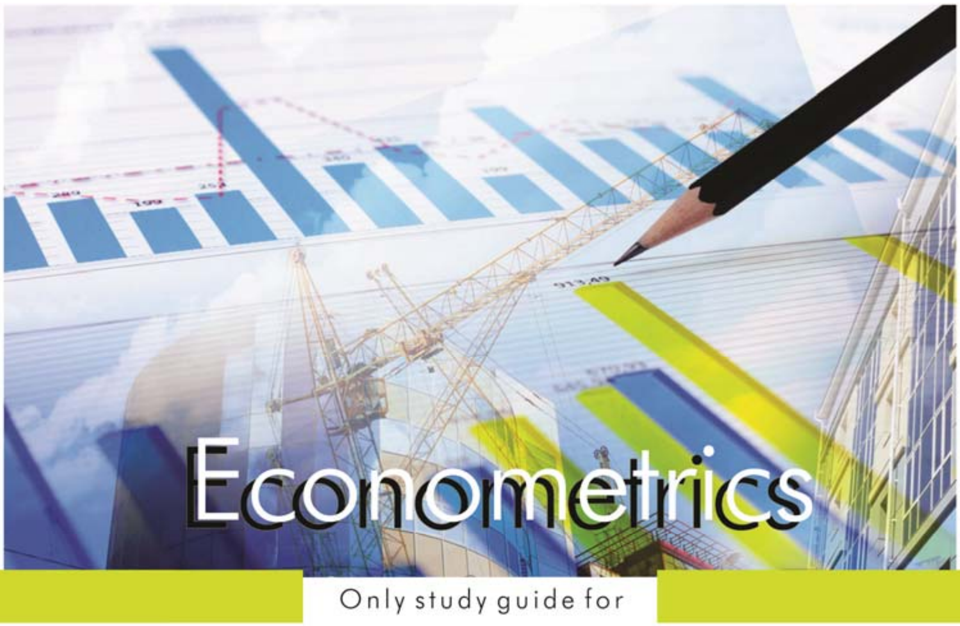


Econometrics

Only study guide for

ECS3706

Compiler
Mr A Bijker

Previous revision by
Dr S Nhamo

Current revision by
Mr MJ Khumalo

Department of Economics

University of South Africa
Pretoria

Printed and published by the
University of South Africa
Muckleneuk, Pretoria

ECS3706/1/2018–2020

70611823

CONTENTS

INTRODUCTION TO ECS3706

WELCOME MESSAGE

Dear Student

We are pleased to welcome you to the Econometrics (ECS3706) module and hope that you will find it both interesting and rewarding. We shall do our best to make your study of this module successful. You will be well on your way to success if you start studying early in the semester and resolve to do the assignments properly. Econometrics is very useful to all who do economic research and analysis and provides a basis for understanding economic literature.

The ECS3706 module is offered by the Department of Economics and is intended for students pursuing a career in the field of economics. It is a semester module that carries 12 credits towards your qualification. This module introduces students to statistical methods used in economics and covers topics such as regression analysis, hypothesis testing etc.

We will use the myUnisa module website (for students who can go online) and the ECS3706/1 study guide to direct you through the various sections.

For this module there is a prescribed textbook, which you will need in order to make your study of this module easy as this document was compiled based on that particular textbook. It is therefore your responsibility to make sure you get a copy of the prescribed book. The material on myUnisa and in the ECS3706/1 study guide is intended to guide you through the textbook and does not replace the textbook.

INTRODUCTION TO ECS3706

Econometrics, like mathematics and statistics, must be properly understood and cannot be learnt by heart. Students who have previously completed modules in mathematics and/or statistics may find this module relatively easy. For the others: you must come to grips with the mathematical language and the statistical principles underlying econometrics. This may add significantly to the already loaded content of this module. It is possible to master the module if, at the very least, you have a sound understanding of grade 12 mathematics.

This module makes extensive use of mathematical notation, which may appear difficult to those not familiar with it. However, once mastered, it proves to be a very concise and efficient form of language. We expect you to master mathematical notation and to use it in the examination.

The approach of the textbook, and of this module, is descriptive and practical, rather than mathematically rigorous. The textbook is simple, intuitive and relatively easy to understand. In fact, it is the most adopted text (United States of America) for introductory econometrics courses. Concepts are explained in nonmathematical terms and by way of practical examples. In contrast, many other books on econometrics focus strongly on statistics, which unfortunately tends to discourage and exclude those students who do not have a solid mathematical or statistical background.

THE PURPOSE OF THE MODULE

The purpose of this module is to gain insight into

- what constitutes regression analysis
- the role of statistics in econometrics
- choosing the appropriate functional form
- how to deal with econometric problems

This module will equip you with the required skills to estimate basic economic functions and evaluate the models.

LEARNING OUTCOMES AND ASSESSMENT CRITERIA

Outcomes	Assessment criteria
Demonstrate an understanding of the purpose and method of econometrics	• Explain the goal and method of econometrics and its three main uses. • Explain whether the specification of a regression equation is mainly objective or subjective. • Use the standard notation of the multivariate linear regression model to explain the meaning of dependent and independent variables, the meaning of parameters of an equation, the difference between causality and correlation, the meaning of the stochastic error term and its main sources of variation. • Explain the meaning and purpose of an ordinary least squares (OLS) estimator, the principle of the OLS estimation technique and the advantage of the OLS estimation technique. • Explain the meaning of degrees of freedom and discuss the meaning of the total, explained and residual sum of squares and how they are related to R^2 (R-squared) and $\overline{R}^2$ (adjusted R-squared). • Explain and execute the steps in an applied regression analysis and interpret the meaning of regression coefficients.
Review the principles of statistics and statistical measures and understand the sampling distribution of the OLS estimator and the power and limitation of hypothesis testing	• Explain the meaning and purpose of each of the seven classical assumptions using standard notation. • Discuss the rationale of including a stochastic error term in a regression equation and the factors that contribute to the error term. • State the relevance of the central limit theorem for the distribution of the error term. • Explain whether or not an unbiased estimator is necessarily better than a minimum variance estimator. • Discuss the meaning and relevance of the Gauss-Markov theorem and a best linear unbiased estimator (BLUE). • Explain the goal and method of hypothesis testing and discuss the application of one-sided and two-sided t-tests, setting the null and alternative hypothesis based on theoretical considerations and calculation of t-values or F-values, looking up critical t-values or F-values in statistical tables and drawing appropriate conclusions.
Identify and describe specification problems, choosing independent	• Discuss the criteria which should be used to select the independent variables in an equation, discuss the problems of selecting variables and explain the impact of an irrelevant variable in a regression equation. • Explain the meaning of a constant term.

variables and choosing a functional form	• Discuss the circumstances under which the linear form, the double-log form, semi-log forms, polynomial forms and inverse forms can be used. Provide the algebraic expressions for each and describe the key characteristics for each of them. • Specify, use and interpret slope dummies and intercept dummies.
Identify and remedy econometric problems of multicollinearity, serial correlation and heteroscedasticity	• Explain the nature of the multicollinearity problem and distinguish between perfect and imperfect multicollinearity; consequences of multicolinearity and how to detect it. • Distinguish between pure and impure serial correlation; discuss consequences and test correlation using the Durbin-Watson test. • Discuss methods of testing heteroskedasticity. • How to remedy multicollinearity, serial correlation and heteroskedasticity in regression.

SUMMARY OF MODULE AND CHAPTER OUTCOMES

Module outcomes Part	Prescribed Textbook	Learning units	Learning unit outcomes
(1) Understanding the purpose and method of econometrics	Page 1	Learning unit 1	Understanding that the essence of econometrics is estimating a regression equation. This requires economic theory, data and statistics
	Page 35	Learning unit 2	Understanding what OLS does, performing OLS in practice and interpreting its output
	Page 71	Learning unit 3	Learning to use regression analysis step by step
(2) Applying statistics with confidence	Page 427	Learning unit 12	Reviewing the principles of statistics and statistical measures
	Page 97	Learning unit 4	Understanding the statistical requirements of OLS and the sampling distribution of the OLS estimator
	Page 127	Learning unit 5	Understanding the power and limitations of hypotheses testing
(3) Specification	Page 177	Learning unit 6	How to choose the independent variables of the regression equation
	Page 219	Learning unit 7	How to choose a functional form
(4) Dealing with econometric problems	Page 261	Learning unit 8	Multicollinearity
	Page 321	Learning unit 9	Serial correlation
	Page 389	Learning unit 10	Heteroskedasticity
	Page 357	Learning unit 11	Running your own regression project

THE PRESCRIBED TEXTBOOK

A prescribed book is the compulsory book that you need to have when studying this module. The prescribed book for this module was compiled from 2010 and 2014 editions of AH Studenmund's "*Using econometrics: a practical guide.* 6th edition. Occidental College: Pearson and customised for UNISA students. The details of the prescribed textbook used in this module are as follows:

Using Econometrics: A practical guide. Harlow: Pearson Education. Published in 2015

(ISBN: 9781784476908)

THE RECOMMENDED TEXTBOOK

The recommended books are not necessarily compulsory but are regarded as the extra books for reading should you wish purchase them. The recommended textbooks for this module are as follows:

Gujarati, DN & Porter, DC. 2009. *Basic econometrics*. 5th edition. New York: McGraw-Hill.

Stock, JH & Watson, MW. 2015. *Introduction to econometrics*. 3rd edition. Harlow: Pearson Education.

THE ECS3706 STUDY GUIDE

The ECS3706 study guide (which is this very document you are reading now) leads you through the prescribed sections in the textbook in a systematic way. You are therefore discouraged from studying the textbook without consulting the study guide.

The purpose of this document is

- certainly NOT to replace the textbook. This is not a self-contained guide, nor does the ECS3706 study guide provide a short summary of the module.
- to outline the module, that is, to indicate which sections are prescribed.
- to provide additional explanations of difficult material.
- to provide some adaptation to South African conditions (the textbook was designed within the context of American economy).
- to strengthen understanding by means of questions and answers, and by providing practical problems.
- to indicate what type of questions can be expected in the examination.

The ECS3706 study guide and textbook must be used together. This document charts a route through the textbook, which is highly readable and relatively easy to understand. Each learning unit in the ECS3706 study guide has the following sections:

Contents

A Prescribed material
B Some important concepts

This section includes a number of practical activities which are meant to reinforce your understanding of the work.

Questions and answers

C True/False questions

Examination questions

D Paragraph questions
E Practical questions

The section on examination questions provides you with representative examination questions. These questions will provide a snapshot of how questions are asked in this module. The examination consists of two sections. Section A requires you to answer paragraph-type questions. They deal mainly with the theory of econometrics. Section B consists of practical problems which require you to apply what you have learnt. You could, for example, be given a regression result which you are required to evaluate, or you could be asked to solve a practical problem.

