Leaving a related variable outside of the model can cause bias in the estimation, and including irrelevant variables in the regression equation leads to higher variances of the estimation. It is mostly difficult to decide whether to include or exclude an independent variable in the regression equation, if it is not included in the economic theory.

Omitting a Relevant Variable

Data unavailability and researchers ignorance may cause a regression equation without relevant independent variable. Omitted variables bias is the bias resulting from leaving a relevant independent variable out of the regression equation. In a multiple regression equation the coefficient $\boldsymbol{\beta_k}$, represents the amount of changes in the dependent variable when independent variable $\boldsymbol{x_k}$ change one unit holding constant the other independent variables in the equation. If a relevant variable is excluded, it is not held constant for the calculation and interpretation of coefficient $\boldsymbol{\beta_k}$. Because of this, omission of a relevant independent variable may cause bias pushing the expected value of the estimated coefficient out from the true value of the population coefficient.

Assume that the independent variable $\boldsymbol{z_i}$ is omitted from the regression equation.

$$y_i = \beta_0 + \beta_1 x_i + e_i \tag{7.1}$$

The error term includes the omitted independent variable as follows:

$$e_i = \beta_2 z_i + \varepsilon_i, \tag{7.2}$$

then Assumption 2 (unbiasedness) does not hold.

$$E(e_i) = \beta_2 z_i \neq 0. \tag{7.3}$$

Also, if $\boldsymbol{x_i}$ and $\boldsymbol{z_i}$ are correlated, Assumption 3 (efficiency) does not hold.

$$cov(e_i, x_i) \neq 0. \tag{7.4}$$

Gauss- Markov theorem is not valid, and ***OLS*** is not ***BLUE***, estimate of regression coefficient is biased. ***OLS*** overestimate the coefficient of independent variable $\boldsymbol{\beta_1}$.

$$E(\hat{\beta}_1^*) \neq \beta_1 \tag{7.5}$$

Omission of a closely related independent variable implies that ***OLS*** is no longer ***BLUE***.

If a relevant variable is omitted from the regression equation:

i- Estimate of the coefficient of that variable is absent

ii- The coefficients of the included independent variables are likely to be biased

If omitted variable is not correlated with the included independent variable and the coefficient of omitted independent variable is zero, then ***OLS*** is ***BLUE***.

Detection

i- Check the economic theory to identify relevant independent variable

ii- Check the compatibility of the sign of estimated coefficients with the related economic theory

Wald Test (Coefficient Restrictions)

The Wald test (Wald, 1943) statistics is the estimation of unrestricted regression without imposing the coefficient restrictions specified by the null hypothesis. It measures how close the unrestricted estimation satisfies the restrictions under the null hypothesis. The unrestricted estimation satisfies the restrictions if the imposed restrictions are true.

The Wald test statistic is computed as follows:

$$y = X\beta + e, \text{ which is linear model}$$

Linear restrictions are as follows:

$$H_0: R\beta - r = 0 \qquad (7.6)$$

R : q x k matrix,
r : a q vector

Wald statistics:

$$W = (Rb - r)^I (s^2 R(X^I X)^{-1} R^I)^{-1} (Rb - r), \chi^2(q) \quad (7.7)$$

If we assume that errors are independent and identically normally distributed, **F-statistic** can be computed as follows:

$$F = \frac{(\bar{u}^I \bar{u} - u^I u)/q}{u^I u/(N-K)} = W/q \quad (7.8)$$

$\bar{u}$: Residual vector of the restricted regression. ***F-statistic*** does compare the residual sum of squares computed with and without the imposed restrictions. ***F-statistic*** is small if the restrictions are true and there is little difference in two residual sums of squares.

Example 7.1

Estimation of Cobb-Douglas production function (Cobb, Douglas, 1928) using the data provided the following results:

Table 7.1 Eviews Output

Dependent Variable: LOG(OUTPUT)
Method: Least Squares
Sample: 1 569
Included observations: 569

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
C	*-1.711459*	*0.096711*	*-17.69672*	*0.0000*
LOG(CAPITAL)	*0.207570*	*0.017188*	*12.07677*	*0.0000*
LOG(LABOR)	*0.714847*	*0.023142*	*30.89002*	*0.0000*

R-squared	*0.837830*	*Mean dependent var*	*1.674907*
Adjusted R-squared	*0.837257*	*S.D. dependent var*	*1.185050*
S.E. of regression	*0.478067*	*Akaike info criterion*	*1.367125*
Sum squared resid	*129.3580*	*Schwarz criterion*	*1.390028*
Log likelihood	*-385.9472*	*Hannan-Quinn criter.*	*1.376062*
F-statistic	*1462.078*	*Durbin-Watson stat*	*1.963300*
Prob(F-statistic)	*0.000000*		

LOG(OUTPUT) = C(1) + C(2)*LOG(CAPITAL) + C(3)*LOG(LABOR)

LOG(OUTPUT) = -1.71145941868 + 0.207570296302*LOG(CAPITAL) + 0.714846786693*LOG(LABOR)

In order to test hypothesis of constant returns to scale, type the following restriction in the dialog box:

c(2)+c(3) = 1

Eviews reports following output:

There is only one restriction. Two test statistics are identical with the p-values of both statistics indicates that the null hypothesis of constant returns to scale can be rejected.

Table 7.2 Eviews Output for Wald Restriction Test

Wald Test:
Equation: Untitled

Test Statistic	*Value*	*df*	*Probability*
F-statistic	*20.33514*	*(1, 566)*	*0.0000*
Chi-square	*20.33514*	*1*	*0.0000*

Null Hypothesis Summary:

Normalized Restriction (= 0)	*Value*	*Std. Err.*
-1 + C(2) + C(3)	*-0.077583*	*0.017205*

Restrictions are linear in coefficients.

To test more than one restriction, separate the restrictions by commas.

If the hypothesis is the elasticity of output with respect to labor is 1/4 and with respect to capital 3/4, type the following restrictions:

c(2) = 1/4, c(3) = 3/4

Eviews reports following result:
Significant F-statistics and a Chi-square indicate that there is enough evidence to reject null hypotheses defined as restrictions.

Table 7.3 Eviews Output for Wald Restriction Test

Wald Test: *Equation: Untitled*			
Test Statistic	*Value*	*df*	*Probability*
F-statistic	*12.27250*	*(2, 566)*	*0.0000*
Chi-square	*24.54501*	*2*	*0.0000*
Null Hypothesis Summary:			
Normalized Restriction (= 0)		*Value*	*Std. Err.*
-1/4 + C(2)		*-0.042430*	*0.017188*
-3/4 + C(3)		*-0.035153*	*0.023142*

Restrictions are linear in coefficients.

Omitted Variables Test

This test gives options to add one or more variables to an existing equation and to test whether the set makes a significant contribution in explaining the variation in the dependent variable. The null hypothesis in this case is:

$\boldsymbol{H_0}$: the additional variable or variables are not significant

The output from ***F-statistics***, Likelihood ratio statistics, ***p-values*** and the estimation results of the unrestricted model is reported.

The ***F-statistic*** is based on the difference between the residual sums of squares of the restricted and unrestricted regressions.

Number of observation of original and test equation should be equal in order to employ the omitted variables test. It also can be employed if the equation is specified by listing regressors not by

formula. It is available for LS, TSLS, ARCH, binary, ordered, censored, truncated and count methods.

Example 7.2

Table 7.4 Eviews output for Omitted Variables Test Results

Dependent Variable: CONS				
Method: Least Squares				
Included observations: 30				
Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
INCOME	*0.000505*	*0.001988*	*0.253942*	*0.8014*
C	*0.316715*	*0.168665*	*1.877771*	*0.0709*
R-squared	*0.002298*	*Mean dependent var*		*0.359433*
Adjusted R-squared	*-0.033334*	*S.D. dependent var*		*0.065791*
S.E. of regression	*0.066878*	*Akaike info criterion*		*-2.507551*
Sum squared resid	*0.125235*	*Schwarz criterion*		*-2.414138*
Log likelihood	*39.61326*	*Hannan-Quinn criter.*		*-2.477667*
F-statistic	*0.064487*	*Durbin-Watson stat*		*0.392752*
Prob(F-statistic)	*0.801396*			
Omitted Variables: PRICE TEMP				
F-statistic	*33.15603*	*Prob. F(2,26)*		*0.0000*
Log likelihood ratio	*38.01235*	*Prob. Chi-Square(2)*		*0.0000*
Test Equation:				
Dependent Variable: CONS				
Method: Least Squares				
Included observations: 30				
Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
INCOME	*0.003308*	*0.001171*	*2.823722*	*0.0090*
C	*0.197315*	*0.270216*	*0.730212*	*0.4718*
PRICE	*-1.044414*	*0.834357*	*-1.251759*	*0.2218*
TEMP	*0.003458*	*0.000446*	*7.762213*	*0.0000*
R-squared	*0.718994*	*Mean dependent var*		*0.359433*
Adjusted R-squared	*0.686570*	*S.D. dependent var*		*0.065791*
S.E. of regression	*0.036833*	*Akaike info criterion*		*-3.641296*
Sum squared resid	*0.035273*	*Schwarz criterion*		*-3.454469*
Log likelihood	*58.61944*	*Hannan-Quinn criter.*		*-3.581528*
F-statistic	*22.17489*	*Durbin-Watson stat*		*1.021170*
Prob(F-statistic)	*0.000000*			

Solution

i- Expected bias analysis may be applied. Omitting a variable may have caused it in the estimated coefficient of one of the independent variable in the model. Expected bias can be estimated as follows:

$$Expected\ bias = \beta_{om}.f(r_{in,om}) \quad (7.9)$$

ii- $\boldsymbol{r_{in,om}}$ denotes the correlation coefficient between included and omitted variable in the regression. Sign of expected bias should be checked with the sign of the unexpected result. If they are same, then the variable could be the reason of bias. This analysis should be used if there is obviously a bias.

iii- If the omitted variable is obvious and available, it should be included in the model.

iv- If the omitted variable is not available, a proxy variable should be found which is closely related to the omitted variable and included in the model.

Including an Irrelevant Variable in the Regression equation

Including an irrelevant variable in the regression equation may cause an increase in the variances of the estimated coefficients of included independent variables. It may reduce the precision of the estimation and does not cause bias.

Suppose that an independent variable $\boldsymbol{z_i}$ is not related or it is obvious that it has no relationship with the dependent variable such as:

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad (7.10)$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i \quad (7.11)$$

If $\boldsymbol{z_i}$ is not related, $\boldsymbol{\beta_2}$**= 0**. Hence Assumption 2 and 3 holds, ***OLS*** estimators are not biased and remain ***BLUE***.

Standard errors on the estimated coefficients are larger in the model which includes irrelevant dependent variable than the optimum model. These results are higher than ***t-ratio***.

When it is detected that it is irrelevant, it should be omitted from the regression equation.

How to Choose Correct Variables?

i- Following the Economic theory

ii- Check whether it is significant with correct sign

iii- Has $\overline{R}^2$ improved

iv- Check the other coefficients sign, after the variable is included

If all the conditions above are satisfied, then variable belongs to the regression equation. Following techniques are strongly recommended to be employed for specification bias problem:

i- Scanning to develop a tesTable theory: it is about analyzing a data set for the purpose of developing a tesTable theory or hypothesis. An economic theory or hypothesis should have been tested on a different data set before giving reference to theory or hypothesis.

ii- Sensitivity analysis: it refers to employing different alternative specifications to determine whether the estimation result is robust. How sensitive an estimation result is to a change in different specifications should be examined.

Redundant Variables Test

This test allows testing for statistical significance of a subset of the included variables. It tests whether the coefficients of the variables in a regression equation are zero. If they are equal to zero, they should be omitted from the equation. It is available for ***LS, TSLS***, ***ARCH***,

binary, ordered, censored, truncated and count methods and can be employed if the equation is specified by listing regressors not by formula.

Example 7.3

Table 7.5 Eviews Result for Redundant Variable Test Results

Dependent Variable: PRICE				
Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
LOTSIZE	*4.040758*	*0.350757*	*11.52011*	*0.0000*
BEDROOMS	*2353.935*	*1084.623*	*2.170280*	*0.0304*
BATHRMS	*15480.96*	*1538.040*	*10.06538*	*0.0000*
DRIVEWAY	*8052.076*	*2107.907*	*3.819939*	*0.0001*
FULLBASE	*5697.870*	*1649.906*	*3.453451*	*0.0006*
PREFAREA	*8873.391*	*1731.846*	*5.123660*	*0.0000*
RECROOM	*4069.431*	*1972.623*	*2.062954*	*0.0396*
AIRCO	*12353.98*	*1592.301*	*7.758570*	*0.0000*
STORIES	*6316.716*	*958.5377*	*6.589950*	*0.0000*
C	*-6619.304*	*3510.132*	*-1.885770*	*0.0599*
R-squared	*0.645637*	*Mean dependent var*		*68121.60*
Adjusted R-squared	*0.639687*	*S.D. dependent var*		*26702.67*
S.E. of regression	*16028.57*	*Akaike info criterion*		*22.22028*
Sum squared resid	*1.38E+11*	*Schwarz criterion*		*22.29908*
Log likelihood	*-6056.136*	*Hannan-Quinn criter.*		*22.25108*
F-statistic	*108.5081*	*Durbin-Watson stat*		*1.562506*
Prob(F-statistic)	*0.000000*			
Redundant Variables: DRIVEWAY				
F-statistic	*14.59193*	*Prob. F(1,536)*		*0.0001*
Log likelihood ratio	*14.66544*	*Prob. Chi-Square(1)*		*0.0001*
Dependent Variable: PRICE				
Method: Least Squares		*Included observations:*		
Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
LOTSIZE	*4.380574*	*0.343556*	*12.75070*	*0.0000*
BEDROOMS	*1893.405*	*1091.458*	*1.734749*	*0.0834*
BATHRMS	*15307.73*	*1556.706*	*9.833411*	*0.0000*
FULLBASE	*5890.468*	*1669.875*	*3.527489*	*0.0005*
PREFAREA	*9731.681*	*1738.805*	*5.596763*	*0.0000*
RECROOM	*4275.759*	*1996.683*	*2.141431*	*0.0327*
AIRCO	*12392.28*	*1612.295*	*7.686113*	*0.0000*
STORIES	*6824.015*	*961.2320*	*7.099238*	*0.0000*
C	*-1099.044*	*3239.084*	*-0.339307*	*0.7345*
R-squared	*0.635990*	*Mean dependent var*		*68121.60*
Adjusted R-squared	*0.630567*	*S.D. dependent var*		*26702.67*
S.E. of regression	*16230.15*	*Akaike info criterion*		*22.24348*
Sum squared resid	*1.41E+11*	*Schwarz criterion*		*22.31440*
Log likelihood	*-6063.469*	*Hannan-Quinn criter.*		*22.27120*
F-statistic	*117.2792*	*Durbin-Watson stat*		*1.583598*
Prob(F-statistic)	*0.000000*			

CHOOSING A FUNCTIONAL FORM

To specify the classical linear regression model, a specific functional form should be chosen. Any functional form that is linear in parameters can be chosen. If the incorrect functional form is chosen, then the model is misspecified. If the model is misspecified then it may not be a reasonable approximation of the true data generation process. We make a functional form specification error when we choose the wrong functional form.

Constant term: constant term should be included in the regression model unless there is some strong reason for opposite such as the data is in the close neighborhood. Not including a constant term causes inflated t-ratio.

Functional Forms:

The Log-log Regression Model (Double Log)

$$lny_i = \beta_0 + \beta_1 lnx_{1i} + \beta_2 lnx_{2i} + \cdots + \beta_k x_{ki} + e_i \qquad (7.12)$$

When
$x_i changes\ 1\%, y\ changes\ \beta_k\ \%, holding\ the\ other\ regressors\ constant.$
Consider the following exponential regression model:

$$y_i = \alpha x_i^{\beta_1} \varepsilon^{e_i}, \qquad (7.13)$$

It can be expressed in logs:

$$lny_i = \beta_0 + \beta_1 lnx_i + e_i, \qquad (7.14)$$

It is called linear in logs and can be estimated by ***OLS*** on the condition that classical assumptions are satisfied.

Cobb-Douglas Production Function

$$Q = \alpha_0 L^{\beta_1} K^{\beta_2} \Longrightarrow logQ = log\alpha_0 + \beta_1 logL + \beta_2 logK \qquad (7.15)$$

When L changes 1%, Q changes by β_1%

Lin-Log Model (semi log)

$$y_i = \beta_0 + \beta_1 lnx_{1i} + \cdots + e_i \quad (7.16)$$

when $\boldsymbol{x_1}$ changes 1%, $\boldsymbol{y}$ changes 0,01 $\times$ $\boldsymbol{\beta_1}$, holding other variables constant**.**
Theimpact of a variation in $\boldsymbol{x_i}$ on **y** decreases as $\boldsymbol{x_i}$ gets larger.

Log-ln Model

$$lny_i = \beta_0 + \beta_1 x_{1i} + \cdots + e_i \quad (7.17)$$

when $\boldsymbol{x_1}$ changes one unit y_i changes 100 × $\boldsymbol{\beta_1}$% holding the other regressors constant. The impact of a variation $in\ \boldsymbol{x_i}\ on\ \boldsymbol{y_i}\ increases\ with\ \boldsymbol{y_i}$.

Quadratic Forms

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{2i} + e_i, \quad i=1,\ldots..,n \quad (7.18)$$

Cost curve is U-shaped and cost is quadratic in output with $\boldsymbol{\beta_1} < 0\ and\ \boldsymbol{\beta_2} > 0$.
$\boldsymbol{y_i}$ increase with $\boldsymbol{x_{1i}}$ but decrease with $\boldsymbol{x_{1i}^2}$**.**

Inverse Form

$$y_i = \beta_0 + \beta_1 \frac{1}{x_{1i}} + \beta_2 x_{2i} + e_i \quad (7.19)$$

$the\ slope\ approaces\ to\ zero\ when\ \boldsymbol{x_{1i}}$ is large.

Intercept Dummy Independent Variable

The intercept dummy independent variable changes the intercept but the slope remains constant. One dummy variable is used for two categories. If there are more than two categories more than one dummy independent variable can be set in the regression equation.

To include two same dummy variable cause perfect multicollinearity and violates Assumption 6.

Slope Dummy Independent Variable

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D_i + \beta_3 x_i D_i + e_i \quad (7.20)$$

There are two equations as follows:

$$y_i = \beta_0 + \beta_2 + (\beta_1 + \beta_3) x_i + e_i, \ \ if\ D_i = 1 \quad (7.21)$$

$$\beta_0 + \beta_1 x_i + e_i, \ \ if\ D_i = 0 \quad (7.22)$$

Each equation can be estimated separately.

Interaction term should be included if there is reason to believe that the slopes are different across categories.

Lags

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + e_i \quad (7.23)$$

The length of time between cause and effect is called a lag.

$\boldsymbol{y_{t-1}}$ is lagged independent variable. $\boldsymbol{\beta_1}$ measures the impact of previous observation on the current observation. If lag structure take place over more than one time period, it is called distributed lags.

Mixed Functional Forms

It is possible to mix functional forms as follows:

$$y_i = \beta_0 + \beta_1 ln x_i + \beta_2 z_i + \beta_3 \varphi_i^2 + e_i \quad (7.24)$$

In this case, $\boldsymbol{y_i}$ are a semi-log function of $\boldsymbol{x_i}$, a quadratic function of $\boldsymbol{\varphi_i}$, and a linear function of $\boldsymbol{z_i}$. The marginal effect and elasticity for each of these variables is given by the formulas above.

How to evaluate functional form?

In the residual plot, whether there is a systematic pattern between e_i and x_i, should be checked. Different functional forms should be used.

Consequences of Choosing the Wrong Functional Form

The ***OLS*** estimator will be biased and not valid.

Detection and Correction of Functional Form Specification Errors

Following two alternative methodologies are used to choose a specific functional form for a model:

i. Maintained hypothesis methodology

ii. Theory/testing methodology

Maintained Hypothesis Methodology

This methodology uses theory and/or tractability to choose a specific functional form. Once a specific functional form is chosen, it is treated as a maintained hypothesis and not tested using the sample data. Choosing functional form based on tractability is not good practice. Relying on theory alone may not be a good practice. This is because there are many situations when theory has nothing to say about the appropriate functional form. If the wrong functional form is chosen, then the parameter estimates will be biased and all tests of hypothesis, strictly speaking, will be incorrect.

Theory/Testing Methodology

This methodology involves the following steps:

i- A set of specific functional forms that are consistent with theory is identified. If a specific functional form is inconsistent with the theory being used to guide the

specification of the statistical model, then it should not be considered.

ii- Statistical tests should be conducted to determine which specific functional form should be chosen.

iii- One of the following two approaches may be used in the third step;
1) Testing-down approach; 2) Testing-up approach.

Testing-Down Approach

When using the testing-down approach, we begin with a general model and test-down to a more specific model. To test for nonlinear terms and interaction terms, begin with a general model that includes one or more of these terms. For example, the general model might include nonlinear terms such as X^2 and/or lnX, or interaction terms such as X*A. A ***t-test*** and/or an ***F-test*** can be employed to test whether these terms belong in the model.

General Functional Forms

A more systematic approach is to begin with a general functional form. This general functional form has one or more specific functional forms as a special case. An ***F-test*** or a ***t-test*** is used to test whether a specific functional form is the appropriate functional form.

Testing-up Approach

When using the testing-up approach, begin with a specific model and test-up to a more general model. To test for nonlinear terms and interaction terms, a specific model that does not include one or more of these terms may be appropriate. For example, the specific model might not include nonlinear terms such as X^2 and/or lnX, or interaction terms such as X*A. We then use a Lagrange multiplier test to test whether these terms should be added to the model.

Other Tests for Functional Form

A number of other criteria and statistical tests are also used to test for functional form. Some of these are the following:

i- Adjusted R^2

ii- Ramsey's Reset Test

iii- Recursive Residual Test

SPECIFICATION TESTS

Residual Tests

Correlograms and Q-Statistics

This test displays the autocorrelation and partial autocorrelations of the squared residuals up to specified number of lags. It is available for LS, TSLS, nonlinear LS, binary, ordered, censored, and count methods.

Example 7.4

Table 7.6 Eviews Result for Correlograms and Q-Statistics

Sample: 1954Q1 1994Q4
Included observations: 164

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
.|****** |	.|****** |	1	0.869	0.869	126.03	0.000
.|***** |	*|. |	2	0.723	-0.130	213.81	0.000
.|***** |	.|* |	3	0.644	0.196	283.88	0.000
.|**** |	*|. |	4	0.543	-0.185	333.97	0.000
.|*** |	.|. |	5	0.432	-0.016	365.96	0.000
.|** |	**|. |	6	0.301	-0.229	381.58	0.000
.|* |	.|. |	7	0.179	-0.014	387.12	0.000
.|* |	.|* |	8	0.131	0.162	390.13	0.000
.|* |	*|. |	9	0.077	-0.101	391.17	0.000
.|. |	.|. |	10	-0.006	-0.047	391.18	0.000
.|. |	.|. |	11	-0.058	-0.004	391.78	0.000
*|. |	.|. |	12	-0.080	0.027	392.93	0.000
*|. |	.|. |	13	-0.089	0.000	394.37	0.000
*|. |	.|. |	14	-0.089	0.032	395.81	0.000
*|. |	.|. |	15	-0.097	-0.002	397.52	0.000
*|. |	.|. |	16	-0.095	-0.012	399.18	0.000
.|. |	.|* |	17	-0.053	0.095	399.71	0.000
.|. |	.|. |	18	-0.009	0.038	399.72	0.000
.|. |	.|. |	19	0.006	-0.049	399.73	0.000
.|. |	.|. |	20	0.027	0.041	399.87	0.000
.|. |	.|. |	21	0.060	0.008	400.56	0.000
.|* |	.|. |	22	0.080	-0.035	401.77	0.000
.|* |	.|. |	23	0.080	-0.052	403.02	0.000
.|. |	.|. |	24	0.061	-0.024	403.74	0.000
.|. |	.|* |	25	0.059	0.077	404.42	0.000
.|* |	.|. |	26	0.081	0.034	405.73	0.000
.|* |	.|. |	27	0.081	-0.002	407.03	0.000
.|* |	.|. |	28	0.078	0.070	408.24	0.000
.|* |	.|. |	29	0.087	0.001	409.78	0.000
.|* |	.|. |	30	0.085	-0.051	411.23	0.000
.|* |	.|. |	31	0.075	-0.028	412.38	0.000
.|. |	.|. |	32	0.050	-0.055	412.90	0.000
.|. |	.|. |	33	0.014	-0.017	412.94	0.000
.|. |	.|. |	34	-0.010	-0.028	412.96	0.000
.|. |	.|. |	35	-0.030	0.010	413.15	0.000
.|. |	.|* |	36	-0.030	0.126	413.34	0.000

Correlograms of Squared Residuals

This test displays the autocorrelation and partial autocorrelations of the squared residuals up to specified number of lags defined. It can be used to check ***ARCH*** in the residuals. If autocorrelation and partial autocorrelation is equal to zero, there is no ***ARCH*** effect in the residuals.

Histogram and Normality test

This test provides you a histogram and descriptive statistics of the residuals, including Jarque-Bera statistics (JB, 1981) for testing normality. If the residuals are normally distributed, the histogram should be bell-shaped and ***Jarque-Bera test*** statistics should not be significant.

Example 7.5

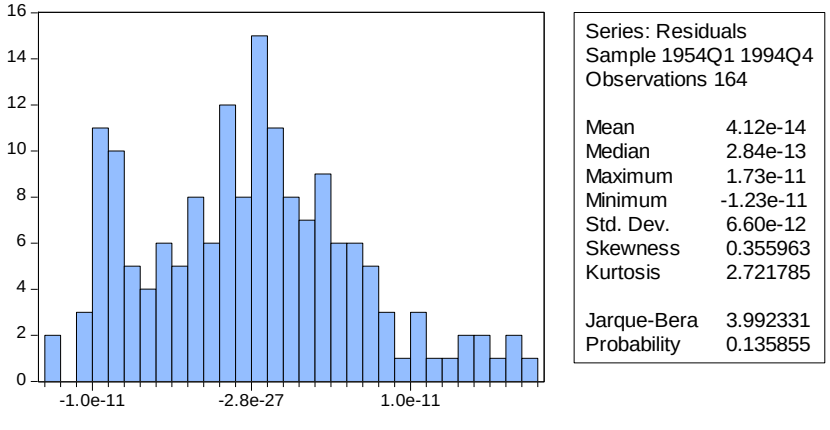


Figure 7.1 Eviews Result for Normality Test

Serial Correlation LM Test

This test is an alternative to the ***Q-statistics*** for testing serial correlation. Unlike Durbin Watson test (AR (1)), ***LM test*** can be used to test higher order ***ARMA*** errors.

The null hypothesis of the ***LM test*** is there is no serial correlation up to chosen lag order. Eviews reports two test statistics. The ***F-statistics***

is an omitted variable test for the joint significance of all lagged residuals. Omitted variables are residuals not independent variables. ***R^2 statistic*** is the Breusch-Godfrey LM test statistics.

Example 7.6

Table 7.7 Eviews Result for Serial Correlation LM Test

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	*2322039.*	*Prob. F(2,156)*	*0.0000*
*Obs*R-squared*	*163.9945*	*Prob. Chi-Square(2)*	*0.0000*

Test Equation:
Dependent Variable: RESID
Method: Least Squares
Sample: 1954Q1 1994Q4
Included observations: 164
Presample missing value lagged residuals set to zero.

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
TBR	*-3.87E-12*	*1.03E-14*	*-375.3065*	*0.0000*
Y	*5.33E-11*	*5.35E-14*	*994.7556*	*0.0000*
MF	*-1.00E-11*	*1.23E-13*	*-81.65197*	*0.0000*
M	*-1.09E-10*	*1.09E-13*	*-997.9981*	*0.0000*
CPR	*1.68E-14*	*9.17E-15*	*1.830093*	*0.0691*
C	*3.58E-10*	*8.06E-13*	*444.1100*	*0.0000*
RESID(-1)	*-0.001390*	*0.001407*	*-0.988141*	*0.3246*
RESID(-2)	*-0.000686*	*0.001019*	*-0.673236*	*0.5018*

R-squared	*0.999966*	*Mean dependent var*	*4.12E-14*
Adjusted R-squared	*0.999965*	*S.D. dependent var*	*6.60E-12*
S.E. of regression	*3.91E-14*	*Sum squared resid*	*2.38E-25*
F-statistic	*663413.6*	*Durbin-Watson stat*	*1.993998*
Prob(F-statistic)	*0.000000*		

ARCH-LM Test

This is ***LM test*** for ***ARCH*** (Engle, 1982) in the residuals. Ignoring ***ARCH*** effect can cause inefficiency in estimation. The null hypothesis is that

there is no ***ARCH*** effect in the residuals. It is computed from the residual test regression as follows:

$$e_t^2 = \beta_0 + \beta_1 e_{t-1}^2 + \beta_2 e_{t-2}^2 + \cdots + \beta_q e_{t-q}^2 + \upsilon_t, \qquad (7.25)$$

e is the residual.