The purpose of this section is to provide you with study goals and to get you used to examination questions right from the start. Mastering this module and preparation for the examination are one and the same action.

Learning units and chapters

The work is divided into 12 learning units that cover all the prescribed chapters in the prescribed textbook. The numbering of the learning units in this document corresponds to the numbering of chapters in the textbook. For example: Learning unit 2 in the ECS3706 study guide corresponds with chapter 2 of the textbook. In each learning unit, you will encounter a section entittled *Econometrics in action*. In this section, an attempt is made to illustrate how econometrics affects our everyday lives.

Study method

The following study method is suggested:

- Use the ECS3706 study guide in parallel with the prescribed textbook. Study the prescribed sections of the textbook.
- Review what you have learnt by performing the activities and answering the true/ false questions in the ECS3706 study guide. You can gain much by doing the activities and some of the practical exercises at the end of each chapter.
- Finally, test yourself by answering the examination questions. The past examination papers are available on myUnisa under **Official Study Material**.

STATISTICS

This module makes extensive use of statistical concepts. We assume, for example, that the data used in regression analysis are random samples drawn from the population. This requires the use of other statistical concepts, like the sampling distribution of an estimator, standard errors and hypotheses testing. You will need to fully understand the meaning of these concepts.

Chapter 15, which is part of the syllabus, focuses on the underlying statistics and explains them in detail. Students with no prior knowledge of statistics may find this particularly helpful. It provides a solid statistical foundation and will enable you to fully understand the concepts. Even students who have previously completed statistics modules may find this chapter useful for brushing up their knowledge.

COMPUTERS

Students of ECS3706 must have access to a personal computer (PC). The most common technique of estimating regression equations is called ordinary least squares (OLS). Even in its simplest form, OLS requires extensive calculations. In the case of estimating multivariate regression equations, the formulas become increasingly complex. The use of a PC drastically reduces the computational load and will allow you to focus on the meaning and interpretation of regression results.

The use of PCs has another advantage. Econometrics is a practical subject and students often learn a lot by solving practical problems. Like flying a plane, important aspects of econometrics are learned by doing.

Please note that in the examination there will be no need for extensive calculations, nor a PC.

Software

Studenmund recognises the educational value of students running their own regressions. No special software is required as most of the techniques can be applied in Microsoft Excel (or any other spreadsheet). We recommend that students use Microsoft Excel to perform regression analysis.

An alternative to Microsoft Excel (but not recommended for this course) is the student version of EViews. The sixth edition of Studenmund encourages students to use EViews. EViews is a powerful and user-friendly econometrics package used by many professionals. In fact, it is the most commonly used econometrics package worldwide.

In this module, we recommend the use of MS Excel and not EViews. The reason for this is the cost of EViews. Since most PC users already have Excel installed, they may as well use it. A spreadsheet has the capacity to perform basic regression analyses quite well. Of course, a spreadsheet is a general purpose tool and it lacks the many specialised functions offered by econometrics packages. This is, however, not that important in an introductory econometrics module.

Whether you choose to use Excel or EViews, we encourage you to perform practical regression analyses, as this will improve your feel for econometrics. It allows you to explore different options quickly and effortlessly. And your future employer and colleagues may be quite impressed by your PC skills and your ability to perform regression analyses purposefully and smoothly.

Although the examination does not require a PC, we expect you to be familiar with a spreadsheet (Excel) when completing Assignment 01. Basic spreadsheet skills, for example entering formulas and using copy and paste, are not dealt with in the ECS3706 study guide. If you have never used a spreadsheet before, then this is the time to familiarise yourself with one of the most powerful calculation tools ever devised.

If you choose to use EViews, you must teach yourself how to use it.

Studenmund data

Studenmund uses many practical examples for which the data files may be downloaded from the web. Go to http://wps.aw.com/aw_studenmund_useecon_6/ and then select "Student resources", "Data Sets" and the relevant chapter. The data files are available in three formats (EViews, Excel and ASCII text format). If you use MS Excel then you would use "Excel". The Excel format can also be read by Quattro-Pro. If you use Lotus 1-2-3 you may use the ASCII text format.

Contact your lecturers if you experience problems in accessing and/or using these data files.

ASSIGNMENTS

There are two assignments in this module. They can be found in Tutorial Letter 101 of ECS3706. Although the completion of both assignments is not compulsory for admission to the examination, it will enhance your understanding of econometrics. Remember, econometrics is a practical module, so you will learn by doing. We strongly recommend that you complete all the assignments as you proceed with the module. To help you, the assignment is structured according to learning units.

SCOPE AND LIMITATIONS OF THE MODULE

This module is an introduction to the theory and practice of econometrics. It provides a basic understanding of the purpose and approach of econometrics. In a nutshell, econometrics uses regression analysis to estimate the coefficients of an economic equation. Everything in this course is focussed on providing estimates that are as accurate as possible.

The module is restricted to single equation models (multiple, or simultaneous equation models also exist). Conventional econometric problems such as specification, multicollinearity, serial correlation and heteroskedasticity are dealt with. It also excludes more advanced topics like time-series models (and cointegration and error correction models) and dummy dependent variable techniques.

PART I

INTRODUCTION TO REGRESSION ANALYSIS

Econometrics is mainly about regression analysis. But what is regression analysis? And how is it applied?

We deal with these issues in three learning units:

- Learning unit 1: An overview of regression analysis
- Learning unit 2: Ordinary Least Squares (OLS)
- Learning unit 3: Learning to use regression analysis

LEARNING UNIT 1

AN OVERVIEW OF REGRESSION ANALYSIS

ECONOMETRICS IN ACTION

Lawrence R Klein is a 1980 Nobel Laureate in Economics for the creation of econometric models.

Ragnar Bentzel, a Swedish economist, said Klein's "Link" project is his "crowning achievement". The project, begun in the late 1960s, aims to coordinate econometric models of various countries to help forecast international trade and capital movements. The model is used to show such things as how an increase in oil prices influences inflation, employment and trade balances. Klein's subsequent work, the "Wharton Econometric Forecasting Model", has become a standard tool for economic forecasters.

How do you win the Nobel prize in economics? Klein has the following to say about his undergraduate training: *The completion of my undergraduate training at the University of California (Berkeley) provided just the needed touches of rigor at advanced levels in both economics and mathematics.*

Do you agree with Klein that a solid foundation in economics and mathematics is invaluable in econometrics? If you cannot make an informed decision now, you hopefully will be able to do so once you have worked through this guide.

STUDY OBJECTIVES

This learning unit explains regression analysis. It looks at its purpose, its uses and its methods.

When you have studied this learning unit you should understand

- the more important uses of econometrics
- the major foundations of econometrics
- regression analysis and the meaning of each of the components of a typical regression equation

(A) PRESCRIBED MATERIAL

This learning unit serves as an introduction to what econometrics is all about, and provides a short outline of its uses.

The following sections are prescribed:

(1) What is econometrics?
(2) What is regression analysis?
(3) The estimated regression equation.

The sections

(4) A simple example of regression analysis,
(5) Using regression to explain housing prices, and providing examples of the basic technique of regression analysis. Although no direct examination questions will be asked from these sections, they are useful, as they give practical examples of concepts introduced in previous sections.

(B) SOME IMPORTANT CONCEPTS

1.1 What is econometrics?

To explain what econometrics is one can look at its goal, its method and its uses. The econometric approach is represented schematically in figure 1. The goal of econometrics is the estimation of economic relationships. Its method is mainly regression analysis using actual data. Econometric models are used mainly for describing economic reality, hypothesis testing, simulation and forecasting.

Econometrics makes use of the following inputs/disciplines:

- economic theory
- economic data
- statistics

Economic theory

In previous economics modules you encountered the following two economic relationships:

- the demand curve: $P = a + bQ_d$ from microeconomics (P: price, Q_d: quantity demanded)
- the consumption function: $C = C + cY$ from macro-economics (C: private consumption, Y: income, c: marginal propensity to consume)

In the previous modules, both were treated in a theoretical fashion. The coefficients of these equations (e.g. a and b in the demand curve) were simply assumed to have some predetermined values. To be of real use, however, one requires accurate estimates thereof. Econometrics provides methods to estimate these coefficients, using actual data.

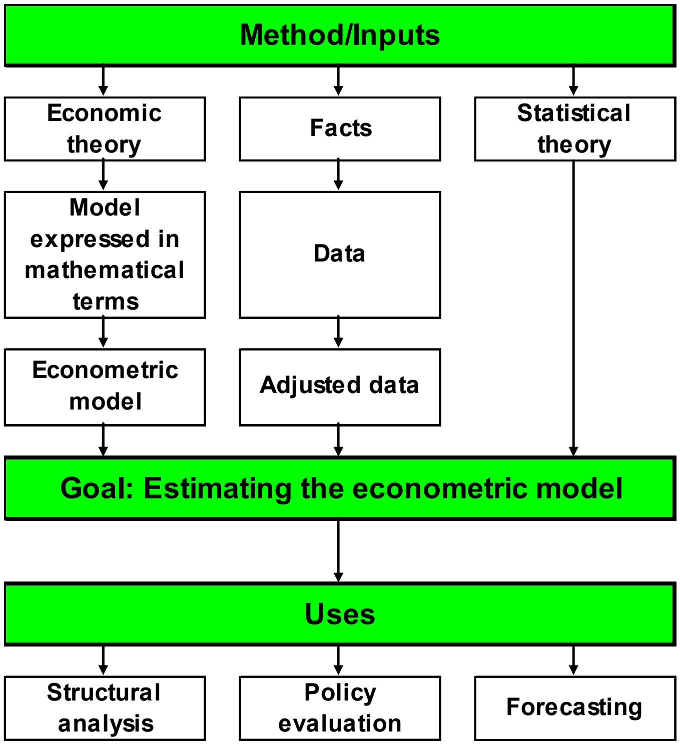


FIGURE 1: The econometric approach and its uses

The theory of economics is important in econometrics, as it provides knowledge about the nature of economic relationships. It helps us to choose the variables as well as the functional form to be used (for example y = a + bX or log(Y) = a + bX). In this module we assume that you are familiar with intermediate microeconomic and macroeconomic models and their specification.

The process of converting economic theory into a mathematical form is called the specification of a model. The specification of an equation involves the selection of the dependent variable (the Y-variable), the variable(s) that cause(s) the effect (the X-variables), as well as the functional form.

The specification of a model can be simple in some cases, and more difficult in others. Monetary theory provides us with a simple specification problem. For many years, in the field of monetary theory, the relationship between income (Y) and the money stock (M) was a hotly debated issue where Y is the real level of economic activity and M reflects the money stock. Of course, changes in M arise mainly because of the amount of net new loans created by banks. The main issue was: does M → Y, or does Y → M? This, of course, affects the way in which the model is specified. The two schools of thought were the monetarists and the post-Keynesians.

- The monetarists believed that Y = f (M) where the direction of causality runs from M → Y and where M is controlled by the central bank. The monetarist transmission mechanism is both direct (for example ΔM affects the prices of assets which affects real spending) and indirect (ΔM affects the interest rate which induces changes in real investment).

- The post-Keynesians believe that M = f(Y), which is the current generally accepted view. The direction of causality is Y → M, which is opposite to the monetarist case. The post-Keynesians believe that M cannot be controlled. If Y increases, this causes an increase in the demand for M in order to finance the increased level of Y.

Econometrics depends on economic theory to provide the variables involved, the direction of causality and the nature of the functional form. Econometrics cannot resolve theoretical differences between different schools of thought. Causality depends only on theory. Econometrics can only determine correlation, which is the strength and nature of a relationship. It cannot say anything about causality. For example, econometrics will happily estimate both forms Y = a + bM and M = c + dY. If there is a good correlation between Y and M, then both forms will indicate a good fit.

Even so-called causality tests – which take into account which variable changed first – cannot prove conclusively the true nature of causality. A simple example is the purchase of wedding rings. Since the purchase of wedding rings normally precedes a marriage, marriages could be seen as being caused by the purchase of wedding rings, which is of course nonsense.

Economic data

In practice an econometrician must know the sources of economic data, and be aware of methods of adjusting data to suit the need. Some knowledge of the methods of compiling economic data for example, the system of national accounts is valuable. The second-year module in Economic Indicators (ECS2603) is useful, but is not a prerequisite for this module.

In econometrics, we use either time-series data, where the same variable is measured over time (e.g. the real GDP for the period 1960–1996). A subscript “t” is conventionally used to show time series data, or cross-sectional data, which provide a measure of several variables at a point in time (e.g. population census data for various provinces as of 1 March 2003). Subscript “i” is used to denote cross section data.

Data are often adjusted in order to enhance their use. Examples of adjusted data are the following:

- A price index time series is adjusted relative to the price of a base year, for example the consumer price index is expressed as 2005 = 100.
- It is often more revealing to view the annual percentage change of a variable, rather than its level. A common example is the inflation rate calculated from the consumer price index.
- Time-series data, which are significantly affected by seasonal influences, are often adjusted to remove the seasonal effect. These are called seasonally adjusted data.
- In order to remove the effect of changes in prices, values are often deflated (calculated at constant prices), for example in the case of real GDP, which allows us to compare the levels of real output of different years, without being affected by a change in the price level.

In contrast to the physical sciences, virtually all economic data are of the non-experimental type. In the physical sciences, experimental data may be collected under controlled conditions. The researcher may isolate the effects of a specific variable of interest, because all other factors may be held constant. In econometrics, controlled experiments are almost impossible. The data observed is of the non-experimental type, which reflects the combined impact of many variables simultaneously. It is left to the econometrician to suggest causality, that is, cause and effect relationships between variables.

Statistics

Econometrics makes extensive use of statistical techniques. Examples are the technique of regression analysis and hypothesis testing.

Econometricians must be familiar with statistical concepts such as a sampling distribution, the normal distribution, t-tests, the expected value of a sample estimate, standard errors and more. These matters are addressed in chapter 17. The focus of this module is on teaching students how to interpret and use these. The focus is not on deriving estimators from first principles as is often the case in econometrics courses.

Because of the unique nature of economic data and/or models, special statistical techniques have been developed to cope with these difficulties. The latter part of the module focuses on these issues, that is, multicollinearity, serial correlation and heteroskedasticity.

1.2 Uses of econometrics

The main uses of econometrics are for structural analysis, forecasting and policy evaluation.

Econometric models are used at different levels and degrees of complexity.

1. Klein's "Link" project, referred to in the introduction, aims to coordinate econometric models of various countries to help forecast international trade and capital movements.
2. Most central banks have large, complex models of their national economies. These are mostly simultaneous-equation type of models which are used to direct monetary and fiscal policy. The South African Reserve Bank uses an econometric model to forecast inflation. The current inflation targeting monetary policy framework is inherently forward looking and relies heavily on the forecasts provided by this model.
3. The department of finance uses an econometric model to forecast tax income and to simulate the effects of alternative policy options.
4. Commercial banks use econometric models to better understand how different economic sectors and industries may react to shocks on the economy.
5. Simple type of models may be used by business and industry for forecasting and planning. A firm might use, for example, a single-equation model to determine the impact of its advertising on sales.

Structural analysis

Structural analysis entails the quantification of economic relationships. A simple example from second-level macro-economics is the consumption function which is specified as:

$$C = \overline{C} + cY$$

where C: consumption expenditure, $\overline{C}$: autonomous consumption, Y: disposable income, and coefficient c reflects the marginal propensity to consume. The value of c is a *structural coefficient* which indicates the tendency of consumers to spend additional income on consumer goods and services (the marginal propensity to consume). Structural analysis means that we gain quantitative knowledge about relationships between economic variables.

Policy evaluation

Economic models can be used by government to compare the effects of policy measures. Alternative policy instruments are quantified and fed into an econometric model. The model is solved to provide a quantitative outcome for each policy option.

Forecasting

Forecasting entails a forward simulation of an econometric model. Assumptions are made regarding the exogenous variables (the level of government expenditure, the gold price, the growth of overseas economies, etc.) and these are fed into the model, which then provides forecasts of the endogenous variables (income, private consumption, etc.).

Although many other means of forecasting and simulation do exist (and can be successful), there are advantages to using an econometric model. A good econometric model represents a logically consistent structure. The critically important variables have already been identified and the relationships between the variables have been quantified. Using this model has the advantage that the results are internally consistent, meaning that no important factors are ignored and that proper account has been taken of the interrelationships between variables. The econometric approach to forecasting is particularly useful in the medium to longer term, when structural relationships are more dominant than short-term or random effects.

1.3 What is regression analysis?

Assume that you have two datasets, a series of private consumption expenditure (CONS) and a series of gross domestic product (GDP) data for, say, the period 1980–2009. You believe that there is a macro relationship between CONS and GDP (the consumption function) such that CONS = a + bGDP where a and b are coefficients of the equation. Regression analysis is simply a technique to estimate the coefficients of the equation from the data.

A typical regression equation with more Xs is as follows

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \varepsilon_i$$

Make sure you understand the meaning of dependent and independent variables, the meaning of the coefficients of a regression equation and the meaning of the stochastic error term. Also make sure you are familiar with regression notation as dealt with in Section 1.2. of Studenmund.

Regression equations (in this module) assume that the direction of causality runs from the Xs to the Y-variable, that is, changes in the Xs cause changes in Y.

The estimated and true regression equation

Section 3 (p. 15 of prescribed book), focuses on the difference between the true regression equation and the estimated regression equation. Make sure you understand figure 3, which summarises these concepts. The two basic concepts are

- the true (population) parameters: β
- the estimated (sample) parameters: $\hat{\beta}$

The true (population) coefficients in the relationship

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \varepsilon_i$$

represents the true relationship between the Xs and Y where the εs take care of deviations in Y not explained by the Xs. Note that the βs are unknown.

The terms population and sample originate from an area of statistics called sampling theory. A sample is a subset drawn from the much larger population. The population is usually too large or costly to enumerate in full, or it may be a theoretical concept that represents all possible outcomes which cannot be observed. In this module, we always apply regression analysis to sample data.

Two concepts that run parallel to that of β and $\hat{\beta}$ are

- the stochastic error term (ε_i) which is the true but unobserved error term as in $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- the residual error term (e_i) which occurs in the estimated equation $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$

1.4 A simple example of regression analysis

This section uses a weight guessing equation:

$$\text{Weight}_i = 103.40 + 6.38(\text{Height}_i) + \varepsilon_i$$

to estimate weight, given the height. Note that height cannot be a perfect predictor of weight because there is variation in weight caused by other factors than height. It is still, however, a useful equation as it improves one's ability to predict weight. In economics we follow much the same approach. We try to estimate Y-variables given a number of Xs. Although the predictions for Y, given the Xs, are not perfect, we do try to account for explanatory variables (Xs), which we believe, are the major determinants of Y.

1.5 Using regression analysis to explain housing prices

This section introduces the house price equation (an example of cross-sectional data):

$$PRICE_i = f(SIZE_I) + \varepsilon_i = \beta_0 + \beta_1 SIZE_i + \varepsilon_i$$

where $PRICE_i$ = the price (in thousands of \$) of the i^{th} house; $SIZE_i$ = the size (in square feet) of that house, ε_i = the value of the stochastic error term for that house

This equation has practical value because it allows one to predict price based on size. Of course, the inclusion of additional X-variables may increase its precision.

LEARNING ACTIVITY 1

Consider equation 23 in the textbook:

$$PRIC\hat{E}_i = 40.0 + 0.138 SIZE_i$$

where $PRICE_i$ = the price (in thousands of \$) of the ith house; $SIZE_i$ = the size (in square feet) of that house.

What is the exact meaning of the slope coefficient (0.138)?

ANSWER

The 0.138 represents the value of $\frac{\Delta PRICE}{\Delta SIZE}$ *, the increase in PRICE per one unit increase in SIZE. In this case, if the size of the house increases by 1 square foot (say from 2 000 ft*^{2} *to 2 001 ft*^{2} *then the price of the house will increase by \$0.138 x 1 000 = \$138. (Multiplying by 1 000 because price is given in thousands of \$).*

The units of measurements of both PRICE and SIZE affect the value of the slope coefficient.

(C) TRUE/FALSE QUESTIONS **(F) = false (T) = true**

(1) Econometrics can resolve theoretical differences in opinion regarding the specification of an equation. (F)

Specification is not totally objective and leaves room for subjectivity, for example, in the way different schools of thought perceive things. Although specification is always based on economic theory, different theories are often based on different assumptions. Econometrics cannot resolve these differences. It can only estimate a given equation.

(2) The more Xs we use in a model, the better. (F)

It depends on the purpose of the model. Given two equally predictive theories (or models), the simplest (the one using less Xs) is usually the best one.

(3) If $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, then $\beta_0 + \beta_1 X_i + \varepsilon_i$ = $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$ (T)

(D) EXAMINATION: PARAGRAPH QUESTIONS

(1) Explain the goal of econometrics and its three main uses. (8)

(2) Explain whether the specification of a regression equation is mainly objective or subjective. Provide an example. (5)

(3) Explain the nature of a linear regression model. Use standard notation of the multivariate linear regression model to explain (10)

- the meaning of dependent and independent variables
- the meaning of the parameters (or coefficients) of an equation
- the difference between causality and correlation

- the difference between causality and correlation
- the meaning of the stochastic error term and its main sources of variation

(4) Concerning the estimation of a linear regression equation, explain (10)

- the meaning of linear in the variables and linear in the coefficients
- the meaning of a true (population) regression equation
- the meaning of an estimated (sample) regression equation
- the difference between β and $\hat{\beta}$
- the difference between the stochastic error term (ε) and the residual error term (e).