This is a regression of the squared residuals on constant and lagged squared residuals up to order q. The ***F-statistic*** is an omitted variable test for the joint significance of all lagged squared residuals. Engels ***LM test*** statistic is equal to number of observations times R^2 ***statistic***.

Example 7.7

Table 7.8 Eviews Result for ARCH-LM Test

Heteroscedasticity Test: ARCH

F-statistic	*122.9920*	*Prob. F(4,155)*	*0.0000*
*Obs*R-squared*	*121.6674*	*Prob. Chi-Square(4)*	*0.0000*

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Sample (adjusted): 1955Q1 1994Q4
Included observations: 160 after adjustments

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
C	*6.23E-24*	*2.91E-24*	*2.141565*	*0.0338*
RESID^2(-1)	*0.997611*	*0.080486*	*12.39489*	*0.0000*
RESID^2(-2)	*-0.245270*	*0.111848*	*-2.192884*	*0.0298*
RESID^2(-3)	*0.229981*	*0.111806*	*2.056967*	*0.0414*
RESID^2(-4)	*-0.120375*	*0.080923*	*-1.487518*	*0.1389*

R-squared	*0.760421*	*Mean dependent var*	*4.23E-23*
Adjusted R-squared	*0.754238*	*S.D. dependent var*	*5.75E-23*
S.E. of regression	*2.85E-23*	*Sum squared resid*	*1.26E-43*
F-statistic	*122.9920*	*Durbin-Watson stat*	*1.983015*
Prob(F-statistic)	*0.000000*		

Whites Heteroscedasticity Test

It is a test for heteroscedasticity in the residuals from a ***LS*** regression (White, 1980). In the presence of heteroscedasticity standard errors are no longer valid but ***OLS*** estimates are still consistent. In order to correct heteroscedasticity, weighted least squares estimation method can be employed or chosen the robust standard error option to correct the standard errors.

White test is test of null hypothesis of that there is no heteroscedasticity.

For example we estimate following regression,

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 z_t + e_t, \tag{7.26}$$

the test statistics based on auxiliary regression.

$$e_t^2 = \alpha_0 + \alpha_1 x_t + \alpha_2 z_t + \alpha_3 x_t^2 + \alpha_4 z_t^2 + \alpha_5 x_t z_t + \upsilon_t \tag{7.27}$$

The ***White test*** statistic is computed the number of observations times R^2 from the regression.

Example 7.8

Table 7.9 Eviews Result for Whites Heteroscedasticity Test

Heteroscedasticity Test: White			
F-statistic	*120.7133*	*Prob. F(18,145)*	*0.0000*
*Obs*R-squared*	*153.7404*	*Prob. Chi-Square(18)*	*0.0000*
Scaled explained SS	*123.4851*	*Prob. Chi-Square(18)*	*0.0000*
Test Equation:			
Dependent Variable: RESID^2			
Method: Least Squares			
Sample: 1954Q1 1994Q4			
Included observations: 164			
Collinear test regressors dropped from specification			

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-2.02E-19	1.57E-20	-12.81748	0.0000
TBR	1.60E-22	9.94E-22	0.160650	0.8726
TBR^2	9.78E-24	7.02E-24	1.392979	0.1658
TBR*Y	-4.17E-22	4.82E-23	-8.644330	0.0000
TBR*MF	-3.64E-22	1.37E-22	-2.658945	0.0087
TBR*M	8.36E-22	1.03E-22	8.125272	0.0000
TBR*CPR	9.93E-24	1.29E-23	0.767886	0.4438
Y	1.35E-20	2.00E-21	6.763355	0.0000
Y^2	2.95E-21	9.05E-23	32.57495	0.0000
Y*MF	2.77E-21	3.16E-22	8.783545	0.0000
Y*M	-1.19E-20	3.70E-22	-32.14765	0.0000
Y*CPR	-1.57E-24	3.91E-23	-0.040015	0.9681
MF	4.74E-20	2.68E-21	17.66015	0.0000
MF*M	-1.08E-20	3.93E-22	-27.41762	0.0000
MF*CPR	1.52E-22	1.33E-22	1.146168	0.2536
M^2	1.25E-20	3.99E-22	31.25539	0.0000
M*CPR	1.64E-23	8.29E-23	0.197514	0.8437
CPR	-1.07E-21	9.46E-22	-1.130305	0.2602
CPR^2	-4.91E-24	6.30E-24	-0.780155	0.4366

R-squared	0.937442	Mean dependent var	4.33E-23
Adjusted R-squared	0.929676	S.D. dependent var	5.71E-23
S.E. of regression	1.52E-23	Sum squared resid	3.33E-44
F-statistic	120.7133	Durbin-Watson stat	1.313970
Prob(F-statistic)	0.000000		

CHAPTER 8

PARAMETER STABILITY

Aim of the structural break and parameter stability test is to check whether the parameters of the model are constant or stable over the subsamples of the data. The stability of the parameter estimates in the sample period can be obtained by computing recursive coefficient estimates and looking at their plots. It provides information about the stability of the estimates. Main idea is repeatedly adding one observation and re-estimating the parameters. From plots we can conclude that the estimates are stable.

One empirical technique is to divide the whole samples into two subsamples, and use first subsample for estimation and second subsample for testing. If there is obvious data points where structural break have taken place, the subsamples might be determined based on these points. If there is no clear structural break, a rule-of-thumb which is 80% to 90% of the observations for estimation and remainder for testing is followed.

CHOW BREAKPOINT TEST FOR STRUCTURAL BREAK

Possible structural breaks points are defined. This test fits the equation separately for each subsample and sees whether there are significant differences in the estimated equations. If significant difference is detected, then it is concluded that there is structural change in the relationship. This test can be used with least squares and two-stage least squares estimation method. ***F-statistic***, with a single breakpoint is computed as follows:

$$F = \frac{(\bar{u}^l\bar{u} - u_1^l u_1 - u_2^l u_2)/K}{(u_1^l u_1 + u_2^l u_2)/(T-2K)} \qquad (8.1)$$

$\bar{u}^l\bar{u}$: the restricted sum of squared residuals,
$u_1^l u_1$: the sum of squared residuals from subsample 1,
$u_2^l u_2$: the sum of squared residuals from subsample 2,
T : the total number of observations
K : the number of parameters in the equation

This formula can be generalized for more than one breakpoint.
The Log likelihood ratio test result is also reported.

Major drawback of the Chow breakpoint test (Chow, 1960) is that each subsample requires the number of observation as many as the number of parameters.

Example 8.1

Table 8.1 Eviews Result for Chow Breakpoint Test

Dependent Variable: INFL				
Method: Least Squares				
Sample: 1954Q1 1994Q4				
Included observations: 164				
Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
TBR	*-0.430756*	*0.406092*	*-1.060735*	*0.2904*
M	*-2.864898*	*1.129140*	*-2.537240*	*0.0121*
CPR	*0.946117*	*0.388299*	*2.436569*	*0.0159*
C	*19.14048*	*7.279982*	*2.629194*	*0.0094*
R-squared	*0.453612*	*Mean dependent var*		*4.280384*
Adjusted R-squared	*0.443367*	*S.D. dependent var*		*2.566663*
S.E. of regression	*1.914931*	*Akaike info criterion*		*4.161328*
Sum squared resid	*586.7136*	*Schwarz criterion*		*4.236934*
Log likelihood	*-337.2289*	*Hannan-Quinn criter.*		*4.192021*
F-statistic	*44.27745*	*Durbin-Watson stat*		*0.848191*
Prob(F-statistic)	*0.000000*			
Chow Breakpoint Test: 1970Q2				
Null Hypothesis: No breaks at specified breakpoints				
Varying regressors: All equation variables				
Equation Sample: 1954Q1 1994Q4				
F-statistic	*15.66846*	*Prob. F(4,156)*		*0.0000*
Log likelihood ratio	*55.38695*	*Prob. Chi-Square(4)*		*0.0000*
Wald Statistic	*62.67385*	*Prob. Chi-Square(4)*		*0.0000*

CHOW FORECAST TEST FOR STRUCTURAL BREAK

If there is only a few observations or the number of observations are less than the number of parameters in the equation, then Chow Forecast test should be used. This test can be used with least squares or two stage least squares estimation method. ***F statistic*** is computed as follows:

$$F = \frac{(\bar{u}^l\bar{u} - u^l u)/T_2}{u^l u/(T_1 - K)} \qquad (8.2)$$

$\bar{u}^l\bar{u}$: the residuals sum of squares when the equation is fitted to all T observations

$u^l u$: the residuals sum of squares when the equation is fitted to T_1 observations

The Log likelihood ratio test result is also reported.

Example 8.2

Table 8.2 Eviews Result for Chow Forecast Test

Chow Forecast Test: Forecast from 1980Q3 to 1994Q4				
F-statistic	*1.498472*	*Prob. F(58,102)*		*0.0374*
Log likelihood ratio	*101.0741*	*Prob. Chi-Square(58)*		*0.0004*
Test Equation:				
Dependent Variable: INFL				
Method: Least Squares				
Sample: 1954Q1 1980Q2				
Included observations: 106				
Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
TBR	*0.756093*	*0.494190*	*1.529963*	*0.1291*
M	*12.49727*	*4.787969*	*2.610141*	*0.0104*
CPR	*0.052019*	*0.457198*	*0.113778*	*0.9096*
C	*-78.75231*	*30.25633*	*-2.602838*	*0.0106*
R-squared	*0.606890*	*Mean dependent var*		*4.421247*
Adjusted R-squared	*0.595328*	*S.D. dependent var*		*2.770335*
S.E. of regression	*1.762317*	*Akaike info criterion*		*4.008141*
Sum squared resid	*316.7876*	*Schwarz criterion*		*4.108648*
Log likelihood	*-208.4315*	*Hannan-Quinn criter.*		*4.048877*
F-statistic	*52.48973*	*Durbin-Watson stat*		*1.321713*
Prob(F-statistic)	*0.000000*			

CUSUM TEST

The CUSUM test statistic (Brown, Durbin, and Evans, 1975) is based on cumulative sums of scaled recursive residuals and is plotting the cumulative sum together with the 5% critical lines against time (Vogelvang, 133). If the cumulative sum goes outside of the 5% critical lines, then the test shows parameter instability. The CUSUM test is based on the following recursive residual test statistics defined as follows:

$$W_t = \sum_{r=k+1}^{t} \frac{w_r}{S_t}, \quad t = k+1, \dots, T \tag{8.3}$$

w : recursive residual
s : the standard error of the regression fitted to all T observations.

The expectations of the CUSUM statistics are zero under the null hypothesis of constant parameters.

Example 8.3

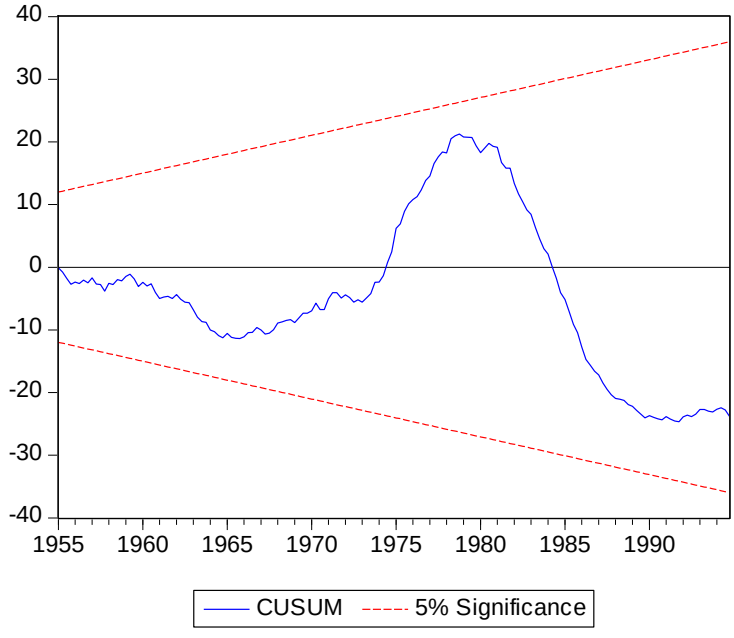


Figure 8.1 Ewievs Result for CUSUM Test

CUSUM OF SQUARES TEST

The CUSUM of squares statistics is a cumulative sum of squared residuals. The expectations of the CUSUM of squares statistics run from zero at the first observation until the value of one at the end of the sample period, under the null hypothesis of constant coefficients and variance. The CUSUM test statistic (Brown, Durbin, and Evans, 1975) is defined as follows:

$$S_t = \sum_{r=k+1}^{t} w_r^2 / \sum_{r=k+1}^{T} w_r^2 \qquad (8.4)$$

The expected value of ***S*** under the null hypothesis of parameter constancy or stability is,

$$E[S_t] = (t-k)/(T-k) \qquad (8.5)$$

If t=k, then it goes zero
If t=T, then it goes to unity
S_t is plotted together with the 5% critical lines against time. Movement of the critical lines shows parameter instability.

Example 8.4

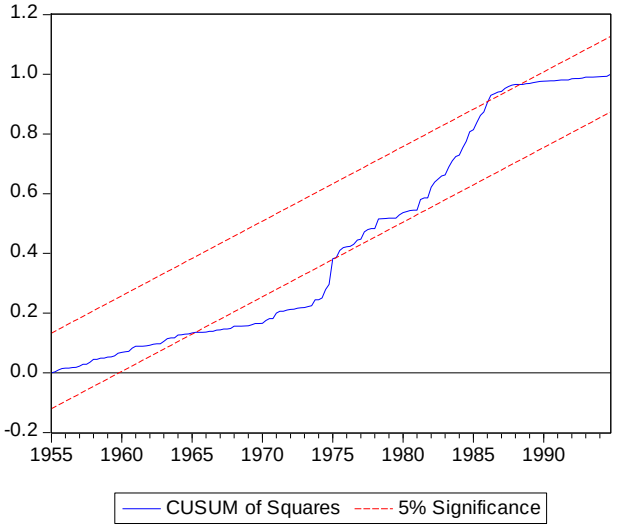


Figure 8.2 Ewievs Result for CUSUM of Squares Test

ONE STEP AHEAD FORECAST TEST

This test produces a plot of the residuals and standard errors of the sample points whose probability value is equal or less than 15%. The upper portion of the plot repeats the recursive residuals and standard errors displayed by the recursive residuals option, and the lower portion of the plot indicates the values for those sample points where the hypothesis of parameter stability would be rejected at the 5%, 10%, or 15% significance levels. If ***p-value*** is less than 0.05, it

means that the recursive residual go outside the two standard error bounds at this point of observation.

Example 8.5

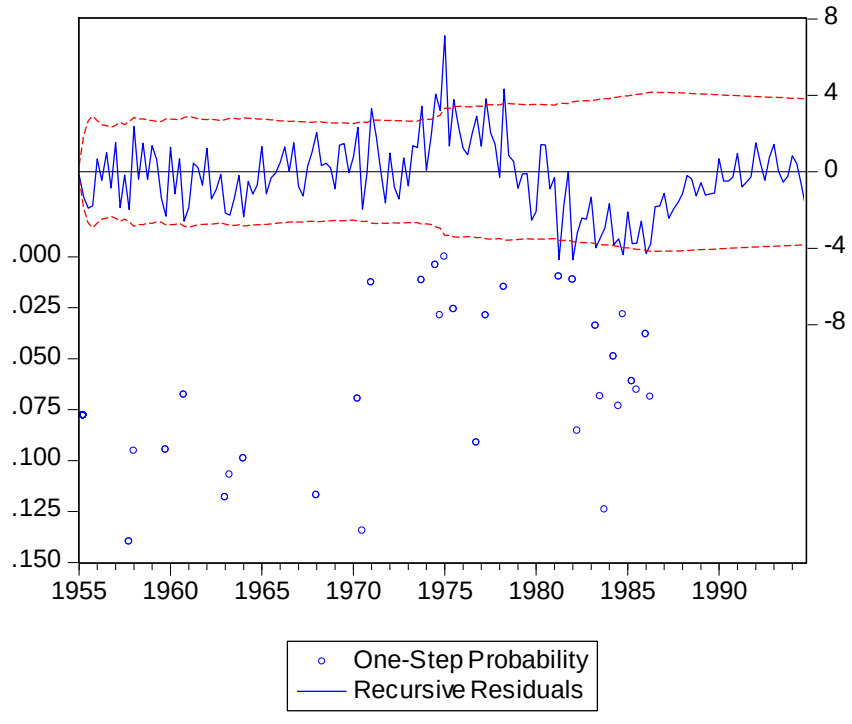


Figure 8.3 Ewievs Result for One Step Ahead Forecast Test

N STEP FORECAST TEST

This test employs the recursive calculations. This test does not require specifying a forecast period; it calculates all feasible cases, starting with the smallest possible sample size in estimating and forecasting equation through adding one observation at a time. Upper portion of the plot indicates recursive residuals and lower portion shows the corresponded ***p-values***.

Example 8.6

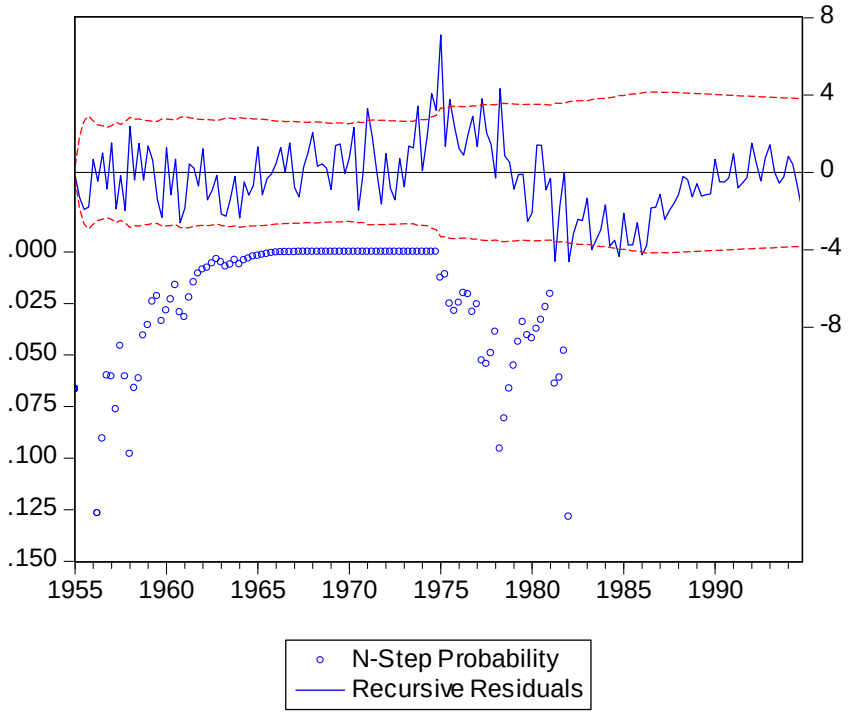


Figure 8.4 Ewievs Result for N Step Forecast Test

RECURSIVE COEFFICIENT ESTIMATES

This test shows the evolution of estimates for all coefficients as more and more of the data are used in the estimation. It gives a plot of selected coefficients in the equation for all feasible recursive estimations including two standard error bands around the estimated coefficients.

If the coefficient shows significant variations as more observation is added to the estimation equation, then there is enough evidence for instability.

Example 8.7

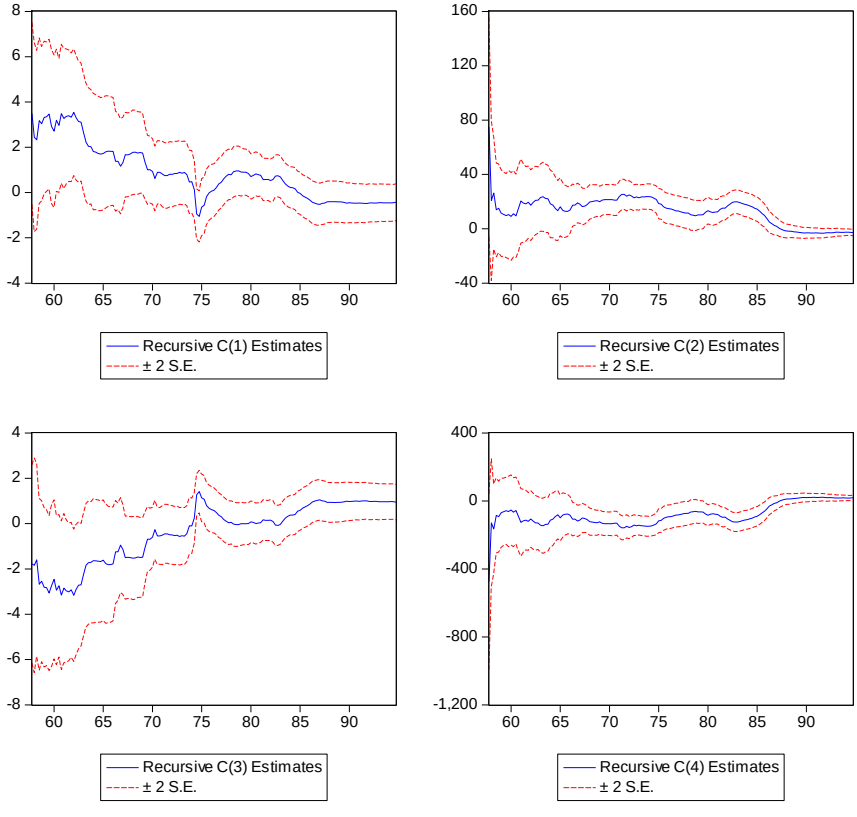


Figure 8.5 Eviews Result for Recursive Coefficient Estimates

WALD TEST FOR STRUCTURAL CHANGE

Chow test for parameter stability is used on the condition that the errors are independent, normally distributed and equal residual variance. Wald test is used in the case of unequal residual variance in different sub-samples.

A Wald statistic for the null hypothesis of no structural change in independent samples is constructed as follows:

$$W = (b_1 - b_2)^l (V_1 - V_2)^{-1} (b_1 - b_2) \qquad (8.6)$$

which has asymptotic χ^2 distribution.

Degrees of freedom are equal to the number of estimated parameters in the estimated ***b*** vector.

Example 8.8

Table 8.3 Eviews Test Result for Wald Test

Ramsey RESET Test:			
F-statistic	*0.382882*	*Prob. F(1,159)*	*0.5370*
Log likelihood ratio	*0.394447*	*Prob. Chi-Square(1)*	*0.5300*

Test Equation:
Dependent Variable: INFL
Method: Least Squares
Sample: 1954Q1 1994Q4
Included observations: 164

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
TBR	*-0.521658*	*0.432586*	*-1.205906*	*0.2296*
M	*-3.500625*	*1.528213*	*-2.290665*	*0.0233*
CPR	*1.160862*	*0.521347*	*2.226659*	*0.0274*
C	*22.87393*	*9.466138*	*2.416395*	*0.0168*
FITTED^2	*-0.023118*	*0.037361*	*-0.618775*	*0.5370*

R-squared	*0.454925*	*Mean dependent var*	*4.280384*
Adjusted R-squared	*0.441212*	*S.D. dependent var*	*2.566663*
S.E. of regression	*1.918635*	*Akaike info criterion*	*4.171118*
Sum squared resid	*585.3042*	*Schwarz criterion*	*4.265626*
Log likelihood	*-337.0317*	*Hannan-Quinn criter.*	*4.209485*
F-statistic	*33.17572*	*Durbin-Watson stat*	*0.847180*
Prob(F-statistic)	*0.000000*		

CHAPTER 9

INTERPRETING THE LINEAR REGRESSION MODELS

SIMPLE LINEAR REGRESSION MODEL

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad (9.1)$$

We assume that ***CLRM*** assumptions are satisfied.

$\boldsymbol{\beta_1}$ measures the expected change in $\boldsymbol{y_i}$ if the $\boldsymbol{x_i}$ changes one unit.

$$\frac{\partial E\{y_i|x_i\}}{\partial x_i} = \beta_1 \quad (9.2)$$

Example 9.1

Table 9.1 Eviews Output for Simple Regression Model Estimates

Dependent Variable: CONS				
Method: Least Squares				
Included observations: 30				
Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
TEMP	*0.003107*	*0.000478*	*6.502305*	*0.0000*
C	*0.206862*	*0.024700*	*8.374902*	*0.0000*
R-squared	*0.601593*	*Mean dependent var*		*0.359433*
Adjusted R-squared	*0.587365*	*S.D. dependent var*		*0.065791*
S.E. of regression	*0.042262*	*Akaike info criterion*		*-3.425533*
Sum squared resid	*0.050009*	*Schwarz criterion*		*-3.332120*
Log likelihood	*53.38299*	*Hannan-Quinn criter.*		*-3.395649*
F-statistic	*42.27997*	*Durbin-Watson stat*		*0.623564*
Prob(F-statistic)	*0.000000*			

Regression result based on the impact of temperature on the ice-cream consumption is reported in Table 9.1. $\beta_0 = 0.206862$, $\beta_1 = 0.003107$ t values are 8.374902 and 6.502305 respectively. It is obvious that T- statistics are statistically significant. It means that there is enough evidence to reject the null hypothesis of there is no impact.

The coefficient 0.003107 means that one unit increase in the temperature causes 0.003107 unit change of ice cream consumption.

MULTIPLE LINEAR REGRESSION MODEL

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 \varphi_i + e_i \qquad (9.3)$$

We assume that ***CLRM*** assumptions are hold.

The regressor $\boldsymbol{\beta_1}$ measures the expected change in $\boldsymbol{y_i}$ if the $\boldsymbol{x_i}$ changes one unit when we hold other regressors constant. This is called *ceteris paribus* condition. The interpretation of multiple linear regression estimates is valid under ceteris paribus condition. It is not possible to interpret one single regressor estimates without knowing

the other regressors of the model. If the research focus is to investigate only the relationship between the regressor and regressand or the impact of any change in regressor on the regressand then the other regressors in the regression equation are called control variables. They should be included in the model to make estimation robust.

Example 9.2

Table 9.2 Eviews Output for Multiple Regression Model Estimates

Dependent Variable: CONS
Method: Least Squares
Included observations: 30

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
TEMP	*0.003458*	*0.000446*	*7.762213*	*0.0000*
PRICE	*-1.044414*	*0.834357*	*-1.251759*	*0.2218*
INCOME	*0.003308*	*0.001171*	*2.823722*	*0.0090*
C	*0.197315*	*0.270216*	*0.730212*	*0.4718*

R-squared	*0.718994*	*Mean dependent var*	*0.359433*
Adjusted R-squared	*0.686570*	*S.D. dependent var*	*0.065791*
S.E. of regression	*0.036833*	*Akaike info criterion*	*-3.641296*
Sum squared resid	*0.035273*	*Schwarz criterion*	*-3.454469*
Log likelihood	*58.61944*	*Hannan-Quinn criter.*	*-3.581528*
F-statistic	*22.17489*	*Durbin-Watson stat*	*1.021170*
Prob(F-statistic)	*0.000000*		

Table 9.2 shows the OLS estimation result for regression and includes three independent variables. $\beta_0 = 0.197315,\ \beta_1 = 0.003458,\ \beta_2 = -1.044414\ and\ \beta_3 = 0.003308.$ In multivariate regression model, estimates are interpreted under ceteris paribus condition. It means that one unit change in temperature causes 0.003458 unit increase in ice-cream consumption when the other variables price and income hold constant.

Sometimes, it is very difficult to maintain the other regressors constant due to possible collinearity between them. For example, if the regression model includes age and experience together as

regressors, it is impossible to control experience and keep constant, while age varies or vice versa. For example:

$$income_i = \beta_0 + \beta_1 age_i + \beta_2 age_i^2 + \beta_3 edu_i + e_i \quad (9.4)$$

In the regression model above, it is impossible to conclude that coefficient $\boldsymbol{\beta_1}$ measures the effect of age given that age-squared as constant. In this situation, we should see the following derivative (write eq. number) which can be interpreted as the marginal effect of a change in age on the regressand when the other variables (excluding age_i^2) in the regression model are held constant.

$$\frac{\partial E\{y_i|x_i\}}{\partial age_i} = \beta_1 + 2\beta_2 age_i \quad (9.5)$$

Example 9.3

Table 9.3 Eviews Output for Multiple Regression Model Estimates

Dependent Variable: WAGE
Method: Least Squares
Sample: 1 3294
Included observations: 3294

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
SCHOOL	*0.599229*	*0.033402*	*17.94010*	*0.0000*
EXPER	*0.157261*	*0.024170*	*6.506507*	*0.0000*
C	*-2.476683*	*0.469997*	*-5.269573*	*0.0000*

R-squared	*0.091489*	*Mean dependent var*	*5.757585*
Adjusted R-squared	*0.090937*	*S.D. dependent var*	*3.269186*
S.E. of regression	*3.116999*	*Akaike info criterion*	*5.112529*
Sum squared resid	*31974.31*	*Schwarz criterion*	*5.118084*
Log likelihood	*-8417.335*	*Hannan-Quinn criter.*	*5.114517*
F-statistic	*165.7051*	*Durbin-Watson stat*	*1.820012*
Prob(F-statistic)	*0.000000*		

If we want to analyze whether the effect of age is different between men and women, we can include a dummy variable for men as follows:

$$income_i = \beta_0 + \beta_1 age_i + \beta_2 age_i^2 + \beta_3 edu_i + \beta_4 agemale_i + e_i \qquad (9.6)$$

The effect of changing age on income between male and female can be interpreted as follows:

$$\frac{\partial E\{y_i|x_i\}}{\partial age_i} = \beta_1 + \beta_4 male_i \qquad (9.7)$$

The effect of changing age is $\boldsymbol{\beta_1}$ for females and $\boldsymbol{\beta_1 + \beta_4}$ for males.