E EXAMINATION: PRACTICAL QUESTIONS

Make sure that you can interpret the coefficients of a regression equation. Note also that the meaning of the coefficients of a regression equation is affected by the units of measurement of the variables.

LEARNING UNIT 2

ORDINARY LEAST SQUARES (OLS)

ECONOMETRICS IN ACTION

The university of Manitoba's department of economics gives the following advice to its students of econometrics:

You can only learn statistical and econometric techniques by practicing them. These techniques are almost always embedded in some software, which looks after the calculations. Most of what you have to do as an econometrician is to learn

(a) how to ask your software for what you want, and then
(b) how to explain and present the output the software gives you.

In this course we use software (Microsoft Excel) to derive OLS estimates. Three important questions arise from our use of OLS:

- Do you understand what OLS does?
- Can you perform an OLS regression on a PC?
- Can you interpret its output?

These matters are addressed in this learning unit.

STUDY OBJECTIVES

The previous learning unit dealt with the purpose and method of econometrics. We learnt that econometrics makes use of three inputs: economic theory, statistics and economic data. The essence of econometrics is regression analysis which is estimating the coefficients of an equation from actual data. This learning unit introduces the most frequently used method, that is, the method of ordinary least squares (OLS).

When you have studied this learning unit you should

- be able to apply the method of OLS
- understand its method, its advantages and its output
- understand the quality of fit of a regression equation
- be able to perform OLS regressions on a PC

(A) PRESCRIBED MATERIAL

This learning unit introduces the method of ordinary least squares (OLS).

The following sections are prescribed:

(1) Estimating single-independent-variable models with OLS
(2) Estimating multivariate regression models with OLS
(3) Evaluating the quality of a regression equation

(4) Describing the overall fit of the estimated model
(5) An example of the misuse of $\overline{R}^2$
(6) Using a PC to perform OLS

Because OLS requires extensive calculations, we also explain in section 6 of the study guide how Microsoft Excel (a PC spreadsheet package) may be used to perform OLS. Of course, using a PC and Excel are not required in the examination. Their use, however, will help you understand econometrics better, help you to answer section B questions in the examination, and help you to complete Assignment 01. You may visit websites such as www.wikihow.com/Run-Regression-Analysis-in-Microsoft-Excel prior to your practical in order to gain more understanding of the process of running a regression using Excel spreadsheet.

(B) SOME IMPORTANT CONCEPTS

2.1 Estimating single-independent-variable models with OLS

The purpose of regression analysis is to derive the coefficients of an econometric equation. For the derivation of these estimates we use formulas presented in mathematical notation. We expect you to be familiar with mathematical notation. Note that a formula sheet will be provided in the examination see appendix 2 so do not spend time on memorising these formulae. You must however understand these formulae and know how to apply them.

Make sure you understand OLS, that is, in principle its method of derivation and its advantages. You must be able to apply the method of OLS in practice.

2.2 Estimating multivariate regression models with OLS

OLS can also deal with multivariate regressions. Multivariate regression involves analysing more than one independent variable and deriving coefficients for each one. These coefficients indicate the effect of each X-variable on the Y-variable.

You should understand the meaning of each of the components of the TSS (the explained sum of squares [ESS] and the residual sums of squares [RSS]). These concepts are also used later on when we come to R^2.

LEARNING ACTIVITY 1

Consider equation 11 in the textbook:

$$\hat{FINAID_i} = 8927 - 0.36PARENT_i + 87.4HSRANK_i$$

Where: $FINAID_i$ = the financial aid (measured in dollars of grant per year) awarded to the i^{th} applicant

$PARENT_i$ = the amount (in dollars per year) that the parents of the i^{th} student are judged able to contribute to college expenses

$HSRANK_i$ = the i^{th} student's GPA rank in high school, measured as a percentage (ranging from a low of 0 to a high of 100)

What is the exact meaning of these coefficients?

ANSWER

The -0.36 implies that the i^{th} student's financial aid grant will fall by $0.36 for every dollar increase in his or her parent's ability to pay, holding constant high school rank.

The 87.4 implies that the i^{th} student's financial aid grant will increase by $87.40 for every percentage point increase in the student's GPA rank, holding constant parents' ability to pay.

2.3 Evaluating the quality of a regression equation

Studenmund warns of the danger of blindly accepting computer output. A regression equation must be properly evaluated. It is important that you familiarise yourself with the list of questions that you should ask when evaluating the quality of a regression model, see the list provided on p. 50.

2.4 Describing the overall fit of the estimated model

Make sure you fully understand the meaning, strengths, weaknesses and misuses of R^2. We expect you to know the formulas of both measures of fit R^2 and $\overline{R}^2$ (they are not included in appendix 2 which will be supplied in the examination).

$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum \left(Y_i - \overline{Y}\right)^2}$	$\overline{R}^2 = 1 - \frac{RSS/(n-K-1)}{TSS/(n-1)}$

When using RSS, ESS and TSS you must also be able to explain their meaning.

LEARNING ACTIVITY 2

Why is $\overline{R}^2$ a better measure of fit than R^2?

ANSWER

It is insufficient to say that $\overline{R}^2$ is better because it includes the degrees of freedom. You must be able to explain exactly what happens to both measures when an additional X-variable is included in the regression equation. These can best be understood with reference to the sums of squares: residual sum of squares (RSS) and total sum of squares (TSS).

Measure	*What happens when an additional X-variable is included in the regression equation?*
R^2	- *TSS remains constant (you can see from the formula that TSS is not affected by X) while RSS never increases (RSS is likely to decrease)* - *This means that R^2 never decreases which may create the illusion of a better fit when an additional X-variable is included.*
$\overline{R}^2$	- *If the equation truly fits better, then the RSS decreases, which increases $\overline{R}^2$.* - *K increases so that n-K-1 decreases and RSS/ (n-K-1) increases (assuming a constant RSS). This means that $\overline{R}^2$ decreases* - *$\overline{R}^2$ is a better measure of fit when an additional X-variable is added because $\overline{R}^2$ may increase or decrease depending which of the two effects is dominant.*

The simple correlation coefficient (r)

Make sure you understand the differences between R^2 and r. R^2 measures the quality of fit of a regression equation, for the cases of one and more than one independent variables. The simple correlation coefficient (r) measures the strength and direction of a linear relationship between two variables. See section 2.4, p. 52.

Formulas

Apart from basic formulas, you need not memorise (complex) formulas. The formulae given in appendix 2 will be provided in the examination. You should be able to apply relevant formulas, which means that you must understand them.

Summation notation

Summation notation is a flexible and shorthand way of writing complex formulas. The summation function is explained in chapter 12, section 1.

LEARNING ACTIVITY 3

Explain exactly what is meant by $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$.

ANSWER

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + X_3 + ... + X_n}{n}$$

$\bar{X}$ is the unweighted average of a set of n numbers X_1 to X_n. The summation sign is often used without its i = 1 to n indicators, which is implied however.

LEARNING ACTIVITY 4

Assume the following data:

Obs	**X**	**Y**
1	3	22
2	5	17
3	7	14
4	9	13
5	11	9
SUM	35	75
Average	7	15

(a) Derive the following sums of squares from basic principles: xy; x^2; y^2 where x = X – $\bar{X}$ and y = Y – $\bar{Y}$.

(b) Estimate the coefficients of Y = $\beta_0 + \beta_1 X + \varepsilon$.

(c) Derive $\sum e_i^2$, TSS, RSS, ESS and R^2.

ANSWERS

The following table provides the required sums:

No	X_i	Y_i	x_i	y_i	$x_i y_i$	x_i^2	y_i^2	$\hat{Y}_i$	e_i	e_i^2	$(\hat{Y}_i - \bar{Y})^2$
1	*3*	*22*	*-4*	*7*	*-28*	*16*	*49*	*21*	*1*	*1*	*36*
2	*5*	*17*	*-2*	*2*	*-4*	*4*	*4*	*18*	*-1*	*1*	*9*
3	*7*	*14*	*0*	*-1*	*0*	*0*	*1*	*15*	*-1*	*1*	*0*
4	*9*	*13*	*2*	*-2*	*-4*	*4*	*4*	*12*	*1*	*1*	*9*
5	*11*	*9*	*4*	*-6*	*-24*	*16*	*36*	*9*	*0*	*0*	*36*
Sum	*35*	*75*	*0*	*0*	*-60*	*40*	*94*		*0*	*4*	*90*

Where $x_i = X_i - \bar{X}$; $y_i = Y_i - \bar{Y}$; $\bar{X} = 35/5 = 7$; $\bar{Y} = 75/5 = 15$

(a) *$\Sigma xy = -60$; $\Sigma x^2 = 40$; $\Sigma y^2 = 94$*

(b) $\hat{\beta}_1 = \sum x_i y_i \big/ \sum x_i^{\ 2} = -60/40 = -1.5$

$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \vec{X} = 15 + 1.5(7) = 25.5$

(c) *To derive $\sum e_i^2$, first derive the e_i. For example,*

$$e_1 = Y_1 - \hat{Y}_1 = Y_1 - \left(\hat{\beta}_0 + \hat{\beta}_1 (X_1)\right) = 22 - \left(25.5 - 1.5(3)\right) = 1$$

The remaining e_s are provided in the table.

The $\sum e_i^2 = 4$.

TSS = $\sum y_i^2 = 94$; RSS = $\sum e_i^2 = 4$; ESS = TSS – RSS = 90 – 4 = 90
or ESS = $\sum \left(\hat{Y}_i - \overline{Y}\right)^2 = 90$

R^2 = ESS/TSS = 90/94 = 0.9574.

2.5 An example of the misuse of $\overline{R}^2$

Equation 2.16 has the higher $\overline{R}^2$ compared to equation 2.17 and thus appears the preferred one. Equation 2.16 however, has a serious shortcoming. Can you identify it? The problem with equation 2.16 is that the coefficient of the price of water variable has an unexpected sign which is likely to cause this equation to forecast badly.

The important lesson is that $\overline{R}^2$ is never the only criteria to judge the quality of a regression equation. In chapter 6, section 6.2 we will learn that there are four major specification criteria which apply, that is, theory, t-test, $\overline{R}^2$ and bias.

2.6 Using a PC to perform OLS

2.6.1 Using Microsoft Excel to perform OLS

Because OLS requires extensive calculations we can use a PC to help us. For the case of estimating the coefficients of Y = β_0 + β_1X + ε, formulas 4 and 5 (see prescribed book section 1, p. 39) are the estimators of its two coefficients. Table 1 in p. 41 of the textbook summarises the calculations required. Even this simple form requires extensive calculations. In the case of estimating multivariate regression equations the formulas become increasingly more complex (see for example, the formulas provided in footnote 4 on page 44).

Of course we can perform the regression analysis by using Microsoft Excel. We will use both Excel version 2003 (also called version 11 – which displays File, Edit, View, … in the main menu bar), and Microsoft version 2007 (also called version 12 which uses the ribbon interface – which may be identified by the Office button in the top-left position of the screen and the ribbon – a broader bar below it which provides easy access to each of the main functions of Excel (Home, Insert, Page Layout, …).

Excel offers two different methods to perform regression analysis; these are the Add-In method and the Linest method.

(a) The Add-In method

Excel requires the appropriate Add-in to be loaded first. An Add-in is a program file which performs, in this case, the OLS calculations automatically. To determine whether the required Add-In has already been loaded, click the following selections from the Excel main menu.

Excel version 2003

Tools; Add-ins.

This opens a list of available Add-ins. Check whether either the Analysis Toolpak (old version) or the Analysis Toolpak-VBA (new version) Add-in is loaded. If not, check its box and click "OK". It is best to select the new version. In cases where the Add-in has not been previously installed, Excel will ask for the installation disk, from which it will then install the Add-in automatically.

To run the Add-In, select from the main menu **Tools**; **Data Analysis**.

This opens a list of Analysis tools. Select "Regression" from this list and click "OK". An input box is then displayed, which allows the user to, amongst others; input the Y-range, X-range and Output-range as explained in section “Specification of input" below.

Excel version 2007 and 2010

To determine whether the required Add-In has already been loaded, click **Data**, which will return "Data Analysis" on the ribbon below if it has already been loaded.

To run the Add-In click on **Data Analysis** which opens a box showing a list of analysis tools. Select "Regression" from this list and click "OK". An input box is then displayed, which allows the user to, among others, input the Y-range, X-range as explained in section “Specification of input" below.

If the "Data Analysis" Add-In is not present, then it must be loaded first. To load it, click the Office button, then select

Excel options; Add-Ins

which opens a list of Add-Ins, both active and non-active (available but not loaded). Select

Manage Excel Add-Ins and click "Go".

This opens a box showing a list of Add-Ins. From this list select "Analysis-Toolpak-VBA" by clicking on the box at its left and then "OK". The Add-In will now be available under "Data".

Specification of input

You are required, at the minimum, to input the Y-range and X-range, By default the output is written to a new worksheet. Assume you wish to run the regression:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_{12} + \varepsilon$$

and that the Y-values are located in range B3:B8 (a single column of data) of the spreadsheet and the X-values in range D3:E8 as indicated below.

TABLE 2.5: Spreadsheet data for regression analysis

Columns

Rows

	B	**C**	**D**	**E**
3	42		12	4
4	56		14	7
5	78		16	9
6	87		18	5
7	107		20	3
8	120		22	1

The X-range could be either a single column of X-values (for example E3:E8), or multiple columns of Xs in the case of a multivariate model. The X-range may not contain empty columns as Excel assumes that each column reflects an X-variable.

The regression input-box is completed as follows:

Input X-range: D3:E8
Input Y-range: B3:B8

The residuals box may also be checked (to display the residuals) but this is optional.

If all is fine the Add-in will return the following output:

	B	C	D	E	F
11	SUMMARY OUTPUT				
12					
13	*Regression statistics*				
14	Multiple R	0.997138			
15	R Square	0.994285			
16	Adjusted R Square	0.990476			

	B	C	D	E	F
17	Standard Error	2.890177			
18	Observations	6			
19					
20	ANOVA				
21		*Df*	*SS*	*MS*	*F*
22	Regression (ESS)	2	4360.2739	2180.1369	260.9965
23	Residual (RSS)	3	25.0593	8.3531	
24	Total (TSS)	5	4385.3333		
25					
26		*Coefficients*	*Standard Error*	*t Stat*	*P-value*
27	Intercept	-58.832044	9.1089	-6.4586	0.0075
28	X Variable 1	8.116479	0.4239	19.1428	0.0003
29	X Variable 2	0.521082	0.5551	0.9386	0.4171

The beauty is that the most important regression results are automatically returned. Let us briefly explore some of these.

- R^2 is returned in cell C15 and $\overline{R}^2$ in cell C16.
- The sums of squares are returned in D22:D24. Of course, R2 = ESS/TSS = 4360.27/ 4385.33 = 0.9943.
- The estimated coefficients are returned in cells C27:C29. Therefore, the estimated regression equation is:

$$\widehat{Y} = -58.832044 + 8.116479X_1 + 0.521082X_2$$

- The standard errors and t-values of each coefficient are also returned, as well as the F-value. We will learn more about these important statistics later.

If you checked the residuals box then the residual terms are automatically returned. Those of the first two observations are indicated below:

		C	D
35	***Observation***	***Predicted Y***	***Residuals***
36	1	40.65004	1.34996
37	2	58.44624	-2.44624

The predicted Y_1 (observation 1) is derived as:

$$\widehat{Y}_1 = -58.832044 + 8.116479(12) + 0.521082(4) = 40.65.$$

The first residual term (e_1) is, if course derived as: $e_1 = Y_1 - \widehat{Y}_1 = 42 - 40.65 = 1.35$

(b) Microsoft Excel: The LINEST method

If you cannot get Excel's Add-in loaded, then there is still the option of using the Excel LINEST function. LINEST can estimate a multiple regression equation and it works the same in MS Excel versions 2003 and 2007.

- The syntax is: LINEST(known_y's,known_x's,const,stats). For example LINEST(B3:B8, D3:E8,1,1) would perform a linear regression to the data in Table 2.5.

Columns	**A**	**B**	**C**		
Rows	X_2	X_1	**constant term**	**Meaning of output**	
25	0.521082	8.116479	-58.832	Coefficients	
25	0.555139	0.423994	9.108985	Standard errors	
27	0.994286	2.890178	#N/A	R^2	se_y
28	260.9965	3	#N/A	F-value	Degrees of freedom
29	4360.274	25.05938	#N/A	ss_{reg}	ss_{resid}

- When you perform the LINEST function, it will initially return a single cell answer (0.521082 in cell A25). The other cells are hidden!
- To unhide them, take the following steps:

(1) Select cell A25.

(2) Starting with cell A25, select a range 3 columns wide x 5 rows down:

- Hold down the Shift-key, then press the Right-key (→) 2x, and the Down-key (↓) 4x.
- This selects the range A25:C29

(3) Press F2 and then

(4) Press Ctrl+Shift+Enter (press and hold Ctrl, then press and hold Shift, then press Enter while Ctrl and Shift are still pressed down. Then release all keys. (Hold means keeping the key pressed down).

- This action should now display all the cell values as shown in the shaded area above.
- LINEST offers fewer options than the Add-in. LINEST, however, offers one important advantage. When the data changes, then LINEST recalculates its output automatically. The Add-in does not automatically recalculate its output when the data changes. It requires a rerun of the Add-In.

(c) How to perform a regression analysis using Quattro Pro or Lotus 1-2-3 for Windows spreadsheets

We recommend that you use the spreadsheet of your choice to perform regression analysis. All spreadsheets have a regression module, which allows you to estimate multivariate regression models.

Specification of input

In each of the packages you are required, amongst others, to input the Y-range, X-range and Output-range.

The **Y-range** is the single column of data of the dependent variable (for example A1:A6). The **X-range** could be either a single column of X-values (for example C1:C6), or multiple columns of Xs in the case of a multivariate model (for example C1:D6).

The **output range** is a top left-hand location (single cell) specifying where the regression results will be written (for example A8) in the case where the results are required to appear on the same sheet below the data.

Quattro Pro (version 8)

From the main menu select

Tools; Numeric tools; Regression

which provides input boxes for the independent, dependent and output cells.

Alternatively, one can also use

Tools; Numeric tools; Analysis; Advanced regression

which, amongst others, provides input boxes for the Y-range; X-range, and Summary output.

Lotus 1-2-3 for Windows (version 1.1)

From the main menu, select

Data > Regression.

An input box opens, which provides for the user's input of the X-range, the Y-range and the Output-range.

LEARNING ACTIVITY 5

Do exercise 2 on p. 60 in the prescribed book.

The answer to exercise 2 is given on p. 69.

C TRUE/FALSE QUESTIONS (F) = false (T) = true

(1) OLS is the only technique available to estimate a regression equation. (F)

Although OLS is the most frequently used regression technique, there are a number of others as well. Some are adapted from OLS, for example, weighted least squares (WLS) and generalised least squares (GLS) which will be dealt with later in the course. Then there are also others such as maximum likelihood estimators (not dealt with in this course).

(2) If $R^2 = 1$ then all the points lie exactly on the regression line. (T)

(3) ESS = $\sum e_i^{2}$ (F)

ESS represents the <u>explained</u> sum of squares and not the residual sum of squares (RSS). (F)

(4) In principle, if we have two (X, Y) points, for example (2, 5) and (6, 7), then we can estimate both the slope and constant term of the linear regression equation: Y = a + bX. (T)

The slope can be derived as ΔY/ΔX

$$\hat{b} = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{7-5}{6-2} = \frac{2}{4} = 0.5$$

The intercept can be derived by substitution. Since $Y_1 = a + 0.5(X_1)$ then

$$\hat{a} = Y_1 - 0.5(X_1) = 5 - 1 = 4$$

Thus, Y = 4 + 0.5(X).

(5) A multivariate regression model is seldom used because of its computational complexity. (F)

(6) A problem in section 2.5 of the textbook is that the slope coefficients of both the estimated equations (2.16 and 2.17) have an incorrect sign. (F)

Only the PR variable in equation 2.16 has an incorrect sign. According to theory, price is inversely related to quantity demanded. Thus, equation 2.17 which has correct signs of all its coefficients is better than equation 2.16 in spite of its lower R^2.

(7) $\bar{R}^2$ is useful, since it allows one to judge whether it is worth adding another X to the regression equation. (T)

(8) The simple correlation coefficient can vary between -1 and +1. (T)

(9) You are required to use a PC in the examination. (F)

(D) EXAMINATION: PARAGRAPH QUESTIONS

(1) Explain (6)

- the meaning and purpose of an OLS estimator
- the principle of the OLS estimation technique
- the advantage of the OLS estimation technique.

(2) Explain the meaning of degrees of freedom. (1)

(3) Explain (6)

- the meaning of the total, explained and residual sum of squares
- their relation to R^2 and $\bar{R}^2$.

(4) Explain (8)

- the meaning of *goodness of fit*
- the difference between R^2 and $\bar{R}^2$, which one is preferred and why
- the potential misuses of $\bar{R}^2$

(E) EXAMINATION: PRACTICAL QUESTIONS

The OLS estimators are provided in the formula sheet (appendix 1) and need not be memorised. You must, however, be able to interpret and apply the formulas correctly.

Practical regression results may be given to you in the examination and you may be asked to interpret these. This means being able to interpret the measures of fit correctly, being aware of the misuse of measures of fit in practical situations, and being able to apply formulas.