Example 9.4

Table 9.4 Eviews Output for Multiple Regression Model Estimates

Dependent Variable: WAGE
Method: Least Squares
Sample: 1 3294
Included observations: 3294

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
SCHOOL	*0.630602*	*0.034653*	*18.19771*	*0.0000*
EXPER	*0.203125*	*0.109467*	*1.855590*	*0.0636*
EXPER^2	*-0.004833*	*0.006596*	*-0.732760*	*0.4638*
MALE	*1.344831*	*0.107685*	*12.48853*	*0.0000*
C	*-3.576700*	*0.536916*	*-6.661567*	*0.0000*

R-squared	*0.132729*	*Mean dependent var*	*5.757585*
Adjusted R-squared	*0.131675*	*S.D. dependent var*	*3.269186*
S.E. of regression	*3.046358*	*Akaike info criterion*	*5.067287*
Sum squared resid	*30522.89*	*Schwarz criterion*	*5.076546*
Log likelihood	*-8340.822*	*Hannan-Quinn criter.*	*5.070602*
F-statistic	*125.8392*	*Durbin-Watson stat*	*1.905606*
Prob(F-statistic)	*0.000000*		

Table 9.4 shows the estimation result for the regression equation specified in Eq.9.7. Now it is possible to interpret marginal effect of a change in experience on wage when the other independent variables, schooling and sex are held constant.

Elasticity measures the relative change in the regressand due to any relative changes in the regressors. Elasticity is estimated in the linear regression model includes logarithms of regressors. For example the coefficient $\boldsymbol{\beta_1}$ in the following regression model measures the relative change in regressand due to relative change in the regressorx_i. Explaining $logy_i$, rather than y_i can be helpful to reduce heteroscedasticity.

$$logy_i = \beta_1(logx_i)' + e_i \quad (9.8)$$

$$logWAGE_i = \beta_0 + \beta_1 SCHOOL_i + ' \beta_2 logEXPER_i + \beta_3 EXPER_i^2 + \beta_4 MALE_i + e_i \quad (9.9)$$

Example 9.5

Table 9.5 Eviews Output for Multiple Regression Model Estimates

Dependent Variable: LOG(WAGE)
Method: Least Squares
Sample: 1 3294
Included observations: 3294

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
SCHOOL	*0.120493*	*0.006517*	*18.48886*	*0.0000*
LOG(EXPER)	*0.249870*	*0.071636*	*3.488034*	*0.0005*
EXPER^2	*9.63E-05*	*0.000612*	*0.157181*	*0.8751*
MALE	*0.242488*	*0.020440*	*11.86326*	*0.0000*
C	*-0.457195*	*0.121089*	*-3.775700*	*0.0002*

R-squared	*0.138814*	*Mean dependent var*	*1.587268*
Adjusted R-squared	*0.137766*	*S.D. dependent var*	*0.622725*
S.E. of regression	*0.578240*	*Akaike info criterion*	*1.743863*
Sum squared resid	*1099.717*	*Schwarz criterion*	*1.753122*
Log likelihood	*-2867.142*	*Hannan-Quinn criter.*	*1.747178*
F-statistic	*132.5375*	*Durbin-Watson stat*	*1.879750*
Prob(F-statistic)	*0.000000*		

Result in the Table 9.5 indicates that there is a relative change in wage because of absolute change in school, relative change in experience, marginal effect of experience and absolute change in

male. One unit relative change in expertise results 0.071636 unit relative changes in wage.

If x_i is dummy variable, or any other variable which may take negative values, then we cannot use log transformation. Regression model can include some variables in log and some in levels. We should include in the model as follows:

$$logy_i = \beta_1 x_i^{'} + e_i \quad (9.10)$$

In this case $\boldsymbol{\beta_1}$ measures the relative change in y_i due to an absolute change of one unit in the regressor $\boldsymbol{x_i}$. If it is a dummy variable for males, it can be interpreted as relative wage differential between men and women.

How our interpretations are effected due to the misspecifying the regressors? Assume that we have following two regression models to estimate.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i \quad (9.11)$$

$$and$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (9.12)$$

If we estimate the first model while, in fact, the second model is the correct one, then it will increase the variance of the estimators for the relevant regressors. Thus we will get higher variance and less reliable estimates.

If the second model is estimated instead of first model which is correct, then it will cause omitted variable bias. In this case the error term includes the omitted regressor. Then Assumption 2 (unbiasedness, $E(e_i) = \beta_2 z_i \neq 0$)and Assumption 3 (efficiency, $cov(e_i, x_i) \neq 0$) do not hold. Gauss- Markov theorem does not apply, and ***OLS*** is not ***BLUE***; estimates of regression coefficient is biased. ***OLS*** overestimate the coefficient of independent variable $\boldsymbol{\beta_1}$.

REGRESSOR SELECTION

To overcome the problem of including irrelevant regressors and omitting relevant regressors in the regression model following methods can be used together or separately:

- The potentially relevant variables should be determined based on the related economic theory
- The relevant regressor should be defined by checking the literature and using the 'common sense'
- The regressors should be determined based on research questions and research focus
- 'From specific to general' approach may be used: the approach suggests to form regression model as simple as possible and add regressors until the adequate model specification is obtained
- 'General to specific modeling' (Charemza and Deadman, 1999) approach may also be used: it starts with very general specification which is called general unrestricted model and reduces size and number of regressors in each step by testing restrictions until the model with adequate specification is obtained. Besides, including more variables and starting with the general model can cause multicollinearity problem which misguides the researchers.

If two regressors are omitted from the model, joint test should be used rather than two tests separately.

Best option is to start with the optimum model and test, whether the restrictions imposed by the model are correct, and whether the restrictions not imposed by the model should be imposed.

In the first category misspecification test for omitted variable bias, autocorrelation and heteroscedasticity test can be employed. Parameter restriction test, such as, whether one or more regressors have zero coefficients can be employed.

CHAPTER 10

ENDOGENEITY

In a statistical model, a parameter or variable is said to be endogenous when there is a correlation between the independent variable and the error term.

Endogeneity can arise as a result of the measurement error, auto regression with auto correlated errors, simultaneity and omitted variables. Broadly, a loop of, bivariate causality between the independent and dependent variables of a model leads to endogeneity.

Tariffs are an example of endogeneity problem.

$$y_i = \beta_1 x_i + e_i \tag{10.1}$$

$$x_i = \beta_2 y_i + \varepsilon_i \tag{10.2}$$

The current value of $\boldsymbol{x_i}$ depends on the current value of $\boldsymbol{y_i}$, $\boldsymbol{x_i}$ is influenced by current shocks to $\boldsymbol{y_i}$ as follows:

$$y_i = \beta_1(\beta_2 y_i + \varepsilon_i) + e_i \tag{10.3}$$

Thus, $\boldsymbol{x_i}$ *and* $\boldsymbol{e_i}$ are correlated *and* it causes endogeneity problem. This is the violation of OLS assumption.

In the presence of endogeneity, the linear model does not correspond to a conditional expectation and ***OLS*** cannot produce unbiased and consistent parameter estimates. Ceteris paribus is not valid and ***OLS*** estimator is biased and inconsistent and does not produce best linear approximation. Hypotheses tests are misleading. It can cause rejecting a hypothesis that is true (Type I Error) and fail to reject a hypothesis that in fact is false (Type II Error). For example, inflation is independent of all other factors within a given period, but influenced by the previous years' export and interest rate. Then we can say that inflation is exogenous within the period, but endogenous over time.

OLS *is no longer* ***BLUE*** *if endogeneity exists*.

SOURCES OF ENDOGENEITY

Autocorrelation with Lagged Dependent Variable

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 y_{t-1} + e_t \tag{10.4}$$

The ***OLS*** estimator provided that $E\{x_t, e_t\} = 0$, $E\{y_{t-1}, e_t\} = 0$ and the other assumptions are met. However, if $\boldsymbol{e_t}$ is depended on the first order autocorrelation as follows:

$$e_t = \rho e_{t-1} + u_t \tag{10.5}$$

we can rewrite the above model as:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 y_{t-1} + \rho e_{t-1} + u_t \tag{10.6}$$

and we also can write:

$$y_{t-1} = \beta_0 + \beta_1 x_{t-1} + \beta_2 y_{t-2} + e_{t-1} \tag{10.7}$$

Indicating that the error term $\boldsymbol{e_t}$ is correlated with $\boldsymbol{y_{t-1}}$. Thus, if $\boldsymbol{\rho = 0}$, ***OLS*** method does not produce consistent estimators.

A possible solution is the use of ***maximum likelihood*** or ***instrumental variables*** techniques. ***Durbin–Watson test*** is not valid to test autocorrelation in the model (Eq.10.7), because the condition that the explanatory variables should be treated as deterministic is violated. An alternative test, ***Breusch–Godfrey Lagrange Multiplier test*** for autocorrelation can be applied. This test statistic is computed as ***T times the R^2*** of a regression of the least squares residuals e_t on e_{t-1} and all included explanatory variables (including the relevant lagged values of y_t). Under H_0, the test statistic asymptotically has a Chi-squared distribution with 1 degree of freedom.

OLS is inconsistent whenever the model which is being estimated does not correspond to a conditional expectation. For example, a lagged dependent variable, combined with autocorrelation of the error term.

Omitted Variable Bias

Let's assume an independent variable (***z***) is correlated with another independent variable (***x***) and error term in the model.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i \qquad (10.8)$$

Omit z_i from the model and run the following regression:

$$y_i = \beta_0 + \beta_1 x_i + u_i, \qquad (10.9)$$

$$u_i = y z_i + e_i \qquad (10.10)$$

In this new regression model x_i is correlated with error term u_i (because there is correlation between x_i and the omitted variable z_i. ***Multicollinearity*** problem occurs.

Measurement Error

If true value of one independent variable is not available, instead of x_i following is defined:

$$x_i^* = x_i + \varepsilon_i, \qquad (10.11)$$

$\boldsymbol{\varepsilon_i}$ is the measurement error, and the following regression model is estimated as:

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad (10.12)$$

Actually, following regression is estimated mistakenly:

$$y_i = \beta_0 + \beta_1 (x_i^* - \varepsilon_i) + e_i \quad (10.13)$$

$$y_i = \beta_0 + \beta_1 x_i^* + (e_i - \beta_1 \varepsilon_i) \quad (10.14)$$

$$y_i = \beta_0 + \beta_1 x_i^* + u_i \quad (10.15)$$

$since\ x_i^*$ and u_i depend on ε_i, they are correlated and ***OLS*** estimator is inconsistent. That is, $\boldsymbol{E\{x_i^*, u_i\} \neq 0}$ and one of the necessary conditions for consistency of ***OLS*** estimator is violated.

Simultaneity

Simultaneity mostly occurs in the dynamic models.

Suppose that we have two structural equations as follows:
$C_t = \beta_0 + \beta_1 Y_t + e_i$, (Keynesian consumption function) (10.16)

Where $\boldsymbol{C_t}$ denotes a country`s real per capita consumption and $\boldsymbol{Y_t}$ is real capita income. $\boldsymbol{\beta_1}$ has a causal interpretation reflecting the impact of income on consumption $(0 < \beta_1 < 1)$. It measures how much individuals increase their consumption if their income increases by one unit.

$\boldsymbol{Y_t}$ is also defined as:

$$Y_t = C_t + I_t \quad (10.17)$$

Where $\boldsymbol{I_t}$ is real per capita investment? This equation says that the total consumption plus total investment should equal total income in a closed economy without government intervention.

Assume that Assumption 11 holds, which says that $\boldsymbol{e_i}$ is i.i.d. over time with zero mean, $\boldsymbol{\sigma^2}$ variance and $\boldsymbol{I_t}$ *and* $\boldsymbol{e_t}$ are independent ($\boldsymbol{E\{I_t, e_t\} = 0}$). This assumption says that investment $\boldsymbol{I_t}$ is exogeneous and determined independently of the error term. In contrast, both $\boldsymbol{Y_t}$ and $\boldsymbol{C_t}$ are endogeneous variables, which are jointly (simultaneously) determined in the model. The above model in Eq. 10.16, and Eq. 10.17, is a simple simultaneous model in structural form (structural model).

$\boldsymbol{Y_t}$ is endogenous in the consumption function Eq.10.16. Because $\boldsymbol{C_t}$ influences $\boldsymbol{Y_t}$ through Eq.10.17. We can no longer argue that $\boldsymbol{Y_t}$ and $\boldsymbol{e_i}$ are uncorrelated. Consequently, the ***OLS*** estimator for $\boldsymbol{\beta_1}$ will be inconsistent.

To overcome this problem, the reduced form of this model may be considered, in which the endogeneous variables $\boldsymbol{C_t}$ and $\boldsymbol{Y_t}$ are expressed as a function of the exogeneous variable $\boldsymbol{I_t}$ and the error term as follows.

$$Y_t = \frac{\beta_0}{1-\beta_1} + \frac{1}{1-\beta_1} I_t + \frac{1}{1-\beta_1} e_t \qquad (10.18)$$

$$C_t = \frac{\beta_0}{1-\beta_1} + \frac{\beta_1}{1-\beta_1} I_t + \frac{1}{1-\beta_1} e_t \qquad (10.19)$$

The usual assumptions of error term cannot be applied to above reduced form model.

ENDOGENEITY TEST

Is there evidence that correlation between the potentially endogenous variables and the error term is strong enough to result in substantively biased estimates?

Instrument relevance test: Are the instruments sufficiently strongly correlated with the potentially endogenous variables?

Exogeneity/excludability of instruments: Are the instruments genuinely uncorrelated with the main equation residuals?

In the presence of weak instruments, the ***TSLS*** estimator can actually produce worse results than ***OLS***. First step is the identify instruments which are strongly correlated with the endogeneous variables.

Hausman (1978) test can be used for endogeneity of independent variables.

$$H_0: the\ independent\ variable\ is\ exoneous$$
$$H_1: the\ independent\ variable\ is\ endogenous$$

i- Run the first stage regressions using ***OLS*** and save the residuals. This is called reduced form equation

ii- Assume that the initial model is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \qquad (10.20)$$

$y_i: individual´s\ wage$
$x_{1i}: individual´s\ personel\ characteristc$
$x_{2i}: number\ of\ hours\ person\ i\ working$

iii- Include the residuals obtained from first stage regression as additional regressors in the main equation and estimate by using ***OLS***.

iv- Test the joint significance of the first stage residuals.

Instruments should also be exogeneous with regard to the dependent variable in the main equation and excludable from the main regression equation.

The test will analyze whether or not there is a correlation between the independent, and that part of the suspect variables variation that is not explained by exogeneous factors.

SOLUTION

Ad hoc Approaches

- A proxy can be used that does not suffer from the endogeneity problem.

- The suspected variable can be lagged by one or more periods.

This class of solution has some advantages. It is easy to implement and does not require additional variables or data.

But it has also following disadvantages:

- Interpretation of proxy variable is difficult.

- It is difficult to figure out the level of endogeneity concern and adequacy of the solution.

Instrumental Variables Estimation

The best way to deal with endogeneity problem is through instrumental variables (***IV***) techniques and the most common ***IV*** estimator is the Two Stage Least Squares (***TSLS***) estimates.

An exogenous variable (instrument) that is strongly correlated with the potentially endogenous regressor should be found. The instrument only influences the dependent variable through the potentially endogenous independent variable. Selection of appropriate instruments is a crucial issue.

When the model is interpreted as a conditional expectation, the ceteris paribus condition only refers to the included variables, while for a causal interpretation it also includes the unobservables (omitted variables) in the error term.

An instrumental variable z_{2i}, is assumed to be uncorrelated with the model error ε_i but correlated with the endogenous regressor x_2 . The main problem is to find appropriate instrument. Another problem is

that standard errors of instrumental variables estimators are typically quite high compared to those of the ***OLS*** estimator. The most important reason for this is that the instrument and regressor have a low correlation.

Advantages:

- Rigor and transparency;
- Amenability to empirical testing, appropriateness of the instruments;

$$y_i = \beta_1 x_i + e_i \tag{10.21}$$

$\boldsymbol{x_i}$ contains some endogeneous variables

$$x_i = \beta_z Z + \varepsilon_i \tag{10.22}$$

Z is the matrix of all endogenous variables in the model, and the number of instruments should be equal to the number of endogenous variables.

The IV estimate is:

$$\hat{\beta}_z = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} \tag{10.23}$$

The $\boldsymbol{R^2}$ in ***IV*** is not useful since it can also be negative. Interpretation of it is not valid.

Example 10.1

Consider a sub-sample of subjects for whom we wish to understand whether college degree have an effect on wages. Unfortunately, to understand the effect of a college degree, we need to have a proxy for the subjects' intrinsic ability. Intrinsic ability may influence the likelihood of obtaining a college degree.

The intrinsic ability may also influence the wages you obtain. As a result a positive estimate on return on college degree may be attributed.

We wish to understand the impact of education on wages. We are unable to measure individuals non-education based capabilities, which not only influence wages but also the choice and ability to complete a degree. Here these capabilities are the unobserved heterogeneity causing bias in the estimated effect of education on wage. In this case probably a positive impact since they are boosting the estimated effect of education.

The problem using an OLS in cases which suffers from endogeneity is that the error term and the explanatory variables become correlated.

$$Cov(x_i, e_i) \neq 0 \qquad (10.24)$$

This is caused by the unobserved element (omitted variable) since it is hidden in the error term.

Instruments (***zi***) are variables used to explain a variable we suspect of being endogenous and which are exogenous with respect to the main equation;

$$Cov(z_i, u_i) = 0 \qquad (20.25)$$

$$Cov(z_i, x_i) \neq 0 \qquad (10.26)$$

For example[25]; Suppose we explain an individual's log wage $\boldsymbol{y_i}$ by a number of personal characteristics, $\mathbf{x_{1i}}$, as well as the number of hours person **i** is working ***(x2i)*** by means of a linear model as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \qquad (10.27)$$

In the Eq. 10.27, $\boldsymbol{\varepsilon_i}$ includes all unobservable factors that affect a person's wage, including things like 'ability' or 'intelligence'. Typically, it is obvious that the number of hours a person is working partly also

[25] Verbeek, M. (2004)

depends on these unobserved characteristics. If this is the case, ***OLS*** is consistently estimating the conditional expected value of a person's wage given, among other things, how many hours he or she is working, but not consistently estimating the causal effect of working hours. That is, the ***OLS*** estimate for $\boldsymbol{\beta_2}$ would reflect the difference in the expected wage of two arbitrary persons with the same observed characteristics in $\boldsymbol{x_{1i}}$, but working $\boldsymbol{x_2}$ and $\boldsymbol{x_{2+1}}$ hours, respectively.

In the above example, variable that is correlated with working hours is $\boldsymbol{x_{2i}}$ but not correlated with the unobserved 'ability' factors that are included in $\boldsymbol{\varepsilon_i}$. Variables relating to the composition of one's family may serve as instrumental variables.

The assumptions captured in the moment conditions are identifying. So that, they cannot be tested statistically. Besides, over identifying restrictions can be tested when there are more conditions than actually needed for identification.

If the instrument $\boldsymbol{z_{2i}}$ is valid, then the endogeneity of $\boldsymbol{x_{2i}}$ can be tested. Housman (1978) proposes to compare the ***OLS*** and ***IV*** estimators for $\boldsymbol{\beta}$. Assuming $E\{\varepsilon_i z_i\} = 0$, the ***IV*** estimator is consistent. If, in addition, $E\{\varepsilon_i, x_{2i}\} = 0$, the ***OLS*** estimator is also consistent and should differ from the ***IV*** one by sampling error only.

In the first step, estimate a regression explaining $\boldsymbol{x_{2i}}$ from $\boldsymbol{x_{1i}}$ and $\boldsymbol{z_{2i}}$, and save the residuals, save $\boldsymbol{v_i}$. This is the reduced form equation. Next, add the residuals to the model of interest and estimate the following regression by ***OLS***.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + v_i\gamma + e_i \qquad (10.28)$$

If $\gamma = 0$, $\boldsymbol{x_{2i}}$ is exogenous. We can easily test the endogeneity of $\boldsymbol{x_{2i}}$ by performing a standard **t-test** on $\boldsymbol{\gamma = 0}$ in the above regression. This test requires the assumption that the instrument is valid and therefore does not help to determine which identifying moment condition, $E\{x_{2i}, e_i\} = 0\ or, E\{z_{2i}, e_i\} = 0$ is appropriate.

To define the variables correlated with the endogenous variable but uncorrelated with the part of the error term that is due to the unobserved heterogeneity is a big challenge. A good instrument should be correlated with the key independent variable, but not with the main equation dependent variable. If there are four endogenous regressors, there should be at least four different instruments.

In a simultaneous equations context, sufficient instruments should be available in the system (order condition for identification). If there are five exogenous variables in the system that is not included in the equation of interest, there can be up to five endogenous regressors. If there is only one endogenous regressor, there are five different instruments to choose from. It is also possible and advisable to estimate more efficiently by using all the available instruments simultaneously[26].

Two Stage Least Squares

Whether the instruments chosen are valid and ***IV*** estimation is necessary can be checked by the help of ***TSLS***:

- In the first step the reduced form is estimated by ***OLS*** (that is: a regression of the endogenous regressors upon all instruments).

- In the second step the original structural equations are estimated by ***OLS***, while replacing all endogenous variables on the right hand side with their predicted values from the reduced form.

Endogeneity test: Is there evidence that correlation between the potentially endogenous variables and the error term is strong enough to result in substantively biased estimates?

Instrument relevance test: Are the instruments sufficiently strongly correlated with the potentially endogenous variables?

[26] Verbeek, M. (2004)

Exogeneity/excludability of instruments: Are the instruments genuinely uncorrelated with the main equation residuals?

Are the instruments individually statistically significant?

Are their signs and magnitudes sensible?

Are the instruments jointly statistically significant? Look for a high F-statistic.

Weak Instrument

Simple ***OLS*** can produce better results than ***TSLS*** estimator if there is weak instrument issue. Whether the instruments are strongly correlated enough with the potentially endogenous variables should be checked in the first step.

To test for instrument relevance, make sure to run the first stage regressions of the potentially endogenous variables on all of the exogenous variables. The properties of the ***IV*** estimator can be very poor, and the estimator can be severely biased, if the instruments exhibit only weak correlation with the endogenous regressor(s). Even if the sample size is large the normal distribution may provide a very poor approximation to the true distribution of the ***IV*** estimator.

Generalized Method of Moment

This approach estimates the model parameters directly from the moment conditions imposed by the model. These conditions can be linear or nonlinear in the parameters and the number of moment conditions should be at least as many as the number of unknown parameters for identification.

The advantages of the generalized method of moments are:

i- It does not require normality assumption.

ii- It allows for heteroscedasticity of unknown form.

iii- It can estimate parameters of variables even if the model cannot be solved analytically from the first order conditions.

The GMM concept is commonly used to estimate and test asset pricing models. An asset pricing model, for example the CAPM should explain the variation in expected returns for different risky investments. Because some investments are more risky than others, investors may require compensation for bearing this risk by means of a risk premium. This leads to variation in expected returns across different assets[27].

[27] Cochrane (2001).

CHAPTER 11

REGRESSIONS WITH DUMMY VARIABLES

It is also called a binary variable. A dummy variable takes only two values, 0 or 1.

A DUMMY INDEPENDENT VARIABLE

A dichotomous factor can be entered into a regression equation by formulating a dummy regressor, giving 1 for one category of the factor and 0 for the other category. Dummy variables also are a way of turning qualitative variables into quantitative variables. Once the variables are quantitative, then the correlation and regression techniques can be used.

Regression with dummy explanatory or independent variables is extremely common and the interpretation of coefficient estimates is different from the other variables. Regression with dummy explanatory variables is closely related to Analysis of Variance (or ANOVA for short). ANOVA is rarely used in economics, but it is an extremely common tool in other social and physical sciences such as sociology, education, medical statistics, and epidemiology.

Regression with dummy variables is a more general and more powerful tool than ANOVA. If you know how to use and understand regression, then you have no need to learn about ANOVA.

Let's assume following regression model includes a continues independent variable and a dummy independent variable:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D + e_i \quad (11.1)$$

It can be interpreted as an intercept shift as follows:

$$If\ D = 0, then\ y_i = \beta_0 + \beta_1 x_i + e_i \quad (11.2)$$

$$If\ D = 1, then\ y_i = (\beta_0 + \beta_2) + \beta_1 x_i + e_i \quad (11.3)$$

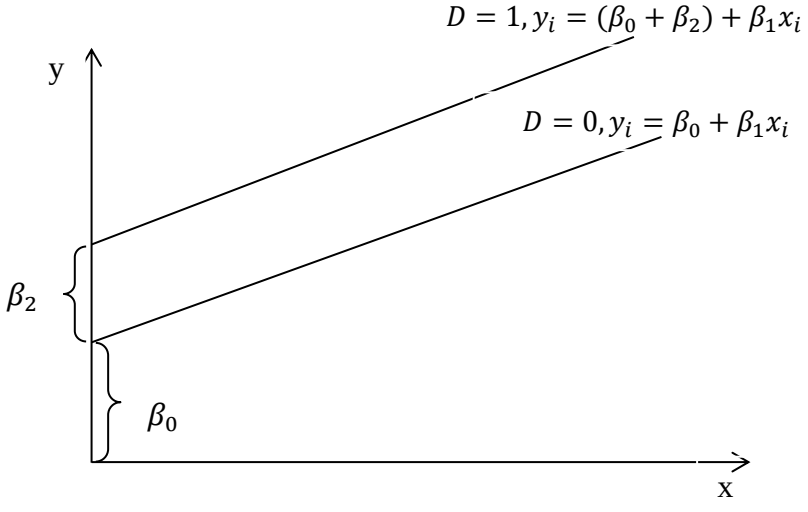


Figure 11.1 Dummy Independent Variable Regression Line

INTERPRETATION

We can say that β is a measure of how much Y tends to change when X is changed by one unit. But, with the present dummy explanatory

variable a "one unit" change implies a change from "No air conditioner" to "Having an air conditioner". That is, we can say "houses with an air conditioner tend to be worth $25,996 more than houses without an air conditioner."

QUALITATIVE VARIABLE WITH MORE THAN TWO CATEGORIES (POLYTOMOUS FACTORS)

Dummy variables can be used as a control in multiple categories.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D_1 + \beta_3 D_2 + \beta_4 D_3 + e_i \qquad (11.4)$$

Example 11.1 Schooling[28]

Table 11.1 Eviews Output for Regression With Dummy Variable

Dependent Variable: LWAGE76
Method: Least Squares
Included observations: 3010

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
ENROLL76	*-0.070539*	*0.026723*	*-2.639635*	*0.0083*
EXP76	*0.053647*	*0.007278*	*7.371461*	*0.0000*
EXP762	*-0.002453*	*0.000355*	*-6.905773*	*0.0000*
BLACK	*-0.326035*	*0.018211*	*-17.90325*	*0.0000*
C	*6.103856*	*0.034063*	*179.1914*	*0.0000*

R-squared	*0.111505*	*Mean dependent var*	*6.261832*
Adjusted R-squared	*0.110323*	*S.D. dependent var*	*0.443798*
S.E. of regression	*0.418602*	*Akaike info criterion*	*1.097867*
Sum squared resid	*526.5590*	*Schwarz criterion*	*1.107850*
Log likelihood	*-1647.290*	*Hannan-Quinn criter.*	*1.101458*
F-statistic	*94.28118*	*Durbin-Watson stat*	*1.715153*
Prob(F-statistic)	*0.000000*		

We can, however, say that β_4 = -0,326 is a measure of the impact of the race on the wage. In other words, if we compare two workers

[28] Verbeek, M. (2004)

with the same experience and number of schooling, black people will always get $0,326 lower wage than the white people with same personal characteristics.

Example 11.2

Suppose everyone from your data is a University dropout, bachelor graduates or master graduates. In order to make comparisons between bachelor and master graduates with university dropouts, you need to include two dummy variables, as follows:

UNGRAD= 1 if university graduates only,
MGRAD=1 if master graduates.

If regression equation includes more dummy variables:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 UNGRAD + \beta_3 MGRAD + e_i \quad (11.5)$$

Example 11.3

Duncan data: Regressing occupational prestige on income and education produces the following regression equation:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 D_1 + \beta_4 D_2 + e_i \quad (11.6)$$

Dependent variable:

$$y_i \; is \; prestige$$

Quantitative independent variables:

$$x_1 \; is \; income \; and \; x_2 \; is \; education$$

Qualitative independent variable:

type (bc, prof, wc)

The *three*-category occupational-type factor can be represented in the regression equation by introducing *two* dummy regressors, employing the following coding scheme:

Type	D1	D2
Blue collar (bc)	0	0
Professional (prof)	1	0
White collar (wc)	0	1

Note: If there are p categories, we should use p − 1 dummy regressors. Blue collar implicitly serves as the baseline category to which the other occupational-type categories are compared.