LEARNING UNIT 3

LEARNING TO USE REGRESSION ANALYSIS

ECONOMETRICS IN ACTION

Professor Kazarosian of Boston College (USA) provides a possible outline of a compulsory research paper which forms part of their introductory econometrics course.

i. *Introduction:* what are the hypotheses being tested? Why are they interesting?
ii. *Brief literature review:* what other work has been done on these issues? What has been found? How does your paper differ from the others?
iii. *Data:* what data would you like to have? What data have you been able to obtain? What special data problems did you encounter?
iv. *Empirical work:* regression analysis & interpretation of the results.
v. *Conclusions and summary:* what have you learned? What are the policy implications? Are there suggestions for future research?

Do you think the compilation of such a research paper could be useful? Can you think of possible topics for such a research paper? This learning unit explores the issues addressed in the outline above.

STUDY OBJECTIVES

By now things start falling into place. You understand that the essence of econometrics is regression analysis. From this perspective the whole process appears relatively simple. You start with a problem requiring regression analysis. You assume a regression equation, find data for it and then run OLS on your PC. It appears to be quick and easy. Or is there more to it?

There is. This chapter explains how regression analysis may be applied in practice. When you have studied this learning unit you should

- understand the practical application of regression analysis by means of the standard six steps
- understand how these principles have been applied in the "Pick restaurant locations" practical example
- know where to look for data

(A) PRESCRIBED MATERIAL

This learning unit summarises the basic practice of regression analysis.

(1) Steps in applied regression analysis.
(2) Using regression analysis to pick restaurant locations is not prescribed but is useful, as it shows how the steps can be applied.
(3) Data: This section provides some additional notes on data and its sources. This is discussed in *section 2, p. 359.*

(B) SOME IMPORTANT CONCEPTS

3.1 Steps in applied regression analysis

These are:

- review the literature
- specify the model
- state the expected signs of the coefficients
- collect the data
- estimate the equation(s) and evaluate the results
- complete the documentation

Make sure you understand the purpose of each of these steps.

3.2 Using regression analysis to pick restaurant locations

We suggest you work through this section in order to better understand the six steps of applied regression analysis. Read through this section and then complete the activities below.

LEARNING ACTIVITY 1

What is the purpose of the regression equation?

The purpose of the equation is to predict gross sales volume of a restaurant. Of course, the idea is that once the equation has been estimated, it may be used to predict the sales volume of a restaurant based on its location. This enables one to find the optimum restaurant location.

LEARNING ACTIVITY 2

What phenomenon does the dependent variable (Y) of the equation measure? Do the data (see Table 3.1) appear to be realistic?

The Y-variable measures gross sales volume. Its precise meaning is not immediately evident. However, if you read Studenmund carefully then the answer becomes clear. Gross sales volume is defined as the number of clients served by a restaurant within a year. It is measured by the number of bills issued to customers. You should realise that this is not an ideal measure. It only indirectly reflects sales value. The amount (value) per bill may vary *greatly since it depends on factors such as the number of people within a group and how much they spend. In practice we often have to settle for a variable which is measurable and available.*

It's always a good idea to inspect the data. Table 3.1 shows that Y varies between say 91 000 and 166 000. The average is about 128 000 bills per year. If we assume restaurants are open, say 350 days per year then this is about 370 bills per day, or 30 bills per hour assuming a 12 hour day. This appears reasonable.

LEARNING ACTIVITY 3

Why is the price of food not included as an X-variable? And what about the quality of service and the quality of food?

This is a family type of restaurant which offers a fast-food service. It is part of a chain and the quality and prices charged do not vary much between these types of restaurants.

LEARNING ACTIVITY 4

Which variables are ultimately included in the model?

There are three: competition (the number of competitors within a two-mile radius), population (the number of people living within a three-mile radius) and income (the average household income of the population). You should realise that not all these variables may be accurately measurable.

LEARNING ACTIVITY 5

What can be said about the quality of the estimated equation?

The coefficients all have the correct signs. The $\overline{R}^2$*-value, however, is relatively low. There are other explanatory variables not included in the model. This implies that the accuracy of its predictions may not be very good.*

LEARNING ACTIVITY 6

Interpret the meaning of each of the slope coefficients of the estimated equation in exact terms (taking account of the units of measurement).

$$\widehat{Y}_i = 102.192 - 9075N_i + 0.355P_i + 1.288I_i$$

Coefficient	*Impact on the number of bills issued by the restaurant in a year*
ΔY/ΔN = -9075	*For each additional competitor restaurant within a two-mile radius, the number of bills will decrease by 9075.*
ΔY/ΔP = +0.355	*If the number of people living within a three-mile radius of the restaurant increases by one, then the number of bills will increase by 0.355.*
ΔY/ΔI = +1.288	*If the average household income of the population increases by $1, then the number of bills will increase by 1.288*

3.3 Data

Proxy variables

A proxy variable is used as a substitute variable for a variable for which data are not available. See section 11.2

Time-series versus cross-sectional data

Observations of variables (e.g. X_j) are usually indexed (j) according to time, or according to observation number if time is not applicable.

Most of the data used in econometrics are time-series data, where the same variable is observed in different periods. Example: GDP for the years 1960–1998. The index of the series represents time, for example GDP_{1994}**.**

Cross-sectional data are data for which the index represents different units for the same time period. The population census of 1997 represents cross-sectional data, for example $POP_{Gauteng.}$

Sources of data

The purpose of this section is to provide some practical information with regard to the sources of data which may be useful in a South African context.

Data is not free and obvious; it has to be actively sought and selected. Some data may have to be purchased and sometimes data must be collected by survey. Nor are data simple, as definitions, the scope, degree of measurement error, primary source, unit of measurement, method of calculation, timeliness, stage of revision, type of adjustment made etc. may vary.

In South Africa the two main sources of economic data are

- the *Quarterly Bulletin* of the South African Reserve Bank (SARB) and
- Statistics South Africa (StatsSA)

The statistical tables of the *Quarterly Bulletin* of the South African Reserve Bank contain extensive data on money and banking, the capital market, public finance, international economic relations, national accounts and economic indicators.

The SARB web site http://www.resbank.co.za provides a wealth of information, and access to the electronic versions of their *Quarterly Bulletins* dating back to 1996 (as on May 2015).

Besides time-series data, access to previous research on the area of interest can provide very useful starting points. The SARB maintains a macroeconomic model of the South African economy and from time to time publishes details of parts of its structure. The June 1995 edition, for example, carries an excellent article on the determinants of private consumption expenditure. There are others on the demand for money et cetera.

Stats SA regularly publishes its data by means of

- the Statistical Yearbook (annually)
- the Bulletin of Statistics (quarterly)
- various publications (monthly leaflets)

Their web address is http://www.statssa.gov.za

Stats SA publishes detailed cross-sectional data on households. They provide such a wealth of detail that most econometricians are almost tempted to start a research project immediately!

A number of **government departments** compile statistics within their fields, for example for agriculture and mining. **Private institutions** like the JSE (Johannesburg Securities Exchange), NAAMSA (National Association of Automobile Manufacturers of SA) and SEIFSA (Steel and Engineering Industries Federation of SA) provide data within their respective fields.

International data are released by the International Monetary Fund (IMF), World Bank, etc.

The web is another rich source of both economic data and articles. Many economic journals can be accessed online.

Data are also provided by a number of commercial firms active amongst others, in the data-distribution business. For example, IHS Global Insight Southern Africa (Pty) Ltd and Quantec offer clients a data service. See, for example, http://www.globalinsight.co.za, and http://www.quantec.co.za.

(C) TRUE/FALSE QUESTIONS (F) = false (T) = true

(1) The number of degrees of freedom is indicated by n-K-1, where n: number of observations and K: number of independent variables excluding the constant term. (T)

(2) When creating a new model, one should first run some preliminary regressions on the data in order to get a proper feel for the data, before one hypothesises the model. (F)

This practice is called data-mining and is not recommended. Prior expectations may be imposed on the data, and the statistical significance of the estimated coefficients is overstated.

(D) PRACTICAL EXERCISES

Attempt problem 8 (GRE, pp. 87–88) and problem 9 (Gasoline mileage, pp. 88–89). Try to answer these questions to sharpen your practical skills.

(E) EXAMINATION: PRACTICAL QUESTIONS

You should be able to

- explain and execute the steps in an applied regression analysis
- interpret the meaning of regression coefficients

ECONOMICS IN ACTION (Looking back)

Professor Kazarosian (referred to at the beginning of this learning unit) offers advice to students regarding possible topics for a research paper. Let's read his advice and see what a wide range of topics econometrics can address.

- Pick an area where data are readily available, and avoid topics requiring new surveys or involving simultaneous equations (for example supply and demand).
- Experience has shown that cross-sectional studies work better than time-series for this assignment, although the latter has been done.
- The Census volumes provide excellent and abundant data on US states and cities.
- Topics from previous years include: State by state variations in divorce rates; Voter behavior; Crime rates in US cities; Wage earnings or poverty differentials by state; The determinants of teenage pregnancies by state; Catholic Church scandal influence on donations; State by state alcohol consumption; Baseball attendance across cities; Salary determination in major league sports; Determination of MCAS results; Determinants of US strike learning activity; Determination of welfare participation; State by state variations in suicide rates; by state variations in traffic fatalities; by state variations in fertility rates; Voter turnout in Presidential election years.

Can you think of typical South African issues which may be added to this list? Do you see how useful econometrics can be?

PART II

STATISTICS

At this point you already have a good idea of the purpose and method of OLS. You know how to derive OLS estimates, given the form of the equation and given suitable data. The question now arises: How good is OLS? How trustworthy are its estimates?

OLS is a statistical estimator which is usually applied to sample data where the sample is a subset of the population. These sample estimates are not necessarily accurate. There is the risk that the sample estimates may be in error. Fortunately, there are limits to the extent of these errors, and their probabilities can be derived based on the laws of chance.

To understand this well, we must use the concepts and measures of statistics. So fasten your seatbelts for some necessary exposure to the language, the concepts and the method of statistics. The journey is, in fact, a short course in statistics!

Part II consists of three learning units

- Learning unit 12: Statistical principles
- Learning unit 4: The classical model
- Learning unit 5: Basic statistics and hypotheses testing

Learning unit 12 is not out of place. It is deliberately included at this point. Learning unit 12 deals with the underlying statistical concepts which are referred to in learning units 4 and 5.