For Blue collar:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e_i \quad (11.7)$$

For Professional:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 + e_i \quad (11.8)$$

$$y_i = (\beta_0 + \beta_3) + \beta_1 x_1 + \beta_2 x_2 + e_i \quad (11.9)$$

For white collar:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 + e_i \quad (11.10)$$

$$y_i = (\beta_0 + \beta_4) + \beta_1 x_1 + \beta_2 x_2 + e_i \quad (11.11)$$

CONSTRUCTING INTERACTION REGRESSORS

Two explanatory variables may *interact* in determining a response variable when the partial effect of one depends on the value of the other. If these regressions are *not* parallel, then the factor interacts with one or more of the quantitative explanatory variables. The dummy-regression model may be constructed to reflect these interactions.

Together with the quantitative regressor ***x*** and the dummy regressor ***D***, an interaction regressor ***xD*** may be included into the regression equation. The interaction regressor is the product of the other two regressors; although ***xD*** is a function of ***x*** and ***D***, not linear, and perfect collinearity is avoided.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D_1 + \beta_3 D_1 * x_i + \beta_4 D_2 + e_i \qquad (11.12)$$

It causes a change in the slope of the estimated regression line.

Example 11.4

Table 11.2 Eviews Output for Regression with Interaction Variable

Dependent Variable: PRICE
Method: Least Squares
Sample: 1 546
Included observations: 546

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
LOTSIZE	*5.336944*	*0.404851*	*13.18248*	*0.0000*
BEDROOMS	*5829.629*	*1195.020*	*4.878270*	*0.0000*
BATHRMS	*18959.39*	*1764.674*	*10.74385*	*0.0000*
RECROOM	*16466.72*	*7353.490*	*2.239306*	*0.0255*
C	*-2713.233*	*3831.495*	*-0.708140*	*0.4792*
*LOTSIZE*RECROOM*	*-1.202176*	*1.230103*	*-0.977298*	*0.3289*

R-squared	*0.505477*	*Mean dependent var*	*68121.60*
Adjusted R-squared	*0.500898*	*S.D. dependent var*	*26702.67*
S.E. of regression	*18864.68*	*Akaike info criterion*	*22.53890*
Sum squared resid	*1.92E+11*	*Schwarz criterion*	*22.58618*
Log likelihood	*-6147.119*	*Hannan-Quinn criter.*	*22.55738*
F-statistic	*110.3921*	*Durbin-Watson stat*	*1.411099*
Prob(F-statistic)	*0.000000*		

Other type of dummy variables is a seasonal dummy, structural breaks, or shocks.

Why is a qualitative independent variable needed?

i- If the research aims to find out the effect of a qualitative independent variable (For example: Do men earn more than women?)

ii- To better predict/describe the dependent variable. The errors can be made smaller by including variables like gender, race, etc.

iii- Qualitative variables may be confounding factors. Omitting them can cause biased estimates of other coefficients.

DUMMY DEPENDENT VARIABLE

In the case of dummy dependent variable, there are better estimation methods than OLS. The two main alternatives are termed Logit and Probit.

TYPES OF VARIABLES[29]

Continuous or Quantitative Variables

Interval - Scale Variables

Interval scale variables take on positive or negative values. The intervals which are equally ordered keep the same importance throughout the whole scale. They can be used for quantification and comparision the magnitudes of of differences. For instance, 40°C is higher than 30°C, and an increase from 30°C to 40°C is twice as much as the increase from 30°C to 35°C. Counts are also interval scale measurements (number of publications or citations, years of education, etc).

[29] http://www.unesco.org

Continuous Ordinal Variables

When a scale is transformed by an exponential, logarithmic or any other nonlinear transformation, it loses its interval - scale property. then, the observations should be ordered by their ranks.

Ratio - Scale Variables

Ratio - Scale variables are continuous positive measurements on a nonlinear scale. For instance, the growth of bacterial population (say, with a growth function Ae^{Bt}). In this model, equal time intervals multiply the population by the same ratio.

Qualitative or Discrete Variables

Discrete variables which are also called categorical variables take a finite number of numerical values, categories or codes and classified as follows;

- Nominal variables
- Ordinal variables
- Dummy variables
- Preference variables
- Multiple response variables

Nominal Variables

Nominal variables which are used for qualitative classification has no order and measured only in terms of certain category. For instance;

Gender:

1. Male 2.Female

Marital Status:

1. Single 2. Married 3. Divorcee 4. Widower

Ordinal Variables

An ordinal variable is a kind of nominal variable. Three-, five-, or seven- point scales answers are used for evaluation of relative

magnitude of quality, importance or relevance in social and behavioral research.

For instance;

1. The economic status of families in the society might be 'upper lower' is lower than 'middle', but 'how much higher' is not known.
2. A question in a questionnaire related with the time involvement of employees in the social activities to measure commitment and satisfaction. The respondents show their involvement by selecting one of the following answers:

    1 = Very low
    2 = Low
    3 = Medium
    4 = High
    5 = Very high

The variable Involvement is an ordinal variable with 5 points scale.

Ordinal variables can be treated as nominal variables or scale variables.

Categorical Variables

A categorical variable can be obtained from quantitative variables by recoding them. For instance, the quantitative variable *Income* can be classified into four intervals such as;

[Up to 500 KM]	1
[500 KM, 1000 KM]	2
[1000 KM, 2000 KM]	3
[2000 KM, 4000 KM]	4
[Above 4000 KM]	5

Preference Variables

Preference variables are the values either in a decreasing or increasing order. For instance, in a survey, a question may be asked

to respondent to grade the importance of the predetermined difficulties of doing research by using the number codes from 1 to 5 for the most important difficulty to the least important difficulty.

Multiple Response Variables

Multiple response variables can take more than one value. For instance, a survey question regarding the purpose of using computers in research. The respondents could score more than one category.

CHAPTER 12

TIME SERIES

STATIC MODELS

Suppose that we have time series data for two variables dated contemporaneously. A static model can be written as follows:

$$y_t = \beta_0 + \beta_1 x_t + e_i \quad (12.1)$$

Static models are used to model contemporaneous relationship between two variables when a change in ***x*** at time **t** is believed to have an immediate effect on ***y.*** Static regression models are also used when exploring the tradeoff between two variables.

Example 12.1

Static Phillips curve, defined by:

$$\boldsymbol{Inf_t = \beta_0 + \beta_1 Unemp_t + e_t} \quad (12.2)$$

Where, $\boldsymbol{Inf_t}$ is the inflation rate, and $\boldsymbol{Unemp_t}$ is the unemployment rate. This form of the Phillips curve assumes a constant natural rate of unemployment and constant inflationary expectations, and it can be used to study the contemporaneous tradeoff between them. [Mankiw (1994, Section 11.2).]. There can be several explanatory variables in a static regression model.

DYNAMIC MODELS

One way to model the dynamic relationships is to include lagged values of regressors on the right hand side of the regression equation; this is the basis of the distributed- lag model. The distributed-lag model takes the form:

$$\boldsymbol{y_t = \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \cdots + \beta_k x_{t-k} + e_t} \qquad (12.3)$$

Given the model above, immediate impact on $\boldsymbol{y}$ is given as $\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1}$, the impact on $\boldsymbol{y}$ after one period, while $\boldsymbol{\beta_2}$ is the impact on $\boldsymbol{y}$ after two periods. The final impact on $\boldsymbol{y}$ is $\boldsymbol{\beta_k}$ that occurs after $\boldsymbol{k}$ periods. It takes $\boldsymbol{k}$ periods for the full effects of the impulse to be completed. The sequence of coefficients constitutes the impulse response function.

TIME SERIES

Time Series Regression with Lag

The researcher which is using time series data faces two problems which do not exist in the cross sectional data:

i- One time series variable can influence another with a time lag; and

ii- If the variables are non-stationary, spurious regression may arise.

Non-stationary time series variables should be transformed into stationary before running a regression using unit root tests for every variable in the regression equation.

The value of the dependent variable at a given point of time can depend not only on the value of the explanatory variable at that time period, but also on values of the explanatory variable in the past. The

simplest model to incorporate dynamic effects is known as the distributed lag model.

A regression model helps in measuring the effect of one or more independent variables on the dependent variable. In the case of time series data, the effect of some explanatory variables on the dependent variable may take time.

Example 12.2[30]

Table 12.1 Eviews Output for Time Series Regression with Lags

Dependent Variable: Y
Method: Least Squares
Sample (adjusted): 2005M07 2010M02
Included observations: 56 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X	-131.9943	47.43609	-2.782571	0.0076
X__1_	-449.8597	47.55659	-9.459460	0.0000
X__2_	-422.5183	46.77785	-9.032445	0.0000
X__3_	-187.1041	47.64089	-3.927384	0.0003
X__4_	-27.77104	47.66190	-0.582668	0.5627
C	91173.32	1949.850	46.75914	0.0000

R-squared	0.759855	Mean dependent var	74067.00
Adjusted R-squared	0.735840	S.D. dependent var	10468.82
S.E. of regression	5380.606	Akaike info criterion	20.11995
Sum squared resid	1.45E+09	Schwarz criterion	20.33695
Log likelihood	-557.3585	Hannan-Quinn criter.	20.20408
F-statistic	31.64143	Durbin-Watson stat	2.240889
Prob(F-statistic)	0.000000		

The result for time series analysis with lag is reported in **Table 12.1**. Inflation in Turkey is effected by its own past values up to three lags. The inflation four months ago does not have any impact on this month´s inflation level.

[30] Koop, G. (2005).

Lag Selection

If the lag length is not clear, there are many different approaches to lag length selection in the econometrics literature.

Step1. Choose the maximum possible lag length, q_{max}, that seems reasonable to you.

Step2. Estimate the distributed lag model

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + \cdots + \beta_{qmax+1} x_{t-qmax} + e_i \quad (12.4)$$

If the ***p-value*** for testing β_{qmax+1} is less than the significance level you choose (e.g. 0.05) then go no further. Use q_{max} as lag length. Otherwise go on to the next step,

Step3. Estimate the distributed lag model

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + \cdots + \beta_{qmax} x_{t-qmax+1} + e_i \quad (12.5)$$

If the p-value for testing β_{qmax} is less than the significance level you choose (e.g. 0.05) then do not go further. Use q_{max} ***-1*** as lag length. Otherwise go on to the next steps

Example 12.3

We estimate that five months is the maximum time period to expect that training may impact on losses due to accidents. We need to start with estimating a distributed lag model with lag length equal to 5 (q_{max}=5). Results are given in **Table 12.2**.

Table 12.2 Eviews Output for Distributed Lag Model

Dependent Variable: Y				
Method: Least Squares				
Included observations: 55 after adjustments				
Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
X	*-145.3045*	*48.18577*	*-3.015506*	*0.0041*
*X__1_*	*-461.8515*	*48.21403*	*-9.579192*	*0.0000*
*X__2_*	*-423.3032*	*47.84485*	*-8.847413*	*0.0000*
*X__3_*	*-199.4649*	*48.25000*	*-4.133988*	*0.0001*
*X__4_*	*-36.26030*	*48.27840*	*-0.751067*	*0.4563*
*X__5_*	*5.332677*	*47.95305*	*0.111206*	*0.9119*
C	*91903.04*	*2207.578*	*41.63072*	*0.0000*
R-squared	*0.770395*	*Mean dependent var*		*74023.99*
Adjusted R-squared	*0.741695*	*S.D. dependent var*		*10560.31*
S.E. of regression	*5367.145*	*Akaike info criterion*		*20.13239*
Sum squared resid	*1.38E+09*	*Schwarz criterion*		*20.38787*
Log likelihood	*-546.6408*	*Hannan-Quinn criter.*		*20.23119*
F-statistic	*26.84252*	*Durbin-Watson stat*		*2.235865*
Prob(F-statistic)	*0.000000*			

Since the ***p-value*** corresponding to the explanatory variable x_{t-5} is greater than 0.05 we cannot reject null hypothesis that $\boldsymbol{\beta_5 = 0}$ at the 5% significance level. Accordingly we drop this variable from the model and re-estimate with lag length set equal to 4; providing the result in Table **12.3**.

Table 12.3 Eviews Output for Distributed Lag Model

Dependent Variable: Y				
Method: Least Squares				
Sample (adjusted): 5 60				
Included observations: 56 after adjustments				
Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
X	*-131.9943*	*47.43609*	*-2.782571*	*0.0076*
*X__1_*	*-449.8597*	*47.55659*	*-9.459460*	*0.0000*
*X__2_*	*-422.5183*	*46.77785*	*-9.032445*	*0.0000*
*X__3_*	*-187.1041*	*47.64089*	*-3.927384*	*0.0003*
*X__4_*	*-27.77104*	*47.66190*	*-0.582668*	*0.5627*
C	*91173.32*	*1949.850*	*46.75914*	*0.0000*
R-squared	*0.759855*	*Mean dependent var*		*74067.00*
Adjusted R-squared	*0.735840*	*S.D. dependent var*		*10468.82*
S.E. of regression	*5380.606*	*Akaike info criterion*		*20.11995*
Sum squared resid	*1.45E+09*	*Schwarz criterion*		*20.33695*
Log likelihood	*-557.3585*	*Hannan-Quinn criter.*		*20.20408*
F-statistic	*31.64143*	*Durbin-Watson stat*		*2.240889*
Prob(F-statistic)	*0.000000*			

We also should drop the explanatory variable x_{t-5} from the model and re-estimate with lag length equal to 3, giving the result in **Table 12.4.**

Table 12.4 Eviews Output for Distributed Lag Model

Dependent Variable: Y
Method: Least Squares
Sample (adjusted): 4 60
Included observations: 57 after adjustments

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
X	*-125.9000*	*46.24049*	*-2.722722*	*0.0088*
*X__1_*	*-443.4918*	*45.88164*	*-9.665999*	*0.0000*
*X__2_*	*-417.6089*	*45.73324*	*-9.131409*	*0.0000*
*X__3_*	*-179.9043*	*46.25205*	*-3.889650*	*0.0003*
C	*90402.22*	*1643.183*	*55.01653*	*0.0000*

R-squared	*0.757447*	*Mean dependent var*	*74153.74*
Adjusted R-squared	*0.738789*	*S.D. dependent var*	*10395.57*
S.E. of regression	*5313.050*	*Akaike info criterion*	*20.07735*
Sum squared resid	*1.47E+09*	*Schwarz criterion*	*20.25657*
Log likelihood	*-567.2045*	*Hannan-Quinn criter.*	*20.14700*
F-statistic	*40.59659*	*Durbin-Watson stat*	*2.234934*
Prob(F-statistic)	*0.000000*		

The ***p-value*** for testing $\boldsymbol{\beta_3}$ is much less than 0.05. We therefore conclude that the lag three is belonging to the model. Hence q=3 is the lag length we choose for this model reported in **Table 12.4**.

Distributed lag models have the dependent variable depending on an explanatory variable and time lags of the explanatory variable. If the variables in the distributed lag model are stationary, then ***OLS*** estimates are reliable and the statistical techniques of multivariate regressions (e.g. looking at ***p-values*** or confidence intervals) can be used in the interpretation of the estimation results. The lag length in a distributed lag model can be selected by sequentially using ***t-tests*** beginning with a reasonable large lag length.

Finite Distributed Lag Models

In a finite distributed lag (FDL) model, we allow one or more variables to affect ***y*** with a lag. For example, for annual observations, consider the model below:

$$gfr_t = \beta_0 + \beta_1 pe_t + \beta_2 pe_{t-1} + \beta_3 pe_{t-2} + e_t \qquad (12.6)$$

Where gfr_t is the general fertility rate and pe_t is the real dollar value of the personnel tax exemption. The objective is to see whether, in aggregate, the decision to have children is linked to tax value of having a child. The decision would not be based on the immediate impact from the changes in the personnel exemption.

All of the functional forms can be used in time series regressions. When using time series data in a regression, the relationship between ***y*** and ***x*** may be concurrent or ***x*** may serve as a leading indicator. Past values of ***x*** appear as a predictor, either with or without the current value of ***x***.

Example 12.5

The effect of advertising on sales takes time to be analyzed and it is cumulative. It can be modeled as follows:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + \beta_3 x_{t-2} + \beta_4 x_{t-3} + e_t \qquad (12.7)$$

Where, x_t is advertising in the current month and the lagged variables x_{t-1}, x_{t-2} *and* x_{t-3} represents advertising in the two previous months.

Autoregressive Modeling (AR)

It is a regression model where the independent variables are lags of the dependent variable (an autoregression is a regression of a variable on lags of itself).

Let's start with an autoregressive model with the explanatory variable which is one period lag of dependent variable. This is called the ***AR(1)*** model:

$$y_t = \alpha + \varphi y_{t-1} + e_t \ for\ t = 2, \ldots, \quad (12.8)$$

$\varphi = 1$ Implies the type of trend behavior is non-stationary. The other values of φ imply stationary behavior. This allows us to provide a formal definition of the concepts of stationarity and nonstationarity, at least for the AR(1) model: For the AR(1) model, we can say that *Y is stationary* if | ϕ | < 1 and is non-stationary if ***φ = 1***. The other possibility, | ϕ | > 1, is formally, "non-stationary" merely means "anything that is not stationary. At this stage it is useful to think of a unit root as implying ***φ = 1*** in the AR(1) model.

Stationarity

Following are different ways to test whether a time series variable, ***y,*** is stationary or has a unit root:

i- In the ***AR(1)*** model, if ***φ = 1***, then ***y*** has a unit root. If ***|φ| < 1*** then ***y*** is stationary.

ii- If ***y*** has a unit root, then its autocorrelations with past values is close to one and does not change as lag length increases.

iii- Non-stationary time series have a long memory, bur stationary time series do not have long memory.

iv- Non-stationary time series have trend behavior.

v- If ***y*** has a unit root, then the first difference of ***y, Δy,*** is mostly stationary. Non-stationary time series are often referred to as difference stationary series.

vi- If we subtract y_{t-1} from both sides of the equation in the ***AR(1)*** model, as follows:
if ***φ = 1***, then ***ρ = 0*** and the previous equation is written in terms of Δy_t. It means that the first difference, Δy_t fluctuates randomly around ***α***. Whether a series has a unit root can be tested for ***ρ= 0***. Furthermore, a time series is stationary if ***-1 < φ < 1*** which is equivalent to ***-2 <ρ < 0***. This is called as the stationarity condition.

Example 12.6

Table 12.6 Eview Outputs for Unit Root Test

Dependent Variable: CROBEX
Method: Least Squares
Sample (adjusted): 1997M10 2012M12
Included observations: 183 after adjustments

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
C	*29.85194*	*26.16873*	*1.140748*	*0.2555*
CROBEX(-1)	*0.984848*	*0.012520*	*78.65950*	*0.0000*

R-squared	*0.971578*	*Mean dependent var*	*1791.031*
Adjusted R-squared	*0.971421*	*S.D. dependent var*	*1083.960*
S.E. of regression	*183.2469*	*Akaike info criterion*	*13.27041*
Sum squared resid	*6077875.*	*Schwarz criterion*	*13.30549*
Log likelihood	*-1212.243*	*Hannan-Quinn criter.*	*13.28463*
F-statistic	*6187.316*	*Durbin-Watson stat*	*1.723101*
Prob(F-statistic)	*0.000000*		

Null Hypothesis: CROBEX has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic based on SIC, MAXLAG=13)

		t-Statistic	*Prob.**
Augmented Dickey-Fuller test statistic		*-1.210194*	*0.6699*
Test critical values:	*1% level*	*-3.466176*	
	5% level	*-2.877186*	
	10% level	*-2.575189*	

**MacKinnon (1996) one-sided p-values.*
Augmented Dickey-Fuller Test Equation
Dependent Variable: D(CROBEX)
Method: Least Squares
Included observations: 183 after adjustments

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
CROBEX(-1)	*-0.015152*	*0.012520*	*-1.210194*	*0.2278*
C	*29.85194*	*26.16873*	*1.140748*	*0.2555*

R-squared	*0.008027*	*Mean dependent var*	*2.755792*

Adjusted R-squared	0.002546	S.D. dependent var	183.4806
S.E. of regression	183.2469	Akaike info criterion	13.27041
Sum squared resid	6077875.	Schwarz criterion	13.30549
Log likelihood	-1212.243	Hannan-Quinn criter.	13.28463
F-statistic	1.464570	Durbin-Watson stat	1.723101
Prob(F-statistic)	0.227783		

Null Hypothesis: D(CROBEX) has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic based on SIC, MAXLAG=13)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-11.74614	0.0000
Test critical values:	1% level	-3.466377	
	5% level	-2.877274	
	10% level	-2.575236	

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(CROBEX,2)
Method: Least Squares
Included observations: 182 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
D(CROBEX(-1))	-0.868009	0.073897	-11.74614	0.0000
C	2.558105	13.55653	0.188699	0.8505

R-squared	0.433912	Mean dependent var	-0.124615
Adjusted R-squared	0.430767	S.D. dependent var	242.3693
S.E. of regression	182.8616	Akaike info criterion	13.26626
Sum squared resid	6018904.	Schwarz criterion	13.30147
Log likelihood	-1205.230	Hannan-Quinn criter.	13.28054
F-statistic	137.9719	Durbin-Watson stat	2.018993
Prob(F-statistic)	0.000000		

Unit root test result given in **Table 12.6** indicate that Croatian stock market index is stationary in first difference (I(1)).

Random Walk

Consider the case where $\varphi = 1$ (or, equivalently, $\rho = 0$) and $\alpha = 0$. In this case the ***AR (1)*** model can be written as:

$$y_t = y_{t-1} + e_t \quad (12.9)$$

This is referred to as the ***random walk*** model. Since ***φ = 1***, ***y*** has a unit root and is non-stationary. Random walk model is commonly thought to hold for stock prices and foreign exchange rates. The price of a stock today is the price of a stock yesterday plus an (unpredicTable) error term. If stock prices do not follow a random walk, then the change in stock price becomes predicTable and investors have arbitrage possibilities.

The non-stationary time series variables contain a unit root. These series contain a stochastic trend. If these time series are differenced, stationary time series are obtained. For this reason, they are also called difference stationary.

It is used mostly for forecasting. The current value of a dependent variable is defined only by its own past values. p_{th} order Autoregressive model is thus:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} + \cdots + \beta_p y_{t-p} + e_t \quad (12.10)$$

Let us subtract y_{t-1} from both sides of the previous equation. We obtain following equation after rearranging:

$$\Delta y_t = \alpha + \rho y_{t-1} + \gamma_1 \Delta y_{t-1} + \cdots + \gamma_{p-1} \Delta y_{t-p+1} + e_t \quad (12.11)$$

ρ = 0 implies that the ***AR(p)*** time series y contains a unit root; if ***-2 < ρ < 0***, then the series is stationary. In Eq. 12.11, ***ρ = 0*** clarifies the logic of unit root series. If ***ρ = 0*** then the term y_{t-1} will drop out of the equation and only terms involving ***Δy*** or its lags will be present in the regression. "If a unit root is present, then the data can be differenced to obtain stationarity".

There is also another regression model which yields trend behavior. The term ***δt*** in the Eq.(12.10) is called a deterministic trend since it is an exact (deterministic) function of time. In contrast, unit root series contain a so-called stochastic trend.

$$y_t = \alpha + \varphi y_{t-1} + \delta t + e_t \quad (12.12)$$

These series can exhibit trend behavior through the incorporation of a deterministic trend and they are called as trend stationary. Stationary models with a deterministic trend can yield time series plots that closely resemble those from non-stationary models having a stochastic trend. Looking at time series plots alone is not enough to tell whether a series has a unit root or not.

Example 12.7

Table 12.7 Eviews Output for Wald Coefficient Restriction Test

Wald Test: *Equation: Untitled*			
Test Statistic	*Value*	*df*	*Probability*
F-statistic	*1.464570*	*(1, 181)*	*0.2278*
Chi-square	*1.464570*	*1*	*0.2262*
Null Hypothesis Summary:			
Normalized Restriction (= 0)		*Value*	*Std. Err.*
-1 + C(2)		*-0.015152*	*0.012520*
Restrictions are linear in coefficients.			

Wald coefficient restriction test reported in **Table 12.7** indicate that coefficient of first lagged variable is not different than 1 meaning that our model follows the random walk.

Seasonality

Seasonal dummies can also be used as explanatory variables. These dummy variables may be included as additional explanatory variables in the ***AR (p)*** with deterministic trend model.

For instance, with quarterly data, you can create the dummy variables:
(1) D1 = 1 if an observation is from the first quarter (= 0 otherwise);

(2) D2 = 1 if an observation is from the second quarter (= 0 otherwise); and
(3) *D3 = 1 if* an observation is from the third quarter (= 0 otherwise).

If the ***AR(p)*** with deterministic trend model includes seasonal dummies, ***OLS*** still provides good estimates of all coefficients and related statistical tests can be used.

Example 12.8[31]

Table 12.8 Eviews Output for Seasonality Analysis

Dependent Variable: DCROBEX
Method: Least Squares
Sample (adjusted): 1997M11 2012M12
Included observations: 182 after adjustments

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
C	*-0.010373*	*14.24460*	*-0.000728*	*0.9994*
DCROBEX(-1)	*0.138657*	*0.074865*	*1.852084*	*0.0657*
D01	*28.98210*	*48.49478*	*0.597633*	*0.5508*

R-squared	*0.019372*	*Mean dependent var*	*2.966044*
Adjusted R-squared	*0.008415*	*S.D. dependent var*	*183.9647*
S.E. of regression	*183.1890*	*Akaike info criterion*	*13.27526*
Sum squared resid	*6006918.*	*Schwarz criterion*	*13.32807*
Log likelihood	*-1205.049*	*Hannan-Quinn criter.*	*13.29667*
F-statistic	*1.768029*	*Durbin-Watson stat*	*2.031438*
Prob(F-statistic)	*0.173636*		

The coefficient of the dummy variable defined for possible seasonality in the model is not significant as reported in **Table 12.8**. It means that the model does not include end of the year effect.

Steps for Testing in the AR(*p*) with Deterministic Tren

Step 1. Choose the maximum lag length, (pmax) that seems reasonable to you;
Step 2. Estimate using ***OLS*** the ***AR(pmax)*** with deterministic trend model as follows:

[31] End of the year effect on Crobex index

$$\Delta y_t = \alpha + \rho y_{t-1} + \gamma_1 \Delta y_{t-1} + \cdots + \gamma_{pmax-1} \Delta y_{t-pmax+1} + \delta t + e_t \quad (12.13)$$

If the ***p-value*** for testing $\gamma_{pmax-1} = 0$ is less than the chosen significance level (e.g. 0.10, 0.05, 0.001) then go to Step 5, using pmax as lag length. Otherwise go on to the next step.

Step 3. Estimate the following ***AR(pmax - 1)*** *model:*

$$\Delta y_t = \alpha + \rho y_{t-1} + \gamma_1 \Delta y_{t-1} + \cdots + \gamma_{pmax-2} \Delta y_{t-pmax+2} + \delta t + e_t \quad (12.14)$$

If the ***p-value*** for testing $\gamma_{pmax-1} = 0$ is less than the predetermined significance level (e.g. 0.10, 0.05 or 0.001) then go to Step 5, using pmax - 1 as lag length. Otherwise go on to the next step.

Step 4. Repeatedly estimate lower order ***AR*** models until you find an ***AR(p)*** model where γ_{p-1} is statistically significant (or run out of lags).

Step 5. Now test for whether the deterministic trend should be omitted. If the ***p-value*** for testing ***δ = 0*** is greater than the chosen significance level then drop the deterministic trend variable.

Example 12.9

Table 12.9 Eviews Output for Testing in The AR(P) with Deterministic Trend Model

Dependent Variable: CROBEX		*Method: Least Squares*		
Sample (adjusted): 1998M06 2012M12		*Included observations: 175 after adjustments*		
Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
C	*-9496.055*	*7984.226*	*-1.189352*	*0.2360*
DATUM	*0.013075*	*0.010929*	*1.196377*	*0.2333*
CROBEX(-1)	*0.958370*	*0.016081*	*59.59740*	*0.0000*
DCROBEX(-1)	*0.159776*	*0.075370*	*2.119878*	*0.0355*
DCROBEX(-2)	*0.087136*	*0.076015*	*1.146303*	*0.2533*
DCROBEX(-3)	*0.118046*	*0.074944*	*1.575127*	*0.1172*
DCROBEX(-4)	*-0.066389*	*0.074777*	*-0.887828*	*0.3759*
DCROBEX(-5)	*0.183187*	*0.074444*	*2.460729*	*0.0149*
DCROBEX(-6)	*-0.141791*	*0.075692*	*-1.873271*	*0.0628*
DCROBEX(-7)	*0.146991*	*0.075739*	*1.940763*	*0.0540*
DCROBEX(-8)	*0.210142*	*0.076206*	*2.757566*	*0.0065*
R-squared	*0.976308*	*Mean dependent var*		*1828.422*
Adjusted R-squared	*0.974864*	*S.D. dependent var*		*1093.873*
S.E. of regression	*173.4277*	*Akaike info criterion*		*13.21019*
Sum squared resid	*4932654.*	*Schwarz criterion*		*13.40912*
Log likelihood	*-1144.892*	*Hannan-Quinn criter.*		*13.29088*
F-statistic	*675.8239*	*Durbin-Watson stat*		*2.008453*
Prob(F-statistic)	*0.000000*			

Result reported in **Table 12.9** shows that the model does not include deterministic trend.

Lagged Dependent Variables

Sometimes a past value of *y* is used as a predictor as well. A relationship of this type might be:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + \beta_3 x_{t-1} + \beta_4 x_{t-2} + \beta_5 x_{t-3} + e_t \qquad (12.15)$$

For instance, this month's sales are defined by four months of advertising expense plus last month's sales.

Example 12.10

Table 12.10 Eviews Output for Lagged Dependent Variables in The Model

Dependent Variable: LCPI
Method: Least Squares
Sample (adjusted): 3 192
Included observations: 190 after adjustments

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
C	*0.120554*	*0.020397*	*5.910510*	*0.0000*
LCPI(-1)	*1.503330*	*0.063013*	*23.85761*	*0.0000*
LCPI(-2)	*-0.506877*	*0.062599*	*-8.097168*	*0.0000*
LIMPORT	*-0.005756*	*0.002572*	*-2.238411*	*0.0264*

R-squared	*0.999901*	*Mean dependent var*	*16.68963*
Adjusted R-squared	*0.999899*	*S.D. dependent var*	*1.250662*
S.E. of regression	*0.012542*	*Akaike info criterion*	*-5.898567*
Sum squared resid	*0.029260*	*Schwarz criterion*	*-5.830208*
Log likelihood	*564.3638*	*Hannan-Quinn criter.*	*-5.870876*
F-statistic	*626343.4*	*Durbin-Watson stat*	*1.918657*
Prob(F-statistic)	*0.000000*		

LCPI = 0.120553908648 + 1.50333021473*LCPI(-1) - 0.506877156909*LCPI(-2) - 0.00575616664906*LIMPORT

Moving Average

The ***MA(1)*** (first order moving average) process is given by the following equation:

$$y_t = \mu + e_t + \alpha e_{t-1} \qquad (12.16)$$

It means that current value of dependent variable, y_1 is a weighted average of current and previous values of error term, e_1 and e_0, and y_2 is a weighted average of e_2 and e_1 .The value of error term in time t, e_t is defined by the white noise process e_t.

The simple moving average model says that observations that deviate two or more periods are not correlated.

$$y_t = \mu + \sum_{j=0}^{\infty} \theta^j e_{t-j} \qquad (12.17)$$

This can be interpreted as the moving average representation of the autoregressive process.

Autoregressive process ***(AR)*** is written as an infinite order moving average processes if $|\theta|<1$.

Moving average representation might be more convenient than an autoregressive in some instances.

Example 12.11

Table 12.10 Eviews Output for Moving Average Method

Dependent Variable: STATEUR
Method: Least Squares
Sample (adjusted): 3 4877
Included observations: 4875 after adjustments

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
C	*7.511632*	*1.09E-05*	*688602.0*	*0.0000*
RES	*0.999995*	*4.36E-06*	*229152.5*	*0.0000*
RES(-1)	*-0.017451*	*4.36E-06*	*-3999.289*	*0.0000*

R-squared	*1.000000*	*Mean dependent var*	*7.511036*
Adjusted R-squared	*1.000000*	*S.D. dependent var*	*2.500353*
S.E. of regression	*0.000762*	*Akaike info criterion*	*-11.52156*
Sum squared resid	*0.002826*	*Schwarz criterion*	*-11.51757*
Log likelihood	*28086.81*	*Hannan-Quinn criter.*	*-11.52016*
F-statistic	*2.63E+10*	*Durbin-Watson stat*	*2.034619*
Prob(F-statistic)	*0.000000*		

STATEUR = 7.51163178608 + 0.999994663914*RES - 0.0174508542575*RES(-1)

United States unemployment rate shows moving average characteristics since the coefficients of the error term and first lag of the error term are statistically significant.

Autocorrelation Function

Defining autocorrelation $\boldsymbol{\rho_k}$ as:

$$\boldsymbol{\rho_k} = \frac{cov\{y_t, y_{t-k}\}}{V\{y_t\}} = \frac{\gamma_k}{\gamma_0} \qquad (12.18)$$

That autocorrelation is a function of $\boldsymbol{k}$ is referred as autocorrelation function (ACF) or correlogram of the series $\boldsymbol{y_t}$. The ACF is employed to model the dependencies between observations by describing the evolution of $\boldsymbol{y_t}$ over time.

The ACF provides following information;

- The value of the process and its correlation with previous observations.

- The length and power of the process memory.
- How long and how strongly a shock in the process e_t impact the values of y_t.

$$y_t = \delta + \theta y_{t-1} + e_t \quad (12.19)$$

$$\rho_k = \theta^k \quad (12.20)$$

$$y_t = \mu + e_t + \alpha e_{t-1} \quad (12.21)$$

$$\rho_1 = \frac{\alpha}{1+\alpha^2} \; and \; \rho_k = 0 \ldots k = 2, 3, 4 \ldots \quad (12.22)$$

In sum, a shock in an ***MA(1)*** process impacts y_t in two periods only. A shock in the ***AR(1)*** process impacts all future observations with a decreasing affect.[32]

Example 12.12

Table 12.12 Eviews Output for Autocorrelation Function

Sample: 1997M09 2012M12
Included observations: 184

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
.|*******	.|*******	1	0.985	0.985	181.37	0.000
.|*******	*|. |	2	0.966	-0.135	356.75	0.000
.|*******	*|. |	3	0.942	-0.141	524.65	0.000
.|*******	*|. |	4	0.915	-0.112	683.84	0.000
.|******|	.|. |	5	0.889	0.069	834.91	0.000
.|******|	*|. |	6	0.858	-0.178	976.31	0.000
.|******|	.|* |	7	0.828	0.097	1109.0	0.000
.|******|	*|. |	8	0.796	-0.145	1232.2	0.000
.|***** |	**|. |	9	0.757	-0.207	1344.1	0.000
.|***** |	.|. |	10	0.717	-0.010	1445.1	0.000
.|***** |	.|* |	11	0.679	0.164	1536.3	0.000
.|***** |	.|. |	12	0.643	0.023	1618.6	0.000
.|**** |	*|. |	13	0.607	-0.084	1692.2	0.000
.|**** |	*|. |	14	0.568	-0.116	1757.1	0.000
.|**** |	.|* |	15	0.533	0.156	1814.7	0.000
.|**** |	*|. |	16	0.497	-0.070	1865.1	0.000

[32] Verbeek, M. (2004)

.|*** |	.|. |	17	0.462	0.017	1908.8	0.000
.|*** |	.|* |	18	0.430	0.113	1946.9	0.000
.|*** |	.|. |	19	0.401	0.036	1980.3	0.000
.|*** |	.|. |	20	0.378	-0.003	2010.2	0.000
.|*** |	.|. |	21	0.354	-0.015	2036.4	0.000
.|** |	.|. |	22	0.330	-0.009	2059.4	0.000
.|** |	.|. |	23	0.310	0.045	2079.8	0.000
.|** |	.|. |	24	0.292	0.053	2098.1	0.000
.|** |	*|. |	25	0.274	-0.108	2114.3	0.000
.|** |	.|. |	26	0.257	-0.052	2128.6	0.000
.|** |	.|. |	27	0.241	-0.012	2141.3	0.000
.|** |	.|. |	28	0.226	-0.019	2152.5	0.000
.|* |	.|* |	29	0.213	0.099	2162.5	0.000
.|* |	.|. |	30	0.198	-0.060	2171.2	0.000
.|* |	.|. |	31	0.186	-0.039	2178.9	0.000
.|* |	.|. |	32	0.174	-0.012	2185.8	0.000
.|* |	.|. |	33	0.162	-0.009	2191.7	0.000
.|* |	.|. |	34	0.149	0.002	2196.8	0.000
.|* |	.|. |	35	0.135	-0.050	2201.0	0.000
.|* |	.|. |	36	0.122	-0.046	2204.4	0.000

Autoregressive Moving Average (ARMA)

MA(q) : $\boldsymbol{y_t = e_t + \alpha_1 e_{t-1} + \cdots + \alpha_q e_{t-q}}$ (12.23)

AR(p): $\boldsymbol{y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \cdots + \theta_p y_{t-p} + e_t}$ (12.24)

ARMA(p,q) :$\boldsymbol{y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \cdots + \theta_p y_{t-p} + e_t + \alpha_1 e_{t-1} + \cdots + \alpha_q e_{t-q}}$ (12.25)

For more parsimonious representation, we may want to work with an ***ARMA*** model that contains both an ***AR*** and ***MA*** part. The general ***ARMA*** model can be also written as:

$$\boldsymbol{\theta(L) y_t = \alpha(L) e_t} \quad (12.26)$$

$$\boldsymbol{L y_t = y_{t-1}} \quad (12.27)$$

$$\boldsymbol{L^p y_t = y_{t-p}} \quad (12.28)$$

Example 12.13

Table 12.13 Eviews Output for ARMA Model

Dependent Variable: DCROBEX
Method: Least Squares
Sample (adjusted): 1998M03 2012M12
Included observations: 178 after adjustments
Convergence achieved after 96 iterations
MA Backcast: 1997M12 1998M02

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
C	*0.367155*	*19.34648*	*0.018978*	*0.9849*
AR(1)	*-0.911170*	*0.096376*	*-9.454333*	*0.0000*
AR(2)	*-0.868000*	*0.103492*	*-8.387151*	*0.0000*
AR(3)	*-0.421543*	*0.137626*	*-3.062956*	*0.0026*
AR(4)	*0.218220*	*0.102726*	*2.124295*	*0.0351*
AR(5)	*0.282092*	*0.078334*	*3.601121*	*0.0004*
MA(1)	*1.103650*	*0.074023*	*14.90948*	*0.0000*
MA(2)	*1.185603*	*0.039084*	*30.33477*	*0.0000*
MA(3)	*0.866090*	*0.073222*	*11.82832*	*0.0000*

R-squared	*0.174800*	*Mean dependent var*	*3.877303*
Adjusted R-squared	*0.135737*	*S.D. dependent var*	*184.6922*
S.E. of regression	*171.7005*	*Akaike info criterion*	*13.17862*
Sum squared resid	*4982298.*	*Schwarz criterion*	*13.33950*
Log likelihood	*-1163.897*	*Hannan-Quinn criter.*	*13.24386*
F-statistic	*4.474844*	*Durbin-Watson stat*	*2.015767*
Prob(F-statistic)	*0.000061*		

Inverted AR Roots	*.55*	*-.07-.97i*	*-.07+.97i*	*-.66-.32i*
	-.66+.32i			
Inverted MA Roots	*-.11-.99i*	*-.11+.99i*	*-.88*	

ARIMA (p,q)

Stationarity of a stochastic process requires that the variances and autocovariances are finite and independent of time.[33]

[33] Verbeek, M. (2004)

A series which becomes stationary after first differencing is said to be integrated of order one, denoted I(1). If Δy_t is described by a stationary ***ARMA(p,q)*** model, we say that y_t is described by an autoregressive integrated moving average (***ARIMA***) model of order p, 1, q or in short an ***ARIMA (p,1,q)*** model.

First differencing quite often can transform a non-stationary series into a stationary one. If a series must be differenced twice before it becomes stationary, then it is said to be integrated of order two, denoted I(2) and it must have two unit roots.

Predicting with ARMA Models

One of main goal of building a time series model is to predict the future path of the economic variables using past observations. the expected quadratic prediction error is minimized to obtain accurate prediction. The accuracy of the prediction decreases as time horizon of the prediction increase.

In the prediction of one period ahead, the ***MA(1)*** model provides more accurate prediction result. In further ahead predictions for the longer time horizon ***ARMA*** type models give better result. Mostly, the autoregressive representation model is most convenient for the computation of the predictor.

The informational value or short and long run memory contained in an ***AR (1)*** process decreases over time. The forecast error variance increases as the forecast horizon increase.

In fact the parameters in ***ARMA*** models are unknown. Hence, the estimated values of the parameters are used which produce additional uncertainty in predictors. However, this uncertainty is ignored in practise.

Autoregressive Distributed Lag (ARDL) Model

In time-series econometric modeling a dynamic regression will usually include both lagged dependent and independent variables as regressors. The dependent variable might be correlated with its lags.

It means that lags of the dependent variable should be included in the regression model. In this model the dependent variable depends on the lags of itself and the explanatory variables as well as lags of the explanatory variables as follows:

$$\boldsymbol{y_t = \alpha + \theta t + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \varphi_0 x_t + \varphi_1 x_{t-1} + \cdots + \varphi_q x_{t-q} + e_t} \quad (12.29)$$

The model described above is called the autoregressive distributed-lag model, abbreviated as ***ARDL(p; k).*** This model also include deterministic trend (***t***). Since the model includes ***p*** lags of ***y*** and ***q*** lags of ***x***, we can write it as ***ARDL(p,q)***. In order to perform this model, series should have same stationarity properties, either they both are stationary or both have a unit root.

The values of ***p*** and ***k*** (lags numbers of ***y*** and ***x*** used) are chosen:

i- On the basis of the statistical significance of the lagged variables, and

ii- So that the resulting model is well specified (e.g. it does not suffer from serial correlation).

The ***ARDL(1,1)***, or alternatively the first order Dynamic Linear Regression Model, takes the following form:

$$\boldsymbol{y_t = \alpha_0 + \alpha_1 y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + e_t} \quad (12.30)$$

Note that ***y*** is sTable (it will converge to its equilibrium) $\boldsymbol{if\ -1 < \alpha <}$ 1. If the above stability condition is satisfied, the long-run solution (or steady state) of Eq. (12.30) is given by:

$$\boldsymbol{y_t = \frac{\alpha_0}{1-\alpha_1} + \left(\frac{\beta_0+\beta_1}{1-\alpha_1}\right) x_t + \frac{e_t}{1-\alpha_1} = c_0 + c_1 x_t + \frac{e_t}{1-\alpha_1}} \quad (12.31)$$

How can we interpret the coefficients in the ARDL model?

The most common way is through the concept of a multiplier. It is common to focus on the long run or total multiplier.

Let's assume that ***x*** and ***y*** is in an equilibrium or steady state (not changing over time). When ***x*** changes one unit, it is affecting ***y***, which starts to change, eventually settling down in the long run to a new equilibrium value. The difference between the old and new equilibrium values for ***y*** can be interpreted as the long run effect of ***x*** *on* ***y*** and is the ***long run multiplier***. This multiplier is often of great interest for policymakers who want to know the eventual effects of their policy changes in various areas.

The long run multiplier measures the effect of a permanent change in ***x***. ***x*** changes permanently to a new level one unit higher than the original value. The long run multiplier measures the effect of this sort of change.

The long run multiplier does not measure the effect of this type of change. The "marginal effect" interpretation of regression coefficients can be used for such temporary changes. Previously, we were interested in the effect of increasing safety training in one particular month on accident losses. But we did not discuss the effect of increasing safety training permanently. The long run multiplier for the ***ARDL(p, q)*** model is:

$$-\frac{\theta}{\rho} \qquad (12.32)$$

Here, we are assuming that ***x and y*** are stationary. How ***ρ = 0*** in the ***AR(p)*** model implied the existence of a unit root has been studied above. The ***ARDL*** model is not the same as the ***AR*** model. If ***ρ = 0*** then the long run multiplier is infinite. In fact, it can be shown that for the model to be sTable, and then we must have ***ρ > 0.4***. Actually, if ***x*** and ***y*** are stationary, this condition is satisfied.[34]

[34] For instance, the effect of computer purchases on sales

Example 12.14

Table 12.14 Microfit output for ARDL model

A. Autoregressive Distributed Lag Estimates

ARDL(2,0,0,1,0) selected based on Schwarz Bayesian Criterion

```
**************************************************************************
Dependent variable is X1
132 observations used for estimation from 2001M2  to 2012M1
**************************************************************************
Regressor          Coefficient     Standard Error      T-Ratio[Prob]
X1(-1)              1.3772          .071186          19.3462[.000]
X1(-2)              -.40579         .071801          -5.6515[.000]
C                  11.1558         17.2889           .64526[.520]
X2                 -.0063771        .018970          -.33617[.737]
X3                  .57778          .12130           4.7634[.000]
X3(-1)              -.54644         .082772          -6.6017[.000]
X4                  .0021061        .0035208          .59819[.551]
**************************************************************************
R-Squared                .98672   R-Bar-Squared              .98608
S.E. of Regression       50.7255  F-stat.   F( 6, 125)    1547.6[.000]
Mean of Dependent Variable 590.7834  S.D. of Dependent Variable   429.9262
Residual Sum of Squares   321634.5  Equation Log-likelihood    -701.9923
Akaike Info. Criterion    -708.9923  Schwarz Bayesian Criterion  -719.0821
DW-statistic             2.1428
**************************************************************************
```

```
                    Diagnostic Tests
**************************************************************************
*   Test Statistics  *      LM Version       *       F Version        *
**************************************************************************
*                    *                       *                        *
* A:Serial Correlation*CHSQ( 12)= 27.9926[.006]*F( 12, 113)=  2.5344[.005]*
*                    *                       *                        *
* B:Functional Form   *CHSQ(  1)= .039857[.842]*F(  1, 124)= .037452[.847]*
*                    *                       *                        *
* C:Normality         *CHSQ(  2)= 12.4107[.002]*      Not applicable     *
*                    *                       *                        *
* D:Heteroscedasticity*CHSQ(  1)= 42.0608[.000]*F(  1, 130)= 60.7955[.000]*
**************************************************************************
```

A:Lagrange multiplier test of residual serial correlation
B:Ramsey's RESET test using the square of the fitted values
C:Based on a test of skewness and kurtosis of residuals
D:Based on the regression of squared residuals on squared fitted values

B. Estimated Long Run Coefficients using the ARDL Approach

```
      ARDL(2,0,0,1,0) selected based on Schwarz Bayesian Criterion
*******************************************************************************
Dependent variable is X1
132 observations used for estimation from 2001M2  to 2012M1
*******************************************************************************
Regressor          Coefficient      Standard Error       T-Ratio[Prob]
C                  389.9973         750.0502             .51996[.604]
X2                 -.22294          .61821              -.36062[.719]
X3                  1.0957          2.9052               .37715[.707]
X4                 .073627          .10250               .71830[.474]
*******************************************************************************
```

C. Error Correction Representation for the Selected ARDL Model

```
      ARDL(2,0,0,1,0) selected based on Schwarz Bayesian Criterion
*******************************************************************************
Dependent variable is dX1
132 observations used for estimation from 2001M2  to 2012M1
*******************************************************************************
Regressor          Coefficient      Standard Error       T-Ratio[Prob]
dX11                .40579          .071801              5.6515[.000]
dC                 11.1558          17.2889              .64526[.520]
dX2                -.0063771        .018970             -.33617[.737]
dX3                 .57778          .12130               4.7634[.000]
dX4                .0021061         .0035208             .59819[.551]
ecm(-1)            -.028605         .016488             -1.7349[.085]
*******************************************************************************
List of additional temporary variables created:
dX1 = X1-X1(-1)
dX11 = X1(-1)-X1(-2)
dC = C-C(-1)
dX2 = X2-X2(-1)
dX3 = X3-X3(-1)
dX4 = X4-X4(-1)
ecm = X1 -389.9973*C +  .22294*X2  -1.0957*X3  -.073627*X4
*******************************************************************************
R-Squared                .43277   R-Bar-Squared              .40554
S.E. of Regression       50.7255  F-stat.   F( 5, 126)   19.0738[.000]
Mean of Dependent Variable  1.5935  S.D. of Dependent Variable  65.7909
Residual Sum of Squares  321634.5  Equation Log-likelihood    -701.9923
Akaike Info. Criterion   -708.9923  Schwarz Bayesian Criterion  -719.0821
DW-statistic             2.1428
*******************************************************************************
R-Squared and R-Bar-Squared measures refer to the dependent variable
dX1 and in cases where the error correction model is highly
restricted, these measures could become negative.
```

If Dependent and Independent Series are Stationary

OLS estimation of ***ARDL (p,q)*** regression model can be carried out in the standard way. Testing of the significance of the parameters can be done using the ***t-stats*** and ***p-values***. However, interpretation of the results is different from the interpretation of standard case.

Macroeconomic time series are often highly correlated with their lags. This implies that original form of ***ARDL*** face multicollinearity problems. With the rewritten form taking first difference, we will not encounter this problem:

$$\Delta y_t = \alpha + \theta t + \rho y_{t-1} + \gamma_1 \Delta y_{t-1} \dots + \gamma_{p-1} \Delta y_{t-p+1} + \delta x_t + \omega_1 \Delta x_{t-1} + \cdots + \omega_q \Delta x_{t-q+1} + e_t \quad (12.33)$$

The ***OLS*** estimation results interpretation under ceteris paribus condition can still be used in ***ARDL***. Another interpretation concept which is called multiplier is commonly used in interpretation of ***ARDL*** regression results focusing on the long run or total multiplier.

Suppose that $x \; and \; y$ are in an equilibrium or steady state (not changing over time). When x changes one unit, it is affecting y, which starts to change, eventually settling down in the long run to a new equilibrium value. Then the difference between the old and new equilibrium values for ***y*** can be interpreted as the long run effect of ***x*** on ***y*** and is the long run multiplier.

The long run multiplier measures the effect of a permanent change in ***x***. For temporary changes standard interpretation method of marginal effect can be used.

The long run multiplier for the ***ARDL (p,q)*** model is: $-\frac{\delta}{\rho}$

Only the coefficients on $x_t \; and \; y_{t-1}$ are important for the long run behavior.

For the model to be sTable, then we must have $\rho > 0$. in practice, if $x \; and \; y$ are stationary, this condition will be satisfied.

Example 15[35]

If dependent and independent series are non-stationary (Spurious regression)

We assume that $\boldsymbol{x}$ and $\boldsymbol{y}$ have unit roots. In practice, we should test stationarity using ADF, DF, PP etc. We start with the case of regression model without lags, then proceed to ***ARDL (p,q)*** model.

Suppose that we want to estimate following regression:

$$\boldsymbol{y_t = \beta_0 + \beta_1 x_t + e_t} \qquad (12.34)$$

If $\boldsymbol{y}$ and $\boldsymbol{x}$ have unit roots then all the usual regression results might be misleading and incorrect. This is called ***Spurious regression problem***. You should never run a regression of $\boldsymbol{y\ on\ x}$ if the variables have unit roots with the exception of ***cointegration*** method.

COINTEGRATION

Bivariate Cointegration (Engle-Granger Cointegration)

The concept of cointegration was first introduced by Granger (1981) and elaborated further by Engle and Granger (1987), Engle and Yoo (1987, 1991), Phillips and Ouliaris (1990), Stock and Watson (1988), Phillips (1991) and Johansen (1988, 1991, 1994).

$$\boldsymbol{e_t = y_t - \beta_0 - \beta_1 x_t} \qquad (12.35)$$

Written in this way, it is clear that the errors are a linear combination of ***y*** and ***x***. However, ***x*** and ***y*** both exhibit non-stationary unit root behavior such that you would expect the error to also exhibit non-stationary behavior. The error usually have a unit root. Statistically, it is this unit root in the error term that causes the spurious regression problem. However, it is possible that the unit roots in ***y*** and ***x*** "cancel each other out" and that the resulting error is stationary. In this

[35] KOOP, G. (2005)

special case, called cointegration, the spurious regression problem disappears and it is valid to run a regression of ***y*** and ***x***.

In sum, if ***y*** and ***x*** have unit roots, but some linear combination of them is stationary, then we can say that ***y*** and ***x*** are cointegrated.

- If ***y*** and ***x*** have unit roots then they have stochastic trends. However, if they are cointegrated, the error does not have such a trend. In this case, the error will not get too large and ***y*** and ***x*** will not diverge from one another; ***y*** and ***x***, in other words, will trend together. This fact motivates other jargon used to refer to cointegrated time series. You may hear them referred as either having common trends or co-trending.

- If we are talking about an economic model involving an equilibrium concept, ***e*** is the equilibrium error. If ***y*** and ***x*** are cointegrated then the equilibrium error remains small. However, if ***y*** and ***x*** are not cointegrated then the equilibrium error will have a trend and departures from equilibrium and will become increasingly large over time. If such departures from equilibrium occur, then many would hesitate to say that the equilibrium is a meaningful one.

- If ***y*** and ***x*** are cointegrated then there is an equilibrium relationship between them. If they are not, then no equilibrium relationship exists.

- Departures from equilibrium should not be too large and there should always be a tendency to return to equilibrium after a shock occurs. Hence, if an economic model which implies an equilibrium relationship exists between ***y*** and ***x*** is correct, then we should observe ***y*** and ***x*** as being cointegrated.

- If ***y*** and ***x*** are cointegrated then their trends will cancel each other out. If cointegration is present, then not only do we avoid the spurious regression problem, but we also have important economic information.

If ***y*** and ***x*** are cointegrated, no need to worry about spurious regression problem.

Regression of ***y*** on ***x*** is called cointegrating regression and is represented as follows:

$$y_t = \beta_1 x_t + e_t \qquad (12.36)$$

If the linear combination of two non-stationary series is stationary, then cointegration exists between these two variables. The unit roots in ***y*** and ***x*** cancel each other out and the resulting error becomes stationary.

Bivariate Cointegration (Engle-Granger) test has following steps:

i- Run the regression of ***y*** on an intercept and ***x*** and save the residuals.

ii- Carry out a Dickey-Fuller test (Dickey, Fuller, 1979) on the residuals (without including a deterministic trend[36]).

iii- If the unit root hypothesis is rejected then conclude that ***y*** and ***x*** are cointegrated. However, if the unit root is accepted then conclude that cointegration does not occur.

System always returns to equilibrium and, hence, that errors never grow too big. The null hypothesis in the Engle Granger test is no cointegration and it is based on Dickey Fuller unit root test. If the structural breaks occur in the data, Engle Granger test has lower power and can be misleading.

Engle-Granger cointegration test tells whether cointegration is present or not. It doesn't provide any information about how many cointegrating relationships are exist.

[36] If it were included it could mean the errors could be growing steadily over time. This would violate the idea of cointegration.

i- The null hypothesis in the Engle–Granger test is "no cointegration" and we conclude "cointegration is present" if we reject this hypothesis.

ii- The Engle–Granger test has low power and can be misleading if there are structural breaks in the data.

Since there is cointegration, no need to worry about the spurious regressions problem. And, we can proceed to an interpretation of our coefficients without worrying that the ***OLS*** estimates are meaningless.

Example 12.14

Table 12.14 Eviews Output for Unit Root Tests

A. France

Null Hypothesis: FRANCE has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 0 (Automatic based on SIC, MAXLAG=30)

		t-Statistic	*Prob.**
Augmented Dickey-Fuller test statistic		*-1.451148*	*0.8458*
Test critical values:	*1% level*	*-3.960237*	
	5% level	*-3.410881*	
	10% level	*-3.127243*	

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(FRANCE)
Method: Least Squares
Included observations: 4175 after adjustments

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
FRANCE(-1)	*-0.001085*	*0.000748*	*-1.451148*	*0.1468*
C	*1.007433*	*1.146994*	*0.878324*	*0.3798*
TREND(12/30/1988)	*0.000810*	*0.000569*	*1.422733*	*0.1549*

R-squared	*0.000557*	*Mean dependent var*	*0.704256*
Adjusted R-squared	*0.000078*	*S.D. dependent var*	*27.85584*
S.E. of regression	*27.85476*	*Akaike info criterion*	*9.492603*

Sum squared resid	*3237004.*	*Schwarz criterion*	*9.497157*
Log likelihood	*-19812.81*	*Hannan-Quinn criter.*	*9.494214*
F-statistic	*1.162196*	*Durbin-Watson stat*	*1.938805*
Prob(F-statistic)	*0.312900*		

Null Hypothesis: D(FRANCE) has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 0 (Automatic based on SIC, MAXLAG=30)

		t-Statistic	*Prob.**
Augmented Dickey-Fuller test statistic		*-62.67179*	*0.0000*
Test critical values:	*1% level*	*-3.960237*	
	5% level	*-3.410882*	
	10% level	*-3.127243*	

**MacKinnon (1996) one-sided p-values.*

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(FRANCE,2)
Method: Least Squares
Sample (adjusted): 1/01/1991 12/29/2006
Included observations: 4174 after adjustments

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
D(FRANCE(-1))	*-0.969971*	*0.015477*	*-62.67179*	*0.0000*
C	*0.260374*	*1.028061*	*0.253267*	*0.8001*
TREND(12/30/1988)	*0.000162*	*0.000358*	*0.453170*	*0.6504*

R-squared	*0.484982*	*Mean dependent var*	*0.001782*
Adjusted R-squared	*0.484735*	*S.D. dependent var*	*38.80164*
S.E. of regression	*27.85256*	*Akaike info criterion*	*9.492446*
Sum squared resid	*3235717.*	*Schwarz criterion*	*9.497000*
Log likelihood	*-19807.73*	*Hannan-Quinn criter.*	*9.494056*
F-statistic	*1963.877*	*Durbin-Watson stat*	*1.998964*
Prob(F-statistic)	*0.000000*		

We want to explore long run relationship between France and Greece stock market indices. After checking the unit root condition of the series, we proceed to find out whether the linear combinations of this two series are stationary or not. Result reported in Table 12.14

shows that there exists long run relationship between France and Greece stock market indices. Furthermore, the long run multiplier is 1.299. This indicates that, in the long run, an increase in the France stock market index by one unit would cause an increase in the Greece stock market index 1.299 units.