LEARNING UNIT 12

STATISTICAL PRINCIPLES

ECONOMETRICS IN ACTION

The Department of Economics at New York University (NYU) has evolved into one of the world's leading centres for research and teaching in economics. Professor C Flinn of NYU teaches the Econometrics I course. Here are some of his comments on his course objectives:

- We will begin by reviewing probability and sampling theory. To be a competent econometrician, one needs to have a solid understanding of basic statistical theory, some familiarity with data, and a good knowledge of economic theory.
- From my perspective, econometrics is essentially the application of standard statistical tools to the analysis of conditional relationships between random variables. What distinguishes econometrics from statistics is the econometrician’s objective to infer something about behaviour from empirical relationships between variables.
- In this course, we will attempt to prepare the student for this kind of research enterprise by carefully covering most or all of the statistical theory [albeit at a basic level] they will need to do competent applied econometric analysis.

The message is clear. One cannot fully understand econometrics without a solid grounding in statistics.

STUDY OBJECTIVES

Econometrics makes extensive use of statistical concepts. Some examples:

- We assume that the data used in regression analysis is a random sample drawn from the population. What exactly is the meaning of "random sample" and of "population"?
- What are the implications of using sample estimates? The concept of the **sampling distribution** of a sample estimator is a fundamental concept you must understand well.
- Related to the sampling distribution are concepts like unbiased estimators and minimum variance. What do these mean?

This module requires you to be familiar with statistical concepts. This chapter deals with the basic statistical concepts required in this regard. This could be particularly helpful to students who have not previously completed statistics courses. Students who have previously completed statistics courses may find this chapter a convenient means to brush up their statistics, and may even learn some new things.

Yes, this learning unit is examination material. Within each of the sections below, we clearly indicate what you must understand.

The approach of this chapter is different to that of other chapters.

- We first provide the headers of sections as discussed in the textbook. We then tell you exactly what you are required to know. Remember, the focus is on understanding the meaning of statistical concepts. There may be examination questions on the material you are required to know. We may, for example, ask you to derive a standard deviation (given some simple data), to explain its meaning, you may be asked to explain what is a sample distribution, or you may be required to explain the meaning of expected value. The major part of this learning unit consists of a number of activities which are practical applications of all the major statistical concepts. The activities are meant to be learning exercises. They may assist you in better understanding statistical concepts. Definitely work through them!
- Although some aspects may be explained in a different way than the textbook, the textbook remains your prime source.

12.1 Probability distributions

This section covers topics on probability, mean, variance and standard deviation, continuous random variables, standardised variables and the normal distribution.

We expect you, in the case of discrete random variables, to understand the meaning of

- a random variable (X) and the probability distribution of X which is denoted by its probability density function P(X)
- the mean (or expected value) of random variable (X)
- the variance and the standard deviation of random variable (X).

In the case of continuous random variables, you must

- why continuous variables arise
- the meaning of the probability distribution (the probability density curve)
- the meaning of mean, variance and standard deviation
- the meaning of standardised variables.

In the case of the normal distribution you must

- understand its meaning and how the central-limit theorem can give rise to a normal distribution
- be able to apply the normal distribution in practice.

Activities 1 to 5 deal with the following major statistical concepts:

- Probability density function (uniform) of a discrete random variable and its expected value
- Mean, variance and standard deviation of a discrete random variable which is not uniformly distributed
- Continuous random variables and their probability density functions; Standardised random variables
- Expected value and bias
- The normal distribution
- The central limit theorem

LEARNING ACTIVITY 1

Consider a normal die with numbers 1 to 6 on its sides. Let X measure the outcome of a throw of the die.

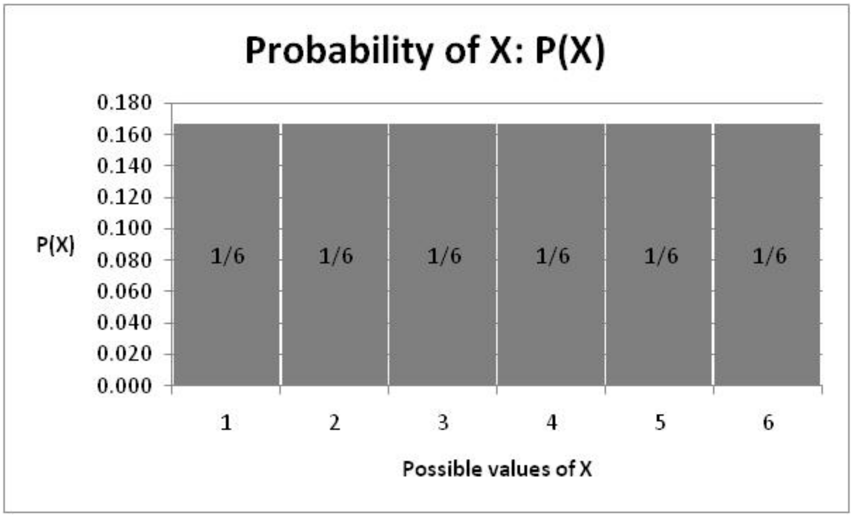
(a) Explain how the concept of a discrete random variable (X) may be applied to the throw of the die.
(b) Derive the probability density function P(X). Explain whether P(X) is normally distributed.
(c) Derive E(X), the expected value of X and explain its practical meaning.
(d) Derive the variance of X and the standard error of X

ANSWERS

(a) The variable (X) can assume 6 possible outcomes when the die is thrown. The range of possible outcomes of X is (1, 2, 3, 4, 5, 6). Because these are a countable number of possible values, X is a discrete variable. Because X assumes values by random chance, X is also a random variable. Thus X is a discrete random variable.

(b) The probability P(X) is the probability of obtaining each of these X-values. Because each number 1 to 6 has an equal chance of occurring, P(X) = 1/6 for all X. Note that ΣP(X) = 1.

Variable X is not normally but uniformly distributed. In the case of the uniform distribution, P(X) is constant for all values of X. In the case of the normal distribution, the chart of P(X) versus X is bell shaped. Loosely speaking, this means that the probability P(X) of realising numbers in the middle range of X is higher than that of the tail ends.



(b) The expected value of X is derived as ∑ X . P(X):

= 1.(1/6) + 2.(1/6) + 3.(1/6) + 4.(1/6) + 5.(1/6) + 6.(1/6)
= (1 + 2 + 3 + 4 + 5 + 6).(1/6)
= 21/6
= 3.5

The meaning of the expected value is the average value of a large number of throws. Because each throw can yield numbers 1 to 6, where the probability of each number is 1/6, we can expect the average of a large number of throws to be 3.5.

(d) The variance of X is $\sum (X - \mu)^2.P(X)$ *where μ is the expected value of X (μ = 3.5).*

X	$P(X)$	$X - \mu$	$(X - \mu)^2$	$(X - \mu)^2.P(X)$
1	*1/6*	*-2.5*	*6.25*	*1.0417*
2	*1/6*	*-1.5*	*2.25*	*0.3750*
3	*1/6*	*-0.5*	*0.25*	*0.0417*
4	*1/6*	*0.5*	*0.25*	*0.0417*
5	*1/6*	*1.5*	*2.25*	*0.3750*
6	*1/6*	*2.5*	*6.25*	*1.0417*
Sum				*2.9167*

The variance of X is 2.9167. The standard error of X is $\sqrt{2.91667}$ *= 1.7078.*

LEARNING ACTIVITY 2

This example deals with a non-uniform probability density function in contrast to example 12.1.1 which deals with a uniform one.

Consider a normal die with numbers 1 to 6 on its sides. Let Y measure the sum of two throws of the die. For example, if two throws realise a 4 and a 2, then Y = 6.

The outcomes of all possible throw 1 and throw 2 values are displayed in the table on the right.

The possible outcomes of Y range from a minimum of Y = 2 (1 + 1) to a maximum of Y = 12 (6 + 6). Each of the 6 x 6 possible outcomes has an equal probability to occur, that is, 1/36. Note that there are more '7' outcomes for Y than for example 5's simply because more combinations of throws have the sum of 7.

Y = $throw_1$ *+* $throw_2$		**Outcome of throw 2**					
		1	**2**	**3**	**4**	**5**	**6**
Outcome of throw 1	**1**	*2*	*3*	*4*	*5*	*6*	*7*
	2	*3*	*4*	*5*	*6*	*7*	*8*
	3	*4*	*5*	*6*	*7*	*8*	*9*
	4	*5*	*6*	*7*	*8*	*9*	*10*
	5	*6*	*7*	*8*	*9*	*10*	*11*
	6	*7*	*8*	*9*	*10*	*11*	*12*

(a) List all possible values of Y as well as their frequency (how many times each occurs). Which value of Y occurs most?

(b) Determine and draw P(Y), the probability density function of Y. Is Y normally distributed?

(c) Derive μ the mean value of Y, σ^2 the variance of Y and σ the standard deviation of Y. Explain the meaning of σ.

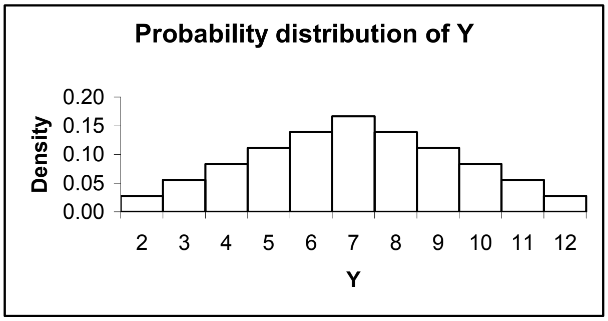
ANSWERS

(a) See the table below for the 11 possible Y_i values which fall between 2 and 12. The Y-value of 7 occurs most (it is called the mode).

Y	*Frequency (F)*	*P(Y) (F/36)*	*Y.P(Y)*	*Y – μ*	*(Y – μ)²*	*(Y – μ)². P(Y)*
2	1	0.0278	0.0556	-5	25	0.6944
3	2	0.0556	0.1667	-4	16	0.8889
4	3	0.0833	0.3333	-3	9	0.7500
5	4	0.1111	0.5556	-2	4	0.4444
6	5	0.1389	0.8333	-1	1	0.1389
7	6	0.1667	1.1667	0	0	0.0000
8	5	0.1389	1.1111	1	1	0.1389
9	4	0.1111	1.0000	2	4	0.4444
10	3	0.0833	0.8333	3	9	0.7500
11	2	0.0556	0.6111	4	16	0.8889
12	1	0.0278	0.3333	5	25	0.6944
Sum	36	1.0000	μ = 7.0000	0	110	5.8333

Frequency: the number of times the specific Y-value occurs.

(b) P(Y) is proportional to the frequency of Y and P(Y) = $Y_{frequency}$/36. ΣP(Y) = 1, that is, the area under the P(Y) curve is 1, which of course also applies to the continuous variable case. The nice thing about P(Y)'s is that if you want to derive the probability of getting numbers say 5 to 9, you simply add their P(Y)'s which is (4 + 5 + 6 + 5 + 4)/36 = 24/36. The probability density function is displayed below.



Is Y normally distributed? Well, not quite! To be normally distributed, Y must be a continuous variable and its probability distribution P(Y) must be bell shaped.