B. GREECE

Null Hypothesis: GREECE has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 1 (Automatic based on SIC, MAXLAG=31)

		t-Statistic	*Prob.**
Augmented Dickey-Fuller test statistic		*-1.758819*	*0.7246*
Test critical values:	*1% level*	*-3.959993*	
	5% level	*-3.410762*	
	10% level	*-3.127172*	

**MacKinnon (1996) one-sided p-values.*
Augmented Dickey-Fuller Test Equation
Dependent Variable: D(GREECE)
Method: Least Squares
Sample (adjusted): 1/03/1989 12/29/2006
Included observations: 4694 after adjustments

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
GREECE(-1)	*-0.001150*	*0.000654*	*-1.758819*	*0.0787*
D(GREECE(-1))	*0.169347*	*0.014392*	*11.76662*	*0.0000*
C	*0.785073*	*1.230216*	*0.638159*	*0.5234*
TREND(12/30/1988)	*0.000958*	*0.000639*	*1.499781*	*0.1337*

R-squared	*0.029146*	*Mean dependent var*	*0.873513*
Adjusted R-squared	*0.028525*	*S.D. dependent var*	*41.93447*
S.E. of regression	*41.33206*	*Akaike info criterion*	*10.28201*
Sum squared resid	*8012112.*	*Schwarz criterion*	*10.28751*
Log likelihood	*-24127.87*	*Hannan-Quinn criter.*	*10.28394*
F-statistic	*46.93212*	*Durbin-Watson stat*	*1.988464*
Prob(F-statistic)	*0.000000*		

Null Hypothesis: D(GREECE) has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 0 (Automatic based on SIC, MAXLAG=31)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		*-57.76232*	*0.0000*
Test critical values:	*1% level*	*-3.959993*	
	5% level	*-3.410762*	
	10% level	*-3.127172*	

**MacKinnon (1996) one-sided p-values.*

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(GREECE,2)
Method: Least Squares
Date: 06/05/13 Time: 18:17
Sample (adjusted): 1/03/1989 12/29/2006
Included observations: 4694 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
D(GREECE(-1))	*-0.831266*	*0.014391*	*-57.76232*	*0.0000*
C	*0.368006*	*1.207416*	*0.304789*	*0.7605*
TREND(12/30/1988)	*0.000153*	*0.000445*	*0.342473*	*0.7320*

R-squared	*0.415633*	*Mean dependent var*	*0.000300*
Adjusted R-squared	*0.415384*	*S.D. dependent var*	*54.06902*
S.E. of regression	*41.34128*	*Akaike info criterion*	*10.28224*
Sum squared resid	*8017396.*	*Schwarz criterion*	*10.28636*
Log likelihood	*-24129.42*	*Hannan-Quinn criter.*	*10.28369*
F-statistic	*1668.243*	*Durbin-Watson stat*	*1.988270*
Prob(F-statistic)	*0.000000*		

Exogenous: Constant
Lag Length: 0 (Automatic based on SIC, MAXLAG=30)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		*-2.993250*	*0.0356*
Test critical values:	*1% level*	*-3.431732*	
	5% level	*-2.862036*	
	10% level	*-2.567077*	

**MacKinnon (1996) one-sided p-values.*

C. Unit Root test for the residual series

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(RESID08)
Method: Least Squares
Date: 06/05/13 Time: 18:10
Included observations: 4175 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
RESID08(-1)	-0.004221	0.001410	-2.993250	0.0028
C	-0.093792	0.756853	-0.123924	0.9014

R-squared	0.002142	Mean dependent var	-0.093794
Adjusted R-squared	0.001903	S.D. dependent var	48.95007
S.E. of regression	48.90346	Akaike info criterion	10.61805
Sum squared resid	9979933.	Schwarz criterion	10.62109
Log likelihood	-22163.18	Hannan-Quinn criter.	10.61913
F-statistic	8.959543	Durbin-Watson stat	1.915883
Prob(F-statistic)	0.002776		

Dependent Variable: GREECE
Method: Least Squares
Date: 06/05/13 Time: 18:19
Sample (adjusted): 12/28/1990 12/29/2006
Included observations: 4176 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-724.2877	21.79627	-33.22990	0.0000
FRANCE	1.299339	0.009047	143.6155	0.0000

R-squared	0.831690	Mean dependent var	2169.725
Adjusted R-squared	0.831649	S.D. dependent var	1308.362
S.E. of regression	536.8283	Akaike info criterion	15.40971
Sum squared resid	1.20E+09	Schwarz criterion	15.41275
Log likelihood	-32173.48	Hannan-Quinn criter.	15.41079
F-statistic	20625.41	Durbin-Watson stat	0.008315
Prob(F-statistic)	0.000000		

GREECE = -724.287661619 + 1.29933920604*FRANCE

The residual of the model is found stationary at the 5% significance level meaning that the model is not spurious and we can accept the model. Greece and France do have long run relationship. The variables in the model are cointegrated or have long run equilibrium relationship.

Error Correction Model (ECM)

If the research aims to explore short run behavior, it is not possible to use the regression of ***y*** on ***x***. In these cases an error correction model (or ECM for short) should be employed.

The Granger Representation Theorem says that if ***y*** and ***x*** are cointegrated, then the relationship between them can be expressed as an ***ECM***. In this section, we assume ***y*** and ***x*** are cointegrated.

In order to understand the properties of ECMs let us begin with the following simple version:

$$\Delta y_t = \varphi + \lambda e_{t-1} + \omega_0 \Delta x_t + v_t \qquad (12.37)$$

where e_{t-1} is the error obtained from the regression model with ***y*** and ***x,***

$$\mathbf{e_{t-1} = y_{t-1} - \alpha - \beta x_{t-1}} \qquad (12.38)$$

and e_t is the error in the ***ECM*** model.
We assume that $\lambda < 0$

Note that, if we know e_{t-1}, then the ***ECM*** is just a regression model. That is, Δy_t is the dependent variable and e_{t-1} and Δx_t are explanatory variables.

The regression model attempts to use explanatory variables to explain the dependent variable. The **ECM** says that Δy_t depends on Δx_t meaning that changes in ***x*** cause ***y*** to change. In addition, Δy_t depends on e_{t-1}. This latter aspect is unique to the ***ECM*** and gives it its name.

In sum, the ECM has both long run and short run properties, such as:

- The former properties are embedded in the e_{t-1} term (which is still the long run multiplier and the errors are from the regression involving ***y*** and ***x***).

- The short run behavior is partially captured by the equilibrium error term, which says that, if ***y*** is out of equilibrium, it is pulled towards it in the next period.

- Further aspects of short run behavior are captured by the inclusion of Δx_t as an explanatory variable. This term implies that, if ***x*** changes, the equilibrium value of ***y*** also change and that ***y*** will also change accordingly.

In ***ECM*** the spurious regression problem does not exist. OLS estimation and ***t-statistics*** and ***p-values*** can be used for interpretation, because of:

- ***y*** and ***x*** both have unit roots; therefore Δy and Δx are stationary.

- Since ***y*** and ***x*** are cointegrated, the equilibrium error is stationary.

- The dependent variable and all explanatory variables in the ECM are stationary.

The inclusion of e_{t-1} as an explanatory variable brings some new statistical issues:

- The errors in the model are not directly observed. How they can be used as an explanatory variable in a regression?

- To replace the unknown errors by the residuals from the regression of ***y*** on ***x*** (replace e_{t-1} by u_{t-1}). A simple technique based on two ***OLS*** regressions proceeds as follows:

  Step 1. Run a regression of ***y*** on ***x*** and save the residuals.

Step 2. Run a regression of ***Δy*** on ***Δx*** and the residuals from Step 1 lagged one period.

Note that ***y*** and ***x*** should have unit roots and be cointegrated before proceeding to ***ECM***. So far we have discussed the simplest error correction model. The ***ECM*** may also have lags and deterministic trend like ***ARDL (p, q)*** model. Incorporating these features into the ***ECM*** yields:

$$\Delta y_t = \varphi + \delta t + \lambda e_{t-1} + \gamma_1 \Delta y_{t-1} + \cdots + \gamma_p \Delta y_{t-p} + \omega_0 \Delta x_t + \cdots + \omega_q \Delta x_{t-q} + v_t \quad (12.39)$$

This expression is still in the form of a regression model and can be estimated using the two-step procedure described above. The adjustment to equilibrium intuition also holds for this model.

As the variables Greece and France stock market indices are cointegrated in the **Example 12.14**, we can proceed to run the following ***ECM***.

$$D(Greece) = \beta_2 + \beta_3 D(France) + \beta_4 e_{t-1} + v \quad (12.40)$$

D(Greece) and D(France) are the first differenced variables;

β_2 is intercept;

β_3 is short run coefficient. It tells us at what rate it corrects the previous period disequilibrium of the system. It validates that there exist a long run equilibrium relationship among the variables.

v is error term

e_{t-1} is the one period lag residual of the model. It is also known as an equilibrium error term. It guides the variables of the system to restore back to equilibrium. Sign of the error correction term should be negative and it should be significant.

After estimating the model reported in the **Table 12.15**, short run coefficient β_3 has been found **0.41** and the coefficient of error term

has been found **0.15** percent meaning that system corrects its previous period disequilibrium at a speed of **0.15** percent monthly. The sign of $\boldsymbol{\beta_4}$ is found negative and significant indicating the non-existence of long run equilibrium relationship between Greece and France stock market indices.

Example 12.15

Table 12.15 Eviews Result for ECM Estimates

Dependent Variable: D(GREECE)
Method: Least Squares
Sample (adjusted): 12/31/1990 12/29/2006
Included observations: 4175 after adjustments

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
C	*0.530286*	*0.654009*	*0.810824*	*0.4175*
D(FRANCE)	*0.413186*	*0.023514*	*17.57155*	*0.0000*
V(-1)	*-0.001512*	*0.001220*	*-1.238985*	*0.2154*

R-squared	*0.068917*	*Mean dependent var*	*0.821274*
Adjusted R-squared	*0.068470*	*S.D. dependent var*	*43.76977*
S.E. of regression	*42.24474*	*Akaike info criterion*	*10.32556*
Sum squared resid	*7445427.*	*Schwarz criterion*	*10.33011*
Log likelihood	*-21551.60*	*Hannan-Quinn criter.*	*10.32717*
F-statistic	*154.4006*	*Durbin-Watson stat*	*1.731312*
Prob(F-statistic)	*0.000000*		

Diagnostic Checking

At first, serial correlation is checked as follows;

Table 12.16 Eviews Output for Serial Correlation Test

Breusch-Godfrey Serial Correlation LM Test:			
F-statistic	*40.63616*	*Prob. F(2,4170)*	*0.0000*
*Obs*R-squared*	*79.81422*	*Prob. Chi-Square(2)*	*0.0000*

Test Equation:
Dependent Variable: RESID
Method: Least Squares
Date: 06/05/13 Time: 18:58
Sample: 12/31/1990 12/29/2006
Included observations: 4175
Presample missing value lagged residuals set to zero.

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
C	*0.000996*	*0.647883*	*0.001537*	*0.9988*
D(FRANCE)	*-0.001451*	*0.023295*	*-0.062292*	*0.9503*
V(-1)	*-0.000639*	*0.001216*	*-0.525342*	*0.5994*
RESID(-1)	*0.139198*	*0.015513*	*8.972856*	*0.0000*
RESID(-2)	*-0.031297*	*0.015526*	*-2.015786*	*0.0439*

R-squared	*0.019117*	*Mean dependent var*	*-6.18E-17*
Adjusted R-squared	*0.018176*	*S.D. dependent var*	*42.23462*
S.E. of regression	*41.84902*	*Akaike info criterion*	*10.30721*
Sum squared resid	*7303092.*	*Schwarz criterion*	*10.31480*
Log likelihood	*-21511.30*	*Hannan-Quinn criter.*	*10.30990*
F-statistic	*20.31808*	*Durbin-Watson stat*	*1.999187*
Prob(F-statistic)	*0.000000*		

Since it is significant we can reject the null hypothesis of there is no serial correlation, and conclude that our model is serially correlated.

In the second step normality assumption is checked as follows;

Check whether the residuals are normally distributed.

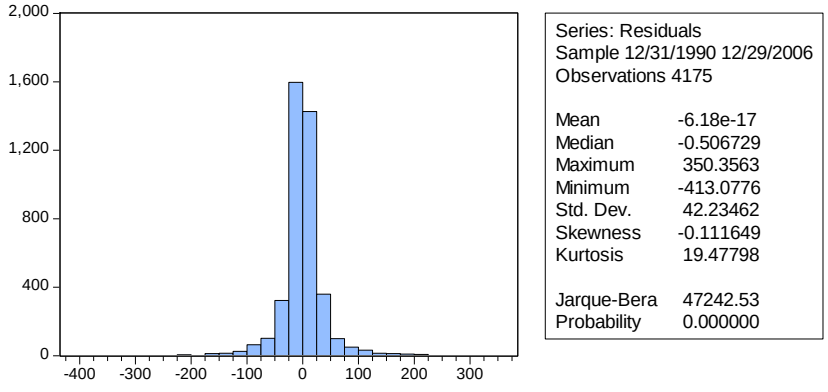


Figure 12.1 Eviews Output for Normality

JB statistics is significant. We reject the null hypothesis of residuals are normally distributed. Residuals of the model are not normally distributed.

Dependent and Independent variables have unit roots but are not cointegrated

Even though the time series have unit root, the Engle–Granger test may indicate no cointegration. In such cases, the series may not be trending together and may not have an equilibrium relationship. Hence, a regression of ***y*** on ***x*** should not be run due to the spurious regression problem. The presence of such characteristics suggests that basic model should be respecified and other explanatory variables included. Instead of working with ***y*** and ***x*** themselves, difference series can be used.

In this case, the following ARDL model can be employed but with changes in the variables:

$$\boldsymbol{\Delta y_t = \alpha + \delta t + \phi_1 \Delta y_{t-1} + \cdots + \phi_p \Delta y_{t-p} + \beta_0 \Delta x_t + \cdots + \beta_q \Delta x_{t-q} + e_t} \quad (12.41)$$

For most time series variables, this specification should not suffer from multicollinearity problems. Or, second variant of the ARDL model based on the differenced data can be estimated.

But if you are working with the differences of your time series and then use the variant of the ***ARDL*** that involves differencing the data you end up with second differenced data:

$$\Delta^2 y_t = \alpha + \delta t + \rho \Delta y_{t-1} + \gamma_1 \Delta^2 y_{t-1} + \cdots + \gamma_{p-1} \Delta^2 y_{t-p+1} + \theta \Delta x_t + \omega_1 \Delta^2 x_t + \cdots + \omega_q \Delta^2 x_{t-q+1} + e_t \quad (12.42)$$

Where:

$$\Delta^2 y_t = \Delta y_t - \Delta y_{t-1} \quad (12.43)$$

OLS estimation and testing can be done in either of these models in a straightforward way. Whatever model is chosen, it is important to emphasize that the interpretation of regression results will likewise change.

More specifically, suppose that:

y: exchange rates and x : interest rates

If ***y*** and ***x*** are cointegrated, or if both are stationary, we can obtain an estimate of the long run effect of a small change in interest rates on exchange rates. If ***y*** and ***x*** are neither stationary nor cointegrated and we estimate either of the two preceding equations, we can obtain an estimate of the long run effect of a small change in the change of interest rates on the change in exchange rates.

This may or may not be a sensible thing to measure depending on the particular empirical exercise, such as:

- If all variables are stationary, then an ***ARDL(p,q)*** model can be estimated using ***OLS***. Statistical techniques are all standard and valid.
- A variant on the ***ARDL*** model is often used to avoid potential multicollinearity problems and provide a straightforward estimate of the long run multiplier.
- If all variables are nonstationary, the spurious regression problem can occur.

- If cointegration is present, the spurious regression problem does not occur.
- If all variables are nonstationary but the regression error is stationary, then cointegration occurs.
- Cointegration implies whether an equilibrium relationship exists.
- Cointegration can be tested using the Engle–Granger test. This test is a Dickey–Fuller test on the residuals from the cointegrating regression.
- If the variables are cointegrated, then an error correction model can be employed. This model captures short run behavior in a way that the cointegrating regression cannot.
- If the variables have unit roots but are not cointegrated, take difference and estimate an ***ARDL*** model using the differenced variables. The interpretation of these models can be awkward.

Multivariate Cointegration

The Johansen test (Johansen, 1988, 1991) is a test for cointegration that allows more than one cointegrating relationship, unlike the Engle–Granger method.

Example 12.16

Table 12.16 Eviews Output for Johansen Cointegration Test

Sample (adjusted): 1/06/1989 12/29/2006
Included observations: 4691 after adjustments
Trend assumption: Linear deterministic trend
Series: GERMANY SPAIN ITALY
Lags interval (in first differences): 1 to 4
Unrestricted Cointegration Rank Test (Trace)

Hypothesized No. of CE(s)	*Eigenvalue*	*Trace Statistic*	*0.05 Critical Value*	*Prob.***
*None **	*0.003696*	*31.25751*	*29.79707*	*0.0337*
At most 1	*0.002596*	*13.88696*	*15.49471*	*0.0861*
At most 2	*0.000361*	*1.692354*	*3.841466*	*0.1933*

Trace test indicates 1 cointegrating eqn(s) at the 0.05 level
** denotes rejection of the hypothesis at the 0.05 level*
***MacKinnon-Haug-Michelis (1999) p-values*
Unrestricted Cointegration Rank Test (Maximum Eigenvalue)

Hypothesized No. of CE(s)	*Eigenvalue*	*Max-Eigen Statistic*	*0.05 Critical Value*	*Prob.***
None	*0.003696*	*17.37056*	*21.13162*	*0.1552*

At most 1	0.002596	12.19460	14.26460	0.1035
At most 2	0.000361	1.692354	3.841466	0.1933

Max-eigenvalue test indicates no cointegration at the 0.05 level
* denotes rejection of the hypothesis at the 0.05 level
**MacKinnon-Haug-Michelis (1999) p-values

Unrestricted Cointegrating Coefficients (normalized by b'*S11*b=I):

GERMANY	SPAIN	ITALY
-0.002378	-0.008764	0.004923
0.002253	-0.003252	-0.001580
0.000603	0.000243	0.000169

Unrestricted Adjustment Coefficients (alpha):

D(GERMANY)	-1.096077	-0.998472	-0.635771
D(SPAIN)	-0.318655	-0.286985	-0.020752
D(ITALY)	-1.959926	-0.175790	-0.230935

1 Cointegrating Equation(s): Log likelihood -59785.64

Normalized cointegrating coefficients (standard error in parentheses)

GERMANY	SPAIN	ITALY
1.000000	3.684964	-2.070029
	(0.84823)	(0.24787)

Adjustment coefficients (standard error in parentheses)

D(GERMANY)	0.002607	
	(0.00149)	
D(SPAIN)	0.000758	
	(0.00027)	
D(ITALY)	0.004661	
	(0.00120)	

2 Cointegrating Equation(s): Log likelihood -59779.54

Normalized cointegrating coefficients (standard error in parentheses)

GERMANY	SPAIN	ITALY
1.000000	0.000000	-1.086768
		(0.08357)
0.000000	1.000000	-0.266830
		(0.02765)

Adjustment coefficients (standard error in parentheses)

D(GERMANY)	0.000358	0.012853
	(0.00205)	(0.00584)
D(SPAIN)	0.000111	0.003726
	(0.00037)	(0.00106)
D(ITALY)	0.004265	0.017748
	(0.00166)	(0.00473)

Johansen cointegration test result reported in Table 12.16 indicate that there is one cointegrating relationships between variables meaning that there is significant long run relationship between variables.

Vector Error Correction Model

Vector Error Correction Model is used to explore dynamic (short run) cointegration relationship among more then three variables.

Example 12.17

Table 12.17 Eviews Output for VECM Test

Vector Error Correction Estimates *Sample (adjusted): 1/04/1989 12/29/2006* *Included observations: 4693 after adjustments* *Standard errors in ( ) & t-statistics in [ ]*			
Cointegrating Eq:	*CointEq1*		
GERMANY(-1)	*1.000000*		
SPAIN(-1)	*3.364462* *(0.75675)* *[4.44593]*		
ITALY(-1)	*-1.974751* *(0.22148)* *[-8.91605]*		
C	*-139.3244*		
Error Correction:	*D(GERMNY)*	*D(SPAIN)*	*D(ITALY)*
CointEq1	*0.002324* *(0.00158)* *[1.47398]*	*0.000753* *(0.00029)* *[2.62824]*	*0.005015* *(0.00128)* *[3.93176]*
D(GERMANY(-1))	*-0.104841* *(0.02289)* *[-4.58118]*	*0.007890* *(0.00416)* *[1.89830]*	*0.026180* *(0.01851)* *[1.41404]*
D(GERMANY(-2))	*-0.042952* *(0.02285)* *[-1.87952]*	*-0.010781* *(0.00415)* *[-2.59759]*	*-0.016748* *(0.01849)* *[-0.90585]*
D(SPAIN(-1))	*0.414903* *(0.13343)*	*0.007206* *(0.02423)*	*0.015514* *(0.10795)*

	[3.10955]	[0.29736]	[0.14372]
D(SPAIN(-2))	0.039475	-0.013438	-0.207321
	(0.13332)	(0.02421)	(0.10786)
	[0.29608]	[-0.55501]	[-1.92210]
D(ITALY(-1))	0.075161	-0.002533	-0.037561
	(0.02933)	(0.00533)	(0.02373)
	[2.56277]	[-0.47560]	[-1.58306]
D(ITALY(-2))	0.023652	0.009558	0.070995
	(0.02933)	(0.00533)	(0.02373)
	[0.80634]	[1.79421]	[2.99176]
C	0.575387	0.271770	0.650120
	(0.62632)	(0.11374)	(0.50670)
	[0.91868]	[2.38932]	[1.28305]
R-squared	0.007436	0.005513	0.006062
Adj. R-squared	0.005953	0.004027	0.004577
Sum sq. resids	8598952.	283602.3	5628036.
S.E. equation	42.84183	7.780367	34.65961
F-statistic	5.014384	3.709923	4.081816
Log likelihood	-24289.09	-16283.25	-23294.46
Akaike AIC	10.35461	6.942787	9.930730
Schwarz SC	10.36561	6.953789	9.941732
Mean dependent	0.662387	0.272514	0.624831
S.D. dependent	42.96993	7.796079	34.73920

Determinant resid covariance (dofadj.)	23839045
Determinant resid covariance	23717340
Log likelihood	-59824.83
Akaike information criterion	25.50685
Schwarz criterion	25.54399

D(GERMANY) = C(1)*(GERMANY(-1) + 3.36446151586*SPAIN(-1) - 1.97475095482*ITALY(-1) - 139.324408384) + C(2)*D (GERMANY(-1)) + C(3)*D(GERMANY(-2)) + C(4)*D(SPAIN(-1)) + C(5)*D(SPAIN(-2)) + C(6)*D(ITALY(-1)) + C(7)*D(ITALY(-2)) + C(8)

Hence, estimate the above model by OLS.

Table 12.18 Eviews Output for Johansen Cointegration Test

dependent Variable: D(GERMANY)
Method: Least Squares
Date: 06/05/13 Time: 19:11
Sample (adjusted): 1/04/1989 12/29/2006
Included observations: 4693 after adjustments
D(GERMANY) = C(1)(GERMANY(-1) + 3.36446151586*SPAIN(-1) - 1.97475095482*ITALY(-1) - 139.324408384) + C(2)*D(GERMANY(-1)) + C(3)*D(GERMANY(-2)) + C(4)*D(SPAIN(-1)) + C(5)*D(SPAIN(-2)) + C(6)*D(ITALY(-1)) + C(7)*D(ITALY(-2)) + C(8)*

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	0.002324	0.001577	1.473983	0.1406
C(2)	-0.104841	0.022885	-4.581177	0.0000
C(3)	-0.042952	0.022853	-1.879518	0.0602
C(4)	0.414903	0.133429	3.109552	0.0019
C(5)	0.039475	0.133325	0.296080	0.7672
C(6)	0.075161	0.029328	2.562772	0.0104
C(7)	0.023652	0.029332	0.806340	0.4201
C(8)	0.575387	0.626318	0.918681	0.3583

R-squared	0.007436	Mean dependent var	0.662387
Adjusted R-squared	0.005953	S.D. dependent var	42.96993
S.E. of regression	42.84183	Akaike info criterion	10.35461
Sum squared resid	8598952.	Schwarz criterion	10.36561
Log likelihood	-24289.09	Hannan-Quinn criter.	10.35848
F-statistic	5.014384	Durbin-Watson stat	2.001250
Prob(F-statistic)	0.000011		

C(1) is the coefficient of VECM. It should be negative and significant. In our model it is positive and not significant. It is unable to validate Johansen cointegration test results reported in the Table 12.18. The other coefficients are short run coefficients.

Diagnostic Checking

At first, whether there is serial correlation is checked as follows;

Table 12.19 Eviews Output for Serial Correlation

Breusch-Godfrey Serial Correlation LM Test:			
F-statistic	*1.150315*	*Prob. F(2,4683)*	*0.3166*
*Obs*R-squared*	*2.304411*	*Prob. Chi-Square(2)*	*0.3159*

Test Equation:
Dependent Variable: RESID
Method: Least Squares
Sample: 1/04/1989 12/29/2006
Presample missing value lagged residuals set to zero.

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
C(1)	*-0.001607*	*0.001902*	*-0.844888*	*0.3982*
C(2)	*0.845628*	*0.576122*	*1.467794*	*0.1422*
C(3)	*-0.012230*	*0.180551*	*-0.067738*	*0.9460*
C(4)	*-0.002758*	*0.133439*	*-0.020668*	*0.9835*
C(5)	*-0.336260*	*0.266225*	*-1.263068*	*0.2066*
C(6)	*-0.002110*	*0.029361*	*-0.071868*	*0.9427*
C(7)	*-0.062456*	*0.052016*	*-1.200713*	*0.2299*
C(8)	*-0.422910*	*0.687281*	*-0.615338*	*0.5384*
RESID(-1)	*-0.844129*	*0.574666*	*-1.468905*	*0.1419*
RESID(-2)	*0.097797*	*0.206616*	*0.473326*	*0.6360*

R-squared	*0.000491*	*Mean dependent var*	*1.22E-15*
Adjusted R-squared	*-0.001430*	*S.D. dependent var*	*42.80986*
S.E. of regression	*42.84045*	*Akaike info criterion*	*10.35497*
Sum squared resid	*8594730.*	*Schwarz criterion*	*10.36872*
Log likelihood	*-24287.94*	*Hannan-Quinn criter.*	*10.35981*
F-statistic	*0.255626*	*Durbin-Watson stat*	*1.999579*
Prob(F-statistic)	*0.985748*		

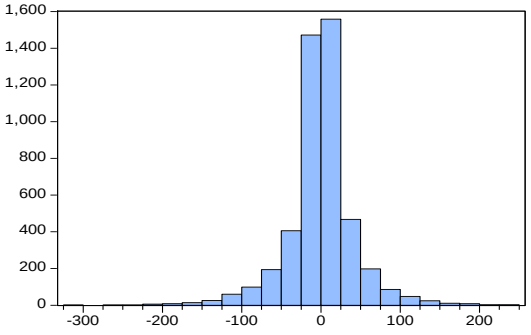


Figure 12.2 Eviews Output for Normality Test

Table 12.20 Eviews Output for Heteroscedasticity

Heteroskedasticity Test: Breusch-Pagan-Godfrey			
F-statistic	*84.61243*	*Prob. F(9,4683)*	*0.0000*
*Obs*R-squared*	*656.3996*	*Prob. Chi-Square(9)*	*0.0000*
Scaled explained SS	*2309.130*	*Prob. Chi-Square(9)*	*0.0000*

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Sample: 1/04/1989 12/29/2006
Included observations: 4693

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
C	*-1652.975*	*163.4808*	*-10.11113*	*0.0000*
GERMANY(-1)	*-8.504171*	*2.413443*	*-3.523667*	*0.0004*
SPAIN(-1)	*-21.73891*	*14.07776*	*-1.544203*	*0.1226*
ITALY(-1)	*1.605723*	*3.094562*	*0.518885*	*0.6039*
GERMANY(-2)	*0.924901*	*3.167677*	*0.291981*	*0.7703*
GERMANY(-3)	*8.457493*	*2.413339*	*3.504478*	*0.0005*
SPAIN(-2)	*21.95260*	*19.60785*	*1.119582*	*0.2629*
SPAIN(-3)	*-6.050174*	*14.09019*	*-0.429389*	*0.6677*
ITALY(-2)	*-1.505254*	*4.240540*	*-0.354968*	*0.7226*
ITALY(-3)	*1.872900*	*3.094881*	*0.605161*	*0.5451*

R-squared	*0.139868*	*Mean dependent var*	*1832.293*
Adjusted R-squared	*0.138215*	*S.D. dependent var*	*4868.971*
S.E. of regression	*4519.983*	*Akaike info criterion*	*19.67253*
Sum squared resid	*9.57E+10*	*Schwarz criterion*	*19.68628*
Log likelihood	*-46151.60*	*Hannan-Quinn criter.*	*19.67737*
F-statistic	*84.61243*	*Durbin-Watson stat*	*1.848660*
Prob(F-statistic)	*0.000000*		

Potential Pitfalls of the Cointegration Method

i- If several lags of the same variable are used, it can cause multicollinearity; if x_t is highly correlated with its own past values, then autocorrelation occurs.

ii- One variable can influence another with a time lag.

iii- If the data are non-stationary, a problem known as spurious regression may arise.

iv- Interpretation depends on whether, ***x*** and ***y***, are stationary or not.

DIAGNOSTIC CHECKING

The model adequacy should be investigated in the last step of the model building cycle. Following residual analysis may be performed;

- If *ARMA(p,q)* model is chosen on the basis of the sample *ACF* and *PACF*, an *ARMA(p+1,q) and an ARMA(p,q+1)* should be estimated in order to test the significance of the additional parameters.

- The residuals of an well specified model should be white noise.

- A plot of the residuals can be used for checking outliers.

- The significance of residual autocorrelation may be examined. The autocorrelations are zero for a white noise series. The Ljung-Box portmanteau test statistics is used in order to test the residual autocorrelation by comparing with approximate two standard error bounds:

$$Q_k = T(T+2)\sum_{k-1}^{K}\frac{1}{T-k}r_k^2 \qquad (12.44)$$

r_k is the estimated autocorrelation coefficients of the residuals
k is a number chosen by the researcher.

The structure of the model should be revised, if the model is rejected at this stage.

CHAPTER 13

VOLATILITY

ARCH MODEL

This model is closely related to the Autoregressive Conditional Heteroscedasticity (ARCH), (Engle, 1982).

Suppose the following random walk model:

$$\boldsymbol{y_t = y_{t-1} + e_t,} \quad (13.1)$$

$$\boldsymbol{\Delta y_t = e_t,} \quad (13.2)$$

Naturally, stock prices are quite fragile and influenced by expected and unexpected factors, reflected in the error (e_t). This result for investors is unable to predict future stock price changes. If the future stock price movements are predicTable, then there would be arbitrage opportunities. Arbitrage opportunities give advantage to the smart investors and they are instantly eliminated.

A more realistic model is:

$$\boldsymbol{\Delta y_t = \alpha + e_t,} \quad (13.3)$$

This model implies that stock prices, on average, increase by $\boldsymbol{e_t}$ per period, but are otherwise unpredicTable. This is known as the ***random walk with drift model***; it is a form of the ***random walk model*** with intercept. It allows stock prices to "drift" upwards over time.

If this is the case, the behavior of the series is not possible to predict. Instead, the volatility of financial time series can be explored. In fact, volatility changes over time are examined.

Mostly, to know the level of macroeconomic indicators is not important, but their variance. For instance, if the level of inflation is high, agents or individuals can plan for the future with a high degree of confidence if the variance of inflation is low. However, if there is a high variance it is hard to predict what the inflation rate might be next period. The low level of variance reflects high level of reliable prediction is possible which also reflect low level of risk. For example, if the exchange rate exhibits low volatility it becomes easy to plan whereas high volatility makes it more difficult. The negative effect of uncertainty partially accounts for the growing number of financial derivatives (futures and options) that firms can use to hedge against the risk posed by currency fluctuations.

The volatility is a measure of the riskiness of a stock. Nonetheless, the riskiness of a portfolio of stocks depends not only on the volatility of the individual stocks, but also on the correlation between the stocks in the portfolio (CAPM).

Let's assume that either asset price follows a pure random walk model or a random walk with drift:

$$\Delta \boldsymbol{Y_t} = \Delta \boldsymbol{y_t} - \Delta \overline{\boldsymbol{y}}, \qquad (13.4)$$

Taking deviations from mean implies that there is no intercept in the model.

$\boldsymbol{\Delta Y_t^2}$ is an estimator of volatility at time $\boldsymbol{t}$. High volatility is associated with big changes. This measure of volatility will be small in sTable times and large in times of crises.

Autoregressive models are normally used to model "clustering in volatility", which exists in financial time series data. Consider, for instance, an AR (1) model that uses volatility as the time series variable of interest:

$$\Delta Y_t^2 = \alpha + \theta \Delta Y_{t-1}^2 + e_t, \qquad (13.5)$$

This model has volatility in a period depending on volatility in a previous period. If, $\theta > 0$ then the volatility was high in last period (e.g. ΔY_{t-1}^2 was very large), it will also tend to be high in this period. Alternatively, if volatility was low last period (e.g. $\Delta Y_{t-1}^2$was near zero) then this period's volatility will also tend to be low.

Low volatility is followed by low volatility and high volatility is followed by high. The level of volatility tends to be similar over time. But, in general, this model implies that intervals or clusters are observed in time where volatility is low and intervals where it is high. In empirical studies of asset prices, such a pattern is very common.[37]

All the series in the model should be stationary. If not, there should be converted into stationary before proceed to estimate ***ARCH*** and ***GARCH*** (Bollerslev, 1986) model. model.

Example 13.1

Let's start with the model that includes Greece stock index.

GREECE C

Estimate the equation and check the residual through graph to identify whether the periods of high volatility is followed by period of high volatility and vice versa.

[37] Koop, G. (2005)

Table 13.1 Eviews Output for Estimation

Dependent Variable: D(GREECE)
Method: Least Squares
Date: 06/06/13 Time: 19:06
Sample (adjusted): 1/02/1989 12/29/2006
Included observations: 4695 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.874198	0.611938	1.428573	0.1532

R-squared	0.000000	Mean dependent var	0.874198
Adjusted R-squared	0.000000	S.D. dependent var	41.93003
S.E. of regression	41.93003	Akaike info criterion	10.31009
Sum squared resid	8252650.	Schwarz criterion	10.31147
Log likelihood	-24201.95	Hannan-Quinn criter.	10.31058
Durbin-Watson stat	1.662471		

Go to ***ARCH*** test under heteroscedasticity test and run. Check R^2 and ***p-values***. If ***p-value*** is less than 5%, we can reject null hypothesis of there is no ***ARCH*** effect.

After rejection of the null hypothesis we can proceed to ARCH models.

Table 13.2 Eviews Output for Heteroscedasticity Test

Heteroskedasticity Test: ARCH

F-statistic	365.0401	Prob. F(1,4692)	0.0000
Obs*R-squared	338.8343	Prob. Chi-Square(1)	0.0000

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Sample (adjusted): 1/03/1989 12/29/2006
Included observations: 4694 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1285.767	108.3720	11.86438	0.0000
RESID^2(-1)	0.268672	0.014062	19.10602	0.0000

R-squared	0.072185	Mean dependent var	1758.125
Adjusted R-squared	0.071987	S.D. dependent var	7504.223
S.E. of regression	7229.077	Akaike info criterion	20.61004
Sum squared resid	2.45E+11	Schwarz criterion	20.61279
Log likelihood	-48369.76	Hannan-Quinn criter.	20.61100
F-statistic	365.0401	Durbin-Watson stat	2.203790
Prob(F-statistic)	0.000000		

Go to estimate and choose ***ARCH*** model from the list. Choose ***1*** for ***ARCH***, ***0*** for ***GARCH***, and normal distribution, then run the model which is ***GARCH(1,0)***.

Table 13.3 Eviews Output for ARCH Test

Dependent Variable: D(GREECE)
Method: ML - ARCH (Marquardt) - Normal distribution
Sample (adjusted): 1/02/1989 12/29/2006
Included observations: 4695 after adjustments
Convergence achieved after 20 iterations
Presample variance: backcast (parameter = 0.7)
*GARCH = C(2) + C(3)*RESID(-1)^2*

Variable	*Coefficient*	*Std. Error*	*z-Statistic*	*Prob.*
C	*1.113585*	*0.278842*	*3.993613*	*0.0001*
	Variance Equation			
C	*774.5477*	*6.553909*	*118.1810*	*0.0000*
RESID(-1)^2	*0.890166*	*0.021965*	*40.52626*	*0.0000*

R-squared	*-0.000033*	*Mean dependent var*	*0.874198*
Adjusted R-squared	*-0.000459*	*S.D. dependent var*	*41.93003*
S.E. of regression	*41.93965*	*Akaike info criterion*	*9.985032*
Sum squared resid	*8252919.*	*Schwarz criterion*	*9.989156*
Log likelihood	*-23436.86*	*Hannan-Quinn criter.*	*9.986482*
Durbin-Watson stat	*1.662417*		

Now estimate GARCH(1,1) model

Table 13.4 Eviews output for GARCH(1,1) test

Dependent Variable: D(GREECE)
Method: ML - ARCH (Marquardt) - Normal distribution
Date: 06/06/13 Time: 19:10
Sample (adjusted): 1/02/1989 12/29/2006
Included observations: 4695 after adjustments
Convergence achieved after 15 iterations
Presample variance: backcast (parameter = 0.7)
*GARCH = C(2) + C(3)*RESID(-1)^2 + C(4)*GARCH(-1)*

Variable	*Coefficient*	*Std. Error*	*z-Statistic*	*Prob.*
C	*0.229580*	*0.175920*	*1.305023*	*0.1919*
	Variance Equation			
C	*0.697177*	*0.099060*	*7.037933*	*0.0000*
RESID(-1)^2	*0.119684*	*0.005399*	*22.16970*	*0.0000*
GARCH(-1)	*0.896867*	*0.003844*	*233.3240*	*0.0000*

R-squared	*-0.000236*	*Mean dependent var*	*0.874198*
Adjusted R-squared	*-0.000876*	*S.D. dependent var*	*41.93003*
S.E. of regression	*41.94839*	*Akaike info criterion*	*9.101550*
Sum squared resid	*8254601.*	*Schwarz criterion*	*9.107049*
Log likelihood	*-21361.89*	*Hannan-Quinn criter.*	*9.103484*
Durbin-Watson stat	*1.662078*		

TARCH model

In the Threshold ARCH (TARCH) model innovations or shocks are divided into intervals and approximated a piecewise linear function for the conditional standard deviation.[38]

Two intervals mean that the division is normally at zero and the impact of positive and negative shocks on the volatility is differentiated.

Table 13.5 Eviews Output for TARCH Test

Dependent Variable: D(GREECE)
Method: ML - ARCH (Marquardt) - Normal distribution
Sample (adjusted): 1/02/1989 12/29/2006
Included observations: 4695 after adjustments
Convergence achieved after 37 iterations
Presample variance: backcast (parameter = 0.7)
*GARCH = C(2) + C(3)*RESID(-1)^2 + C(4)*RESID(-1)^2*(RESID(-1)<0) + C(5)*GARCH(-1)*

Variable	*Coefficient*	*Std. Error*	*z-Statistic*	*Prob.*
C	*0.266773*	*0.188900*	*1.412241*	*0.1579*
		Variance Equation		
C	*0.617096*	*0.091341*	*6.755947*	*0.0000*
RESID(-1)^2	*0.122972*	*0.006082*	*20.21843*	*0.0000*
RESID(-1)^2(RESID(-1)<0)*	*-0.016922*	*0.006908*	*-2.449444*	*0.0143*
GARCH(-1)	*0.900832*	*0.003799*	*237.1060*	*0.0000*

R-squared	*-0.000210*	*Mean dependent var*	*0.874198*
Adjusted R-squared	*-0.001063*	*S.D. dependent var*	*41.93003*
S.E. of regression	*41.95231*	*Akaike info criterion*	*9.101390*
Sum squared resid	*8254382.*	*Schwarz criterion*	*9.108264*
Log likelihood	*-21360.51*	*Hannan-Quinn criter.*	*9.103807*
Durbin-Watson stat	*1.662122*		

[38] See Glosten et al. (1993) for more information.

EGARCH model

Standard GARCH models are based on the same (symmetric) effect of negative and positive shoks on the volatility which is mostly violated in practice. Normaly, stock returns are more responsive to the negative shoks than positive shocks. This is called Leverage Effect (Black, 1976).

The exponential GARCH (EGARCH) model is a GARCH family model provided by Nelson (1990). This model includes modelling logarithm and leverage term to detect asymmetry in volatility clustering as follows;

$$logY_t^2 = \alpha + \theta logY_{t-1}^2 + \gamma \frac{e_{t-1}}{Y_{t-1}} + \delta \frac{|e_{t-1}|}{y_{t-1}} \qquad (13.6)$$

The model is asymmetric as long as $\gamma \neq 0$. If $\gamma < 0$, then positive shocks generate less volatility then negative shocks.[39]

Table 13.6 Eviews Output for EGARCH Test

Dependent Variable: D(GREECE)
Method: ML - ARCH (Marquardt) - Normal distribution
Sample (adjusted): 1/02/1989 12/29/2006
Included observations: 4695 after adjustments
Convergence achieved after 18 iterations
Presample variance: backcast (parameter = 0.7)
*LOG(GARCH) = C(2) + C(3)*ABS(RESID(-1)/@SQRT(GARCH(-1))) + C(4)*
**LOG(GARCH(-1))*

Variable	*Coefficient*	*Std. Error*	*z-Statistic*	*Prob.*
C	*0.222382*	*0.115135*	*1.931490*	*0.0534*
		Variance Equation		
C(2)	*-0.098750*	*0.005062*	*-19.50920*	*0.0000*
C(3)	*0.208688*	*0.007636*	*27.32960*	*0.0000*
C(4)	*0.992058*	*0.000647*	*1532.764*	*0.0000*
R-squared	*-0.000242*	*Mean dependent var*		*0.874198*
Adjusted R-squared	*-0.000881*	*S.D. dependent var*		*41.93003*
S.E. of regression	*41.94850*	*Akaike info criterion*		*9.103164*
Sum squared resid	*8254645.*	*Schwarz criterion*		*9.108663*
Log likelihood	*-21365.68*	*Hannan-Quinn criter.*		*9.105098*
Durbin-Watson stat	*1.662069*			

[39] See Verbeek (A guide to modern econometrics, 2004) for more information.

We have estimated four models from ***ARCH*** family models. Among them ***GARCH(1,1)*** is the best because it has lowest ***AIC*** and ***SIC*** values.

DIAGNOSTIC CHECKING

1- serial correlation

Table 13.7 Eviews Output for Serial Correlation Test

Sample: 1/02/1989 12/29/2006				Included observations: 4695			
Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
| |	| |	1	0.014	0.014	0.9418	0.332
| |	| |	2	0.006	0.006	1.1159	0.572
| |	| |	3	0.023	0.023	3.6682	0.300
| |	| |	4	-0.002	-0.002	3.6835	0.451
| |	| |	5	-0.004	-0.005	3.7701	0.583
| |	| |	6	-0.015	-0.016	4.8983	0.557
| |	| |	7	-0.018	-0.017	6.4174	0.492
| |	| |	8	-0.017	-0.016	7.7068	0.463
| |	| |	9	-0.024	-0.023	10.378	0.321
| |	| |	10	0.014	0.016	11.292	0.335
| |	| |	11	-0.010	-0.009	11.745	0.383
| |	| |	12	-0.019	-0.019	13.528	0.332
| |	| |	13	-0.006	-0.007	13.695	0.396
| |	| |	14	-0.031	-0.032	18.368	0.191
| |	| |	15	-0.005	-0.005	18.503	0.237
| |	| |	16	-0.032	-0.032	23.267	0.107
| |	| |	17	-0.021	-0.019	25.321	0.088
| |	| |	18	-0.021	-0.021	27.357	0.073
| |	| |	19	-0.031	-0.030	31.839	0.033
| |	| |	20	-0.006	-0.007	32.008	0.043
| |	| |	21	-0.010	-0.011	32.443	0.053
| |	| |	22	-0.011	-0.012	33.040	0.061
| |	| |	23	-0.010	-0.013	33.482	0.073
| |	| |	24	-0.016	-0.018	34.710	0.073
| |	| |	25	0.002	-0.002	34.724	0.093
| |	| |	26	0.009	0.006	35.096	0.110
| |	| |	27	0.007	0.004	35.297	0.131
| |	| |	28	-0.011	-0.016	35.830	0.147
| |	| |	29	-0.001	-0.004	35.835	0.178
| |	| |	30	0.004	-0.001	35.930	0.210
| |	| |	31	-0.017	-0.021	37.261	0.203
| |	| |	32	-0.034	-0.038	42.780	0.097
| |	| |	33	0.002	-0.002	42.798	0.118
| |	| |	34	0.011	0.009	43.375	0.130
| |	| |	35	0.012	0.008	44.023	0.141
| |	| |	36	-0.019	-0.025	45.818	0.127

Since all the correspondent ***p-values*** are more than 5% then we do not have any evidence to reject null hypothesis of there is no serial correlation.

2- whether our model has ARCH effect or not

Table 13.8 Eviews Output for Heteroscedasticity Test

Heteroskedasticity Test: ARCH			
F-statistic	*0.940887*	*Prob. F(1,4692)*	*0.3321*
*Obs*R-squared*	*0.941100*	*Prob. Chi-Square(1)*	*0.3320*

Test Equation:
Dependent Variable: WGT_RESID^2
Method: Least Squares
Sample (adjusted): 1/03/1989 12/29/2006
Included observations: 4694 after adjustments

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
C	*0.985305*	*0.039104*	*25.19718*	*0.0000*
WGT_RESID^2(-1)	*0.014159*	*0.014597*	*0.969994*	*0.3321*

R-squared	*0.000200*	*Mean dependent var*	*0.999463*
Adjusted R-squared	*-0.000013*	*S.D. dependent var*	*2.485467*
S.E. of regression	*2.485483*	*Akaike info criterion*	*4.659237*
Sum squared resid	*28985.42*	*Schwarz criterion*	*4.661987*
Log likelihood	*-10933.23*	*Hannan-Quinn criter.*	*4.660204*
F-statistic	*0.940887*	*Durbin-Watson stat*	*2.000046*
Prob(F-statistic)	*0.332100*		

P value is bigger than 5% it means that we do not have enough evidence to reject null hypothesis of there is no ARCH effect.

3- Whether the residuals are normally distributed

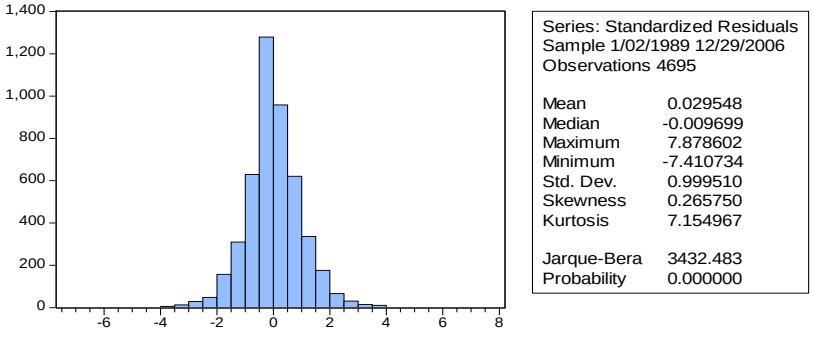


Figure 13.1 Eviews Output for Normality Test

Correspondent ***p-value*** is less than 5%. It means that we have enough evidence to reject null hypothesis of residuals are normally distributed.

This model does not have serial correlation, ***ARCH*** effect but it is not normally distributed. This model still can be accepted although residuals are not normally distributed.

VOLATILITY SPILLOVER, GARCH (1,1) MODEL

Mean Equation

$$D(GREECE) = C1 + C2D(FRANCE) \qquad (13.7)$$

Variance Equation

$$H_t = C3 + C4H_{t-1} + C5e_{t-1}^2 + C6D(TURKEY) + C7D(GERMANY) + C8D(ENGLAND) \qquad (13.8)$$

First plot the residuals derived from the mean equation.

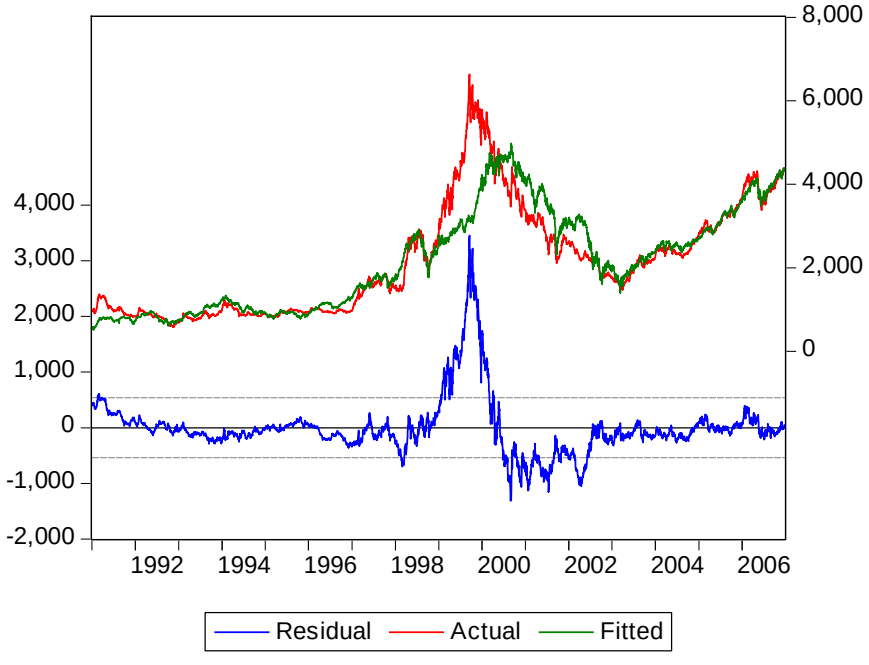


Figure 13.2 Eviews Output for Residuals Derived from Mean Equation

H_t: variance of the residual derived from mean equation. It is known as current days variance or volatility of GREECE.

H_{t-1}: previous days residual variance or volatility. It is known as ***GARCH*** term.

e_{t-1}^2: previous periods squared residual derived from mean equation. It is known as previous days bond return information about volatility. It is *ARCH* term.

TURKEY, GERMANY and ENGLAND are known as variance regressors as they can also contribute in the volatility of GREECE (H_t) in the variance equation.

This model is ***GARCH(1,1)*** model which refers to first order ***ARCH*** term and first order ***GARCH*** term.

Here we estimate mean equation and variance equation simultaneously.

Table 13.9 Eviews Output for Volatility Spillover Test

Dependent Variable: GREECE
Method: ML - ARCH (Marquardt) - Normal distribution
Sample (adjusted): 12/28/1990 12/29/2006
Included observations: 4176 after adjustments
Convergence achieved after 56 iterations
Presample variance: backcast (parameter = 0.7)
*GARCH = C(3) + C(4)*RESID(-1)^2 + C(5)*GARCH(-1) + C(6)*TURKEY + C(7)*GERMANY + C(8)*ENGLAND*

Variable	*Coefficient*	*Std. Error*	*z-Statistic*	*Prob.*
FRANCE	*1.263438*	*0.005808*	*217.5493*	*0.0000*
C	*-712.8147*	*21.51183*	*-33.13594*	*0.0000*
		Variance Equation		
C	*287684.6*	*9989.890*	*28.79758*	*0.0000*
RESID(-1)^2	*0.793084*	*0.084408*	*9.395841*	*0.0000*
GARCH(-1)	*-0.188672*	*0.016241*	*-11.61692*	*0.0000*
TURKEY	*-1.204154*	*0.107242*	*-11.22839*	*0.0000*
GERMANY	*-4.213912*	*2.776550*	*-1.517679*	*0.1291*
ENGLAND	*-35.63601*	*1.747883*	*-20.38809*	*0.0000*

R-squared	*0.828314*	*Mean dependent var*	*2169.725*
Adjusted R-squared	*0.828025*	*S.D. dependent var*	*1308.362*
S.E. of regression	*542.5753*	*Akaike info criterion*	*14.15328*
Sum squared resid	*1.23E+09*	*Schwarz criterion*	*14.16542*
Log likelihood	*-29544.05*	*Hannan-Quinn criter.*	*14.15758*
F-statistic	*2872.692*	*Durbin-Watson stat*	*0.007986*
Prob(F-statistic)	*0.000000*		

ARCH is significant. It means that previous days Greece stock market return information (ARCH) can influence today's Greece stock market volatility.

GARCH is also significant. It means that previous days Greece stock market index volatility (***GARCH***) can influence today's Greece stock market volatility.

It means that Greece stock market volatility is influenced by its own ***ARCH*** and ***GARCH*** factors or own shocks.

Turkey stock market and England stock market are also significant meaning that Turkey and England stock market volatility influence the volatility of Greece stock market. In In summary, volatility spillover from Turkey and England stock markets to Greece stock market is detected. Greece stock market volatility is mainly defined by its own shocks such as ***ARCH*** and ***GARCH***. It is also influenced by the Turkey and England stock market volatility. Volatility of the Germany stock market does not contribute in the volatility of Greece stock market.

CHAPTER 14

GRANGER CAUSALITY

If event A happens before event B, then it is possible that A is causing B. In other words, events in the past can cause events to happen today. Future events cannot cause past events.

Causality concern is examined by Granger causality (Granger, 1969). The basic idea is that a variable *x* Granger causes *y* if past values of *x* can help explain *y*. Of course, if Granger causality holds this does not guarantee that ***x*** causes ***y***. Nevertheless, if past values of ***x*** have explanatory power for current values of ***y***, it at least suggests that ***x*** might be causing ***y***.

Granger causality is only relevant with time series variables.

Let's consider Granger causality between two variables (***x*** and ***y***) which are both stationary. A nonstationary case, where ***x*** and ***y*** have unit roots but are cointegrated, will be mentioned below.

Granger Causality in simple ***ARDL*** model is:

$$\boldsymbol{y_t = \alpha + \theta_1 y_{t-1} + \omega_1 x_{t-1} + e_t,} \qquad (14.1)$$

This model implies that last period's value of ***x*** has explanatory power for the current value of ***y***. The coefficient $\boldsymbol{\omega_1}$ measures the level of the influence of $\boldsymbol{x_{t-1}}$ on $\boldsymbol{y_t}$. If $\boldsymbol{\omega_1 = 0}$, then past values of ***x*** have no effect on ***y*** and ***x*** does not Granger cause ***y***. An alternative way of expressing this concept is to say that if $\boldsymbol{\omega_1 = 0}$ then past values of ***x*** have no explanatory power for ***y*** over sample period". Since we know how to estimate the ***ARDL*** and carry out hypothesis tests, it is simple to test for Granger causality.

OLS estimation and correspondent ***t-statistics*** and ***p-values*** can be conducted for the coefficients. If $\boldsymbol{\omega_1}$ is statistically significant (***p-value*** is higher than chosen significance level) then we conclude that ***x*** Granger causes ***y***.

The null hypothesis being tested here is $\mathbf{H_0}$: $\boldsymbol{\omega_1 = 0}$ which is a hypothesis that Granger causality does not occur. The test of $\boldsymbol{\omega_1 = 0}$ may be referred as a test of Granger non-causality.

Granger Causality in ***ARDL (p,q)*** model:

$$\boldsymbol{y_t = \alpha + \delta t + \theta_1 y_{t-1} + \ldots + \theta_p y_{t-p} + \omega_1 x_{t-1} + \ldots + \omega_q x_{t-q} + e_t,} \quad (14.2)$$

Here ***x*** Granger causes ***y*** if any or all of $\boldsymbol{\omega_1, \ldots, \omega_q}$ are statistically significant. Since we are assuming ***x*** and ***y*** does not contain unit roots, ***OLS*** regression analysis can be used to estimate this model and correspondent ***p-values*** of the individual coefficients can be used to determine whether Granger causality is present. If any of the ***p-values*** for the coefficients $\boldsymbol{\omega_1, \ldots, \omega_q}$ were less significance level, then Granger causality is present. If none of the ***p-values*** is less than significance level, then Granger causality is not present.

If any or all of the coefficients $\boldsymbol{\omega_1}$, $\boldsymbol{\omega_2, \ldots, \omega_q}$ are significant using ***t-statistics*** or the ***p-values*** of individual coefficients, then ***x*** Granger causes ***y***. If none of these coefficients is significant, then **x** does not Granger cause **y**.

CAUSALITY IN BOTH DIRECTIONS

Mostly, it is not clear where the direction of the causality is. There can be also bidirectional causality. For instance, should past interest rate changes cause inflation or should the reverse hold? If there is possibility of causality in either direction, it should be checked running a regression of ***x*** on lags of itself and lags of ***y***.

Bidirectional causality is also equally reasonable. For example, exchange rates may also affect future interest rate policy.

GRANGER CAUSALITY IN COINTEGRATED VARIABLES

Testing for Granger causality among cointegrated variables is very similar to the method outlined above. It is common to work with a variant of the error correction model (ECM):

$$\Delta y_t = \varphi + \delta t + \lambda e_{t-1} + \gamma_1 \Delta y_{t-1} + \cdots + \gamma_p \Delta y_{t-p} + \omega_0 \Delta x_t + \cdots + \omega_q \Delta x_{t-q} + \varepsilon_t, \quad (14.3)$$

This is an ***ARDL*** model except the term λe_{t-1}.

$e_{t-1} = y_{t-1} - \alpha - \beta x_{t-1}$ an estimate of which can be obtained by running a regression of ***y*** on ***x*** and saving the residuals.

x Granger causes ***y*** if past values of ***x*** have explanatory power for current values of ***y***. Applying this interpretation to the **ECM**, the past values of ***x*** appear in the terms $\Delta x_{t-1}, \ldots, \Delta x_{t-q}$ and e_{t-1}. This implies that ***x*** does not Granger cause ***y*** if $\omega_1 = \cdots = \omega_q = \lambda = 0$. Thus, **t-statistics** and ***p-values*** can be used to test for Granger causality. Also, the ***F*-tests** can be used to carry out a formal test of: $H_0: \omega_1 = \omega_2 = \ldots = \omega_q = 0$.

Testing whether ***y*** Granger causes ***x*** can be done by reversing the roles of ***x*** and ***y*** in the ***ECM***. If ***x*** and ***y*** are cointegrated then Granger

causality must occur (Granger Representation Theorem). That is, either *x* must Granger cause *y* or *y* must Granger cause *x* (or both).

Example 14.1

Whether Greece stock market causes France stock market or France stock market causes Greece stock market?

$$Greece_t = C1France_{t-i} + C2Greece_{t-j} + e_{1t} \quad (14.4)$$
$$France_t = C3France_{t-i} + C4Greece_{t-j} + e_{2t} \quad (14.5)$$

The variables in the model should be stationary. If not, they should be converted into stationary variables by taking first difference.

We assume that $\boldsymbol{e_{1t}}$ and $\boldsymbol{e_{2t}}$ are not correlated.

Null Hypothesis is Greece stock market does not granger cause France stock market. ***F- statistic*** can be used to test null hypothesis.

Table 14.1 Eviews Outputs for Granger Causality Test

Pairwise Granger Causality Tests
Sample: 12/30/1988 12/29/2006
Lags: 2

Null Hypothesis:	*Obs*	*F-Statistic*	*Prob.*
DFRANCE does not Granger Cause DGREECE	*4173*	*25.3000*	*1.E-11*
DGREECE does not Granger Cause DFRANCE		*0.32394*	*0.7233*

Pairwise Granger Causality Tests
Sample: 12/30/1988 12/29/2006
Lags: 4

Null Hypothesis:	*Obs*	*F-Statistic*	*Prob.*
DFRANCE does not Granger Cause DGREECE	*4171*	*16.7342*	*1.E-13*
DGREECE does not Granger Cause DFRANCE		*2.15275*	*0.0718*

CHAPTER 15

VECTOR AUTOREGRESSION MODEL (VAR)

DEFINITION

In estimating Granger causality between ***x*** and ***y***, restricted version of an ***ARDL (p, q)*** model is used with ***y*** as the dependent variable to investigate whether ***x*** Granger cause ***y***. Causality in the other direction can also be examined switching the roles of ***x*** and ***y*** in the ***ARDL*.** The two equations can be written as follows:

$$y_t = \alpha_1 + \delta_1 t + \theta_{11} y_{t-1} + \ldots + \theta_{1p} y_{t-p} + \omega_{11} x_{t-1} + \ldots + \omega_{1q} x_{t-q} + e_{1t} \quad (15.1)$$

$$x_t = \alpha_2 + \delta_2 t + \theta_{21} y_{t-1} + \ldots + \theta_{2p} y_{t-p} + \omega_{21} x_{t-1} + \ldots + \omega_{2q} x_{t-q} + e_{2t} \quad (15.2)$$

The first one is used to test whether ***x*** Granger causes ***y*** and the second one, whether ***y*** Granger causes ***x***.

These two equations comprise a ***VAR***. A ***VAR*** is the extension of the autoregressive (***AR***) model to the case in which there is more than one variable under study.

The AR model involves one dependent variable, which depended only on lags of itself (and possibly a deterministic trend). A VAR has more than one dependent variable and, has more than one equation. Each equation uses its explanatory variables lags of all the variables under study (and possibly a deterministic trend).

The Eq. (15.1) and Eq. (15.2) form a VAR model with two variables.

In the first equation ***y*** depends on ***p*** lags of itself and on ***q*** lags of ***x***. The lag lengths are ***p*** and ***q***. They can be selected using the sequential testing methods. However, especially if the VAR has more than two variables, many different lag lengths need to be selected (i.e. one for each variable in each equation).

It is common to set ***p = q*** and use the same lag length for every variable in each equation. The resulting model is known as a ***VAR(p)*** model.

For example, the following ***VAR (p)*** has three variables, **y, x** and ***z***;

$$y_t = \alpha_1 + \delta_1 t + \theta_{11} y_{t-1} + \ldots + \theta_{1p} y_{t-p} + \omega_{11} x_{t-1} + \ldots + \omega_{1q} x_{t-q} + \vartheta_{11} z_{t-1} + \cdots + \vartheta_{1p} z_{t-p} + e_{1t} \quad (15.3)$$

$$x_t = \alpha_2 + \delta_2 t + \theta_{21} y_{t-1} + \ldots + \theta_{2p} y_{t-p} + \omega_{21} x_{t-1} + \ldots + \omega_{2q} x_{t-q} + \vartheta_{21} z_{t-1} + \cdots + \vartheta_{2p} z_{t-p} + e_{2t} \quad (15.4)$$

$$z_t = \alpha_3 + \delta_3 t + \theta_{31} y_{t-1} + \ldots + \theta_{3p} y_{t-p} + \omega_{31} x_{t-1} + \ldots + \omega_{3q} x_{t-q} + \vartheta_{31} z_{t-1} + \cdots + \vartheta_{3p} z_{t-p} + e_{3t} \quad (15.5)$$

In addition to an intercept and deterministic trend, each equation contains ***p*** lags of all variables in study.

It is assumed that all the variables in the ***VAR (p)*** are stationary. Estimation and testing can be carried out by using ***OLS***. ***P-values*** or **t-statistics** help to examine whether individual coefficients are significant.

The reasons to use VARs are:

- It is easy to use.

- It provides a framework for testing for Granger causality between each set of variables.

Economic theories or common sense help to interpret the results obtained through VAR model. For instance, ***x*** (exam score) caused ***y*** (class size) or ***x*** (lot size) influenced ***y*** (house price). In both cases it is not plausible for us to say that ***y*** influenced or caused ***x***.

However, there are many circumstances in which neither economic theory nor common sense can provide sensible interpretation reflecting causality. For instance, does ***y*** (interest rate) cause ***x*** (inflation) or vice versa? Economic theory and common sense tells us that either can happen and that Granger causality tests helps to answer these questions.

If cointegration is present one should be careful when interpreting regression results as reflecting causality.

In the VAR model, the explanatory variables might influence the dependent variable, but there is no possibility that the dependent variable influences the explanatory variable.

One of the drawback in VARs is they are not theoretical which are not based on economic theory. There is theory in selecting the variables for the VAR.

For example, the interaction between interest rates, the price level, money supply and real GDP. There have been created many sophisticated models for this relationship (The IS–LM model). However, a VAR model says: Interest rates, price level, money supply and real GDP are related. This relationship can be modeled as each variable depends on lags of itself and all other variables. There is no need to establish any link between the empirical VAR and a theoretical macroeconomic model.

The VAR has better forecasting performance than sophisticated macroeconomic models. The regression-based methods can outperform complicated macroeconomic models. This is a strong motivation for using VARs.

LAG LENGTH SELECTION

The following sequential testing strategy can be used in lag selection:

Step1. Choose the maximum possible lag length, *p*max, that seems reasonable to you.

Step2. Estimate a VAR(*p*max). If **any** of the variables lagged *p*max periods are significant, use the VAR(*p*max), otherwise proceed to the next step.

Step3. Estimate a VAR(*p*max - 1). If **any** of the variables lagged *p*max - 1 periods are significant, use the VAR(*p*max - 1), otherwise proceed to the next step.

Step4. Estimate a VAR(*p*max - 2). If **any** of the variables lagged *p*max - 2 periods are significant, use the VAR(*p*max - 2), otherwise proceed to the next step and try *p*max - 3, etc.

FORECASTING WITH "VAR"

Forecasting in econometrics is done using time series variables. The idea is that using observed future is predicted.

To provide some notions for how forecasting is done, consider a ***VAR(1)*** involving two variables, ***y*** and ***x***:

$$\boldsymbol{y_t = \alpha_1 + \delta_1 t + \theta_{11} y_{t-1} + \beta_{11} x_{t-1} + e_{1t}} \qquad (15.6)$$

$$\boldsymbol{x_t = \alpha_2 + \delta_2 t + \theta_{21} y_{t-1} + \beta_{21} x_{t-1} + e_{2t}} \qquad (15.7)$$

$\boldsymbol{y_{t+1}}$ is not observable but it can be predicted. Using the Eq. (15.6) and setting ***t = T + 1***, following equation is obtained for $\boldsymbol{y_{t+1}}$:

$$\boldsymbol{y_{T+1} = \alpha_1 + \delta_1(T+1) + \theta_{11} y_T + \beta_{11} x_T + e_{1T+1}} \quad (15.8)$$

This equation cannot be directly used to obtain $\boldsymbol{y_{t+1}}$since we don't know what $\boldsymbol{y_{t+1}}$is. It is not clear that what unpredicTable shock or surprise will hit the economy next period. Furthermore, the coefficients are not known. But, if the error term (which cannot be predicted) is ignored and the coefficients are replaced by ***OLS*** estimates a forecast can be obtained which is denoted as $\boldsymbol{\hat{y}_{T+1}}$:

$$\boldsymbol{\hat{y}_{T+1} = \hat{\alpha}_1 + \hat{\delta}_1(T+1) + \hat{\theta}_{11} y_T + \hat{\beta}_{11} x_T} \quad (15.9)$$

The strategy for how to forecast one period into the future can be used for two periods on the condition that one is an extension. In the one period, $\boldsymbol{x_T}$ and $\boldsymbol{y_T}$ are used to create $\boldsymbol{\hat{y}_{T+1}}$ and $\boldsymbol{\hat{x}_{T+1}}$. In the two period, $\boldsymbol{\hat{y}_{T+2}}$ and $\boldsymbol{\hat{x}_{T+2}}$ subject to $\boldsymbol{y_{T+1}}$ and $\boldsymbol{x_{T+1}}$. Since the data runs until period ***T***, what is $\boldsymbol{y_{T+1}}$ and $\boldsymbol{x_{T+1}}$ are not known. Consequently $\boldsymbol{y_{T+1}}$ and $\boldsymbol{x_{T+1}}$ are replaced by $\boldsymbol{\hat{y}_{T+1}}$ and $\boldsymbol{\hat{x}_{T+1}}$. That is, use the relevant equation from the ***VAR***, is used ignoring the error, replacing the coefficients by their ***OLS*** estimates and replacing past values of the variables that are not observed by forecast.

In a formula:

$$\boldsymbol{\hat{y}_{T+2} = \hat{\alpha}_1 + \hat{\delta}_1(T+2) + \hat{\theta}_{11} \hat{y}_{T+1} + \hat{\beta}_{11} \hat{x}_{T+1}} \quad (15.10)$$

$\boldsymbol{\hat{x}_{T+2}}$ can also be calculated using the following formula:

$$\boldsymbol{\hat{x}_{T+2} = \hat{\alpha}_2 + \hat{\delta}_2(T+2) + \hat{\theta}_{21} \hat{y}_{T+1} + \hat{\beta}_{21} \hat{x}_{T+1}} \quad (15.11)$$

The general strategy which are ignoring the error, replacing coefficients by ***OLS*** estimates and replacing lagged values of variables that are unobserved by forecasts, can be used in order to obtain forecasts for any number of future periods for any ***VAR(p).***

OLS provides estimates only for coefficients which cannot be precisely corrected, it is also recommended to report confidence intervals.

VECTOR AUTOREGRESSIONS WITH COINTEGRATED VARIABLES

So far, we have assumed that all variables are stationary. If some of the original variables are non-stationary and are not cointegrated, then non-stationary variables should be differenced and used in the ***VAR***. This is valid except if the variables are non-stationary in levels and are cointegrated.

In Granger causality, it is recommended to work with an ***ECM***. Instead of working with a vector autoregression (***VAR***), a vector error correction model (***VECM***) should be employed. Like the ***VAR***, the ***VECM*** has one equation for each variable in the model.

If there are two variables, ***y*** and ***x***, the ***VECM*** is:

$$\Delta y_t = \varphi_1 + \delta_1 t + \lambda_1 e_{t-1} + \gamma_{11}\Delta y_{t-1} + \cdots + \gamma_{1p}\Delta y_{t-p} + \omega_{11}\Delta x_{t-1} + \cdots + \omega_{1q}\Delta x_{t-q} + e_{1t} \quad (15.12)$$

$$\Delta x_t = \varphi_2 + \delta_2 t + \lambda_2 e_{t-1} + \gamma_{21}\Delta y_{t-1} + \cdots + \gamma_{2p}\Delta y_{t-p} + \omega_{21}\Delta x_{t-1} + \cdots + \omega_{2q}\Delta x_{t-q} + e_{2t} \quad (15.13)$$

$$e_{t-1} = y_{t-1} - \alpha - \beta x_{t-1}. \quad (15.14)$$

The ***VECM*** is the same as a ***VAR*** with differenced variables, except for the term e_{t-1}. This error correction variable can be estimated by running an ***OLS*** regression of ***y*** on ***x*** and saving the residuals. Then ***OLS*** can be used to estimate ***ECM***s, and ***p-values*** and correspondent confidence intervals can be obtained. Lag length selection and forecasting also can be done in a similar way to the VAR, with the forecasts of the error correction term, e_t must be calculated.

In summary, this is simple using ***OLS*** estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and replacing the error, $\boldsymbol{e_t}$, by the residual $\boldsymbol{u_t}$.

Note the following issues:

i- If ***x*** and ***y*** are stationary, standard statistical methods based on an ***ARDL*** model can be used to test for Granger causality.

ii- If ***x*** and ***y*** are non-stationary and are cointegrated, statistical methods based on an ECM can be used to test for Granger causality.

iii- Vector autoregressions have one equation for each variable. Each equation uses one variable as the dependent variable. The explanatory variables are lags of all the variables under study.

iv- ***VARs*** are useful for forecasting, and Granger causality test is useful for understanding the relationships between several series.

v- If all the variables in the ***VAR*** are stationary, ***OLS*** can be used to estimate each equation and standard statistical methods can be employed (***p-values*** and **t-statistics** can be used to test for significance of variables).

vi- If the variables under study non-stationary and are cointegrated, a variant on the ***VAR*** called the Vector Error Correction Model, or ***VECM***, can be used.

Example 15.1

Table 15.1 Eviews Output for VAR Analysis

Vector Autoregression Estimates
Sample (adjusted): 1/01/1991 12/29/2006
Included observations: 4174 after adjustments
Standard errors in ( ) & t-statistics in [ ]

	GREECE	FRANCE	TURKEY
GREECE(-1)	1.142527	0.007954	-0.240347
	(0.01585)	(0.01028)	(0.11145)
	[72.0657]	[0.77401]	[-2.15652]
GREECE(-2)	-0.143405	-0.004824	0.262410
	(0.01587)	(0.01029)	(0.11160)
	[-9.03349]	[-0.46878]	[2.35140]
FRANCE(-1)	0.190865	1.024415	0.679475
	(0.02523)	(0.01635)	(0.17736)
	[7.56520]	[62.6386]	[3.83107]
FRANCE(-2)	-0.190844	-0.029875	-0.709881
	(0.02518)	(0.01632)	(0.17704)
	[-7.57802]	[-1.83003]	[-4.00973]
TURKEY(-1)	-0.007488	-0.000914	1.016284
	(0.00228)	(0.00148)	(0.01602)
	[-3.28514]	[-0.61831]	[63.4212]
TURKEY(-2)	0.007555	0.001036	-0.015816
	(0.00228)	(0.00148)	(0.01604)
	[3.31170]	[0.70028]	[-0.98615]
C	1.872509	4.919013	24.41604
	(2.19474)	(1.42269)	(15.4287)
	[0.85318]	[3.45755]	[1.58251]
R-squared	0.998927	0.999085	0.999321
Adj. R-squared	0.998926	0.999083	0.999320
Sum sq. resids	7663596.	3220236.	3.79E+08
S.E. equation	42.88491	27.79919	301.4754
F-statistic	646709.7	757954.6	1021740.
Log likelihood	-21607.21	-19797.73	-29747.21
Akaike AIC	10.35659	9.489566	14.25693
Schwarz SC	10.36722	9.500193	14.26756
Mean dependent	2170.302	2227.884	9341.112
S.D. dependent	1308.410	918.1297	11559.02

The corresponding equations are;

GREECE = C(1)*GREECE(-1) + C(2)*GREECE(-2) + C(3)*FRANCE(-1) + C(4)*FRANCE(-2) + C(5)*TURKEY(-1) + C(6)*TURKEY(-2) + C(7)

(15.15)

FRANCE = C(8)*GREECE(-1) + C(9)*GREECE(-2) + C(10)*FRANCE(-1) + C(11)*FRANCE(-2) + C(12)*TURKEY(-1) + C(13)*TURKEY(-2) + C(14)

(15.16)

TURKEY = C(15)*GREECE(-1) + C(16)*GREECE(-2) + C(17)*FRANCE(-1) + C(18)*FRANCE(-2) + C(19)*TURKEY(-1) + C(20)*TURKEY(-2) + C(21)

(15.17)

Table 15.2 Eviews Output for VAR Analysis

Estimation Method: Least Squares
Sample: 1/01/1991 12/29/2006
Included observations: 4174
Total system (balanced) observations 12522

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	1.142527	0.015854	72.06569	0.0000
C(2)	-0.143405	0.015875	-9.033488	0.0000
C(3)	0.190865	0.025229	7.565201	0.0000
C(4)	-0.190844	0.025184	-7.578016	0.0000
C(5)	-0.007488	0.002279	-3.285137	0.0010
C(6)	0.007555	0.002281	3.311698	0.0009
C(7)	1.872509	2.194735	0.853182	0.3936
C(8)	0.007954	0.010277	0.774005	0.4389
C(9)	-0.004824	0.010290	-0.468784	0.6392
C(10)	1.024415	0.016354	62.63860	0.0000
C(11)	-0.029875	0.016325	-1.830032	0.0673
C(12)	-0.000914	0.001478	-0.618314	0.5364
C(13)	0.001036	0.001479	0.700279	0.4838
C(14)	4.919013	1.422688	3.457548	0.0005
C(15)	-0.240347	0.111451	-2.156518	0.0311
C(16)	0.262410	0.111598	2.351398	0.0187
C(17)	0.679475	0.177359	3.831065	0.0001
C(18)	-0.709881	0.177040	-4.009731	0.0001
C(19)	1.016284	0.016024	63.42118	0.0000
C(20)	-0.015816	0.016038	-0.986153	0.3241
C(21)	24.41604	15.42870	1.582508	0.1136
Determinant residual covariance		1.11E+11		

Equation: GREECE = C(1)*GREECE(-1) + C(2)*GREECE(-2) + C(3)
*FRANCE(-1) + C(4)*FRANCE(-2) + C(5)*TURKEY(-1) + C(6)*TURKEY(
-2) + C(7)
Observations: 4174

R-squared	0.998927	Mean dependent var	2170.302
Adjusted R-squared	0.998926	S.D. dependent var	1308.410
S.E. of regression	42.88491	Sum squared resid	7663596.
Durbin-Watson stat	1.995159		

Equation: FRANCE = C(8)*GREECE(-1) + C(9)*GREECE(-2) + C(10)
*FRANCE(-1) + C(11)*FRANCE(-2) + C(12)*TURKEY(-1) + C(13)
*TURKEY(-2) + C(14)
Observations: 4174

R-squared	0.999085	Mean dependent var	2227.884
Adjusted R-squared	0.999083	S.D. dependent var	918.1297
S.E. of regression	27.79919	Sum squared resid	3220236.
Durbin-Watson stat	1.998785		

Equation: TURKEY = C(15)*GREECE(-1) + C(16)*GREECE(-2) + C(17)
*FRANCE(-1) + C(18)*FRANCE(-2) + C(19)*TURKEY(-1) + C(20)
*TURKEY(-2) + C(21)
Observations: 4174

R-squared	0.999321	Mean dependent var	9341.112
Adjusted R-squared	0.999320	S.D. dependent var	11559.02
S.E. of regression	301.4754	Sum squared resid	3.79E+08
Durbin-Watson stat	2.000858		

REFERENCES

Baltagi, B.H. (2008).*Econometric Analysis of Panel Data*. Fourth Edition. Great Briatin: John Wiley & Sons

Fischer, B. (1976). The pricing of commodity contracts, *Journal of Financial Economics* 3.p.167-179.

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* 31.p.307-327.

Box,G.E.P. & Pierce, D.A. (1970). Distribution of the Autocorrelations in Auotoregressive moving Average Time Series Models. *Journal of American Statistical Association* 65. p. 1509-1526.

Breusch, T.S. (1979). Testing for Autocorrelation in Dynamic Linear Models. *Australian Economic Papers* 17.p. 334–355.

Breusch, T.S.&Pagan, A.R.(1980). The Lagrange Multiplier Test and Its Applications to Model Specification in Econometrics. *Review of Economic Studies* 47(1). p. 239-253.

Brown, R.L.; Durbin, J. & Evans, J.M. (1975). Techniques for Testing the Constancy of Regression Relationships over Time. *Journal of the Royal Statistical Society*. Ser.B. 37.p.149 – 192.

Charemza, W.W & Deadman, D.F.(1997). *New Directions in Econometric Practice*. Cheltenham: Edward Elgar.

Chow, G. C. (1960). Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica* 28 (3).p. 591–605.

Cobb, C. W.& Douglas, P. H. (1928). A Theory of Production. *American Economic Review* 18.p. 139–165.

Cochrane, J.H.(2001). *Asset Pricing*. USA: Princeton University Press.

Collis, J & Hussey, R. (2009). *Business Research: A Practical Guide for Undergraduate and Postgraduate Students.* Third Edition. China. Palgrave Macmillan

Dickey, D. A. & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association* 74 (366).p. 427–431.

Durbin, J. & Watson, G.S. (1950). Testing for Serial Correlation in Least-Squares Regression *Biometrika* 37. p.409-428.

Engle, R.F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation , Econometrica. 50(4).p. 987-1007.

Engle, R.F. (1982). Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U.K. Inflation. *Econometrica* 50.p. 987-1008.

Engle, R.F. & C.W.J. Granger (1987). Cointegration and Error Correction: Representation, Estimation and Testing. *Econometrica* 55.p.251-76.

Engle, R.F. & B.S. Yoo (1991). Cointegrated Economic Time Series: An Overview with New Results" in R.F. Engle and C.W.J. Granger (eds.), Long-run Economic Relationships: Readings in Cointegration, Oxford: Oxford University Press.

Fisher, R. A. (1948). Conclusions fiduciaires. *Ann. Inst. H. Poincaré* 10. p. 191–213.

Fox, J.H.(1958). Criteria of Good Research. *Phi Delta Kappan*. Vol. 39. P. 285–86.

Goldfeld, S. M. & Quandt, R. E. (1965). Some Tests for the Homoscedasticity. *Journal of the American Statistical Association* 60. p. 539-547.

Godfrey, L.G. (1978). Testing for Higher Order Serial Correlation in Regression Equations when the Regressors Include Lagged Dependent Variables. *Econometrica* 46. p.1303-1310.

Glejser, H. (1969). A New Test for Heteroscedasticity. *Journal of American Statistical Association*. 64. p. 316-323.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *Journal of Finance* 48.p. 1779-1801.

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* 37 (3).p. 424–438.

Granger, C.W.J. (1981). Some Properties of Time Series Data and Their Use in Econometric Model Specification. *Journal of Econometrics* 16.p. 121-30.

Harvey, A.C. (1976). Estimating Regression Models with Multiplicative Heteroscedasticity. *Econometrica* 44(3).p. 461

Hausman, J.A. (1978). Specification Tests in Econometrics. *Econometrica* 46 (6).p. 1251-1271.

Jarque, C. M. & Bera, A. K. (1981). Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals: Monte Carlo Evidence. *Economics Letters* 7 (4).p. 313–318.

Johansen, S. (1988). Statistical Analysis of Cointegration Vectors. *Journal of Economic Dynamics and Control* 12.p.231-254.

Johansen, S. (1991). Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. *Econometrica* 55.p. 1551-1580.

Johansen, S.& Juselius, K. (1994). Identification of the Long-Run and the Short Run Structure. An Application to the ISLM Model. *J. Econom* 63.p. 7-36.

Koop, G. (2005). Analysis of Financial Data. Second Edition. England: John Wiley &Sons.

Kothari, C.R.(2004). *Research Methodology: Methods and Techniques*. Second Edition. New Age International

Kumar, R. (2011). *Research Methodology: A Step-by-Step Guide for Beginners*. Third Edition. Great Britain. SAGE.

Lee, H.B & Kerlinger, F.N. (2000) *Foundations of Behavioral Research*. Third Edition. USA, Earl McPeek.

Ljung, G.M. & Box, G.E.P.(1978). On a measure of Lack of Fit in Time Series. *Biometrica* 65. p. 297-303.

Nelson, D.B. (1990). ARCH models as diffusion approximations, *Journal of Econometrics* 45.p. 7 - 38.

Phillips, P. C. B. & Ouliaris, S. (1990). Asymptotic Properties of Residual Based Tests for Cointegration. *Econometrica* 58, 165–193.

Phillips, P.C.B. (1991). Optimal Inference in Cointegrated Systems. *Econometrica* 59.p. 283-306.

Ramsey, J.B. (1969). Tests for Specification Errors in Classical Linear Least Squares Regression Analysis. *Journal of the Royal Statistical Society, Series B.*, 31(2), 350–371.

Savin, N.E. & White, K.J. (1977). The Durbin-Watson Test for Serial Correlation with Extreme Sample Sizes or Many Regressors. *Econometrica* 45.p. 1989-1996.

Stock, J.H. & Watson, M.W. (1988). Testing for Common Trends. *Journal of the American Statistical Association* 83.p. 1097-1107.

Stock, J.H & Watson, M.W. (2008). *Introduction to Econometrics*. Brief Edition. USA: Greg Tobin

Verbeek, M.(2004). *A Guide to Modern Econometrics*. Second Edition.England: John Wiley & Sons.

Vogelvang, B. (2005). *Econometrics Theory and Applications with Eviews*. England: Pearson

Wald, A. (1943). Tests of Statistical Hypotheses Concerning Several Parameters when the Number of Observations is Large. *Transactions of the American Mathematical Society* 54. p. 426-482.

White,H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica* 48 (4).p.817-838.

White, H. (1980). A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity. *Econometrica*. 48.p.817-838.

Standard Normal Curve Probability Distribution

The table is based on the upper right 1/2 of the Normal Distribution; total area shown is 0.5

The Z-score values are represented by the column value + row value, up to two decimal places

The probabilities up to the Z-score are in the cells

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

Critical Values of t-Distribution

The table shows the critical t-values for a given alpha level (one-tailed) and degrees of freedom

The degrees of freedom are the rows (denoted by df)

Note: The probability levels represent the whole of alpha (you must divide alpha by 2 if you want the t-value for a two-tailed test)

df	0,1000	0,0500	0,0250	0,0100	0,0050	0,0010	0,0005
1	3,078	6,314	12,706	31,821	63,657	318,309	636,619
2	1,886	2,920	4,303	6,965	9,925	22,327	31,599
3	1,638	2,353	3,182	4,541	5,841	10,215	12,924
4	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	1,476	2,015	2,571	3,365	4,032	5,893	6,869
6	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	1,415	1,895	2,365	2,998	3,499	4,785	5,408
8	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	1,330	1,734	2,101	2,552	2,878	3,610	3,922
19	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	1,319	1,714	2,069	2,500	2,807	3,485	3,768
24	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	1,310	1,697	2,042	2,457	2,750	3,385	3,646
40	1,303	1,684	2,021	2,423	2,704	3,307	3,551
50	1,299	1,676	2,009	2,403	2,678	3,261	3,496
60	1,296	1,671	2,000	2,390	2,660	3,232	3,460
100	1,290	1,660	1,984	2,364	2,626	3,174	3,390
120	1,289	1,658	1,980	2,358	2,617	3,160	3,373
Infinity	**1,282**	**1,645**	**1,960**	**2,326**	**2,576**	**3,090**	**3,291**

F Table for Alpha=0.5

alpha=.05

df lower \ df upper	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	inf
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,9	245,9	248,0	249,1	250,1	251,1	252,2	253,3	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
inf	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00