Y is a discrete variable and its probability distribution P(Y) is not bell shaped.

(c) $\mu = \Sigma Y.P(Y) = 7$

$\sigma^2 = \Sigma(Y_i\text{-}\mu)^2.P(Y_i) = 5.8333$

$\sigma = \sqrt{\sigma^2} = 2.4152$. σ is a measure of the dispersion or variation of Y. In the case of a normal distribution, about 2/3 of its values fall within the range $\mu - \sigma$ to $\mu + \sigma$. Applied to the case of Y, $\mu - \sigma = 4.6$ and $\mu + \sigma = 9.4$. Well, Y is a discrete variable and not normally distributed either but let's approximate it by the range 5–9. The probability of Y falling within this range is indeed 2/3 (Sum its P(Y) values: 1/36(4 + 5 + 6 + 5 + 4) = 24/36 = 2/3).

LEARNING ACTIVITY 3

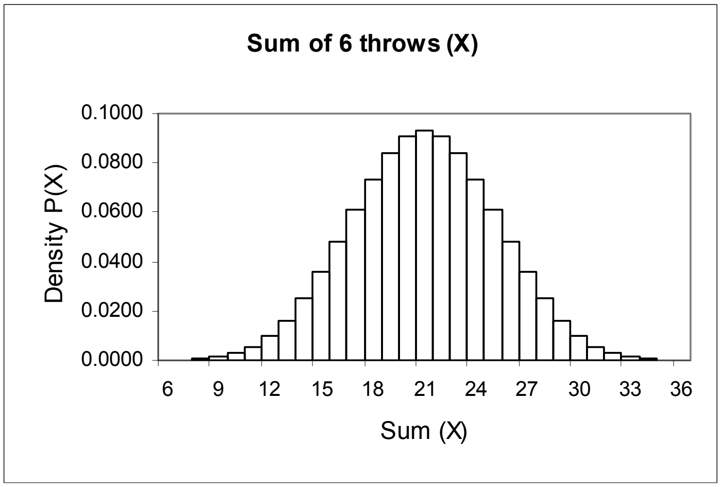
Explain why there is a need for continuous variables. How do we interpret P(X) for continuous variables? When is a continuous variable normally distributed? What is a standardised variable?

ANSWER

In real life the outcomes of random variables are often not countable numbers. Often the values of random variables are rational numbers which may include decimal fractions. For example, a continuous random variable u may assume the value of -4.7636 (rounded to 4 decimals). In regression analysis, the error term values typically include rational numbers which fluctuate around an average value of 0.

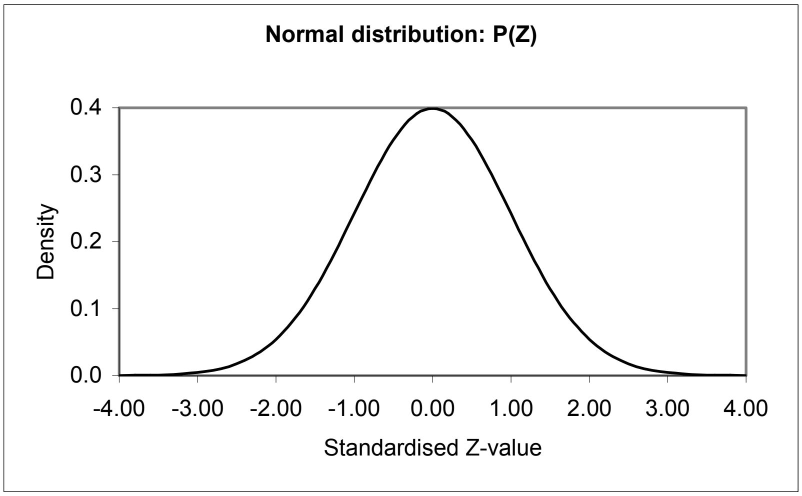
Continuous random variables, say variable X, allow for rational numbers. Continuous random variables often occur over an interval, say from -20.8 to +30.2. It is even possible that we do not even specify their minimum or maximum X-values! For example, it is possible that $-\infty \leq X \leq +\infty$ where ∞ indicates infinity, as in the case of the normal distribution.

But how do we deal with their probability density functions P(X)? The P(X) curve is defined such that the total area under the curve = 1. We cannot speak of the probability of obtaining a, say, X = 7 value. The probability of P(X = 7) would be very small. We instead deal with the probability across a range of X-values, for example: $4 \leq X \leq 7$.



An example of a discrete distribution that is approximately normally distributed is provided on the left. This refers to the case where X is the sum of six throws of the die. In this case, the minimum value of X = 6 (6 x 1) and the maximum value of X = 36 (6 x 6).

The value of X = 21 occurs most frequently. The sum of the probability of obtaining values in both tail ends (that is, relative large deviations from the average), say X ≤ 12 plus X ≥ 30 is relatively small.



We use standardised Z-values to look up probabilities of the normal distribution in which case Z = (X − μ)/σ. The probability of say -1 ≤ Z ≤ 1 is represented by the area under the P(Z) curve from Z = -1 to +1. See the chart at the left. According to table B7 the area under the curve for Z ≥ +1 is 0.1587, similarly the area below Z ≤ -1 is 0.1587. Thus the area below the curve for -1 ≥ Z ≤ 1 is 0.6826. Consequently, the probability that a continuous normal distributed random variable will fall within the range μ − σ and μ + σ is 68.26%.

LEARNING ACTIVITY 4

(a) Explain what is meant by the expected value of a random discrete variable (X). Its P(X) and X.P(X) are provided in table 17.1.4

(b) What is the meaning of bias in the case of a sample distribution (X) used to measure an unknown population parameter μ?

(c) Explain how an expected value is derived in the case of a random continuous variable Z of which P(Z) is known.

Table 17.1.4

X	P(X)	X.P(X)
3	1/216	0.0139
4	3/216	0.0556
5	6/216	0.1389
6	10/216	0.2778
7	15/216	0.4861
8	21/216	0.7778
9	25/216	1.0417
10	27/216	1.2500
11	27/216	1.3750
12	25/216	1.3889
13	21/216	1.2639
14	15/216	0.9722
15	10/216	0.6944
16	6/216	0.4444
17	3/216	0.2361
18	1/216	0.0833
Sum	1.000	10.5000

ANSWERS

The expected value of a random discrete variable (X), E(X) is its weighted average:

$\sum X.P(X) = 10.5$.

Assume that variable X is the sample estimate of a population parameter μ. Also assume the sample estimates vary from 3 to 18 and that their P(X) as in table 17.1.4.

If $E(X) = \mu = 10.5$ then the estimator is unbiased. If, say, $E(X) = 12$, while $\mu = 10.5$, then the estimator is biased. Bias occurs when the estimator tends to overestimate or underestimate the true value.

In the case of a random continuous variable Z:

$$E(Z) = \int Z.P(Z).dZ \text{ where } \int P(Z).dZ = 1.$$

LEARNING ACTIVITY 5

Psychologists tell us that the intelligence quotient (IQ) of the population is normally distributed with average $\mu = 100$ and the standard deviation $\sigma = 15$.

(a) Compile a table which indicates which proportion of the population has an IQ exceeding (or equal to) 100, 110, 120, 130, 140 and 145 respectively. In the process, also indicate the standardised Z-values. Look up the probabilities in table B-7 (the normal distribution). Also indicate how many persons of a population of 10 000 persons fall within each group.

(b) Explain why IQ is normally distributed within a population. Refer to the central limit theorem.

ANSWERS

(a)

IQ (X) (greater or equal to)	*z = (X − μ)/σ*	*Probability that Z > z*	*Number of persons in population of 10 000*
100	*0.00*	*0.5000*	*5 000*
110	*0.67*	*0.2514*	*2 514*
120	*1.33*	*0.0918*	*918*
130	*2.00*	*0.0228*	*228*
140	*2.67*	*0.0038*	*38*
145	*3.00*	*0.0013*	*13*

(b) The central limit theorem states that if Z is a standardised sum of N independent and identically distributed random variables, then the probability distribution of Z approaches the normal distribution. See p. 552. IQ is normally distributed because it reflects the cumulative outcome of a large number of hereditary and environmental factors. See p. 554.

12.2 Sampling

This section deals with topics on selection bias, survivor bias, non-response bias and the power of random selection.

The textbook provides a good overview of some sample selection methods and of sampling error. Our interest, however, does not lie with the different sample selection methods. Within econometrics, sampling is important because we use the concept of the sampling distribution. Thus, you need to focus only on the following aspects:

- the difference between the population and the sample
- the meaning of sampling error and
- the meaning of statistical inference

LEARNING ACTIVITY 6

Explain:

- What is sampling in general?
- Why do sampling concepts arise in econometrics?

ANSWERS

Sampling is the process of selecting only some units e.g. people, organisations) from a total population of interest. For example, we can select a sample of, say, 50 students from the population of 250 000 Unisa students. The beauty of sampling is that the characteristics of the sample quite often accurately reflect those of the population. Statistical inference refers to the process of estimating population parameters (mean, total, ratios, etc.) from the sample estimates, and of providing suitable measures of their accuracy.

In econometrics, we use sample data for estimation. An example is the house price regression of chapter 1 where the sample of 43 houses[1] *is a subset of all houses sold in Southern California during a given time period. In econometrics, we make the distinction between the population regression function (PRF) and the sample regression function (SRF). The PRF refers to the true but unknown regression equation. The PRF is a theoretical construct. It is not something we would normally estimate because often not all population values are known, or the population is impractical to measure. In contrast, the SRF is a practical concept. The SRF is based on data which we observe. In practice, we estimate the SRF.*

Given that we use sampling, we can expect that the sample estimates of parameters will fluctuate round their true population parameters. This is called sampling error. Parameters refer to statistical measures such as the mean or standard deviation. In econometrics our interest lies mainly with the coefficients of a regression equation – which may also be called parameters.

12.3 Estimation

This section deals with sampling distributions, the mean of the sampling distribution, the standard deviation of the sampling distribution, the t-distribution, confidence intervals and sampling from finite populations.

We expect you to understand the meaning of

- a sampling distribution, and its expected value and standard deviation
- systematic error (or bias)
- the t-distribution

The importance of the sampling distribution

[1] The sample in the book only includes real estate transactions of the past four weeks.

Please refer back to the statement made by Kennedy at the beginning of this learning unit. We sometimes have different estimators which have different sampling distributions. For example, we will come across the econometric problem of serial correlation (chapter 9) which affects the accuracy of estimates. In this case we then have the choice of two estimators, normal OLS, and the method of GLS. The choice of the better estimator then rests upon the characteristics of its sampling distribution. To determine the best estimator, we ask four questions:

- Is the estimator unbiased?
- What is the size of its standard error?
- Are the estimates of its standard error unbiased?
- What impact does an increased sample size have on these characteristics?

LEARNING ACTIVITY 7

This learning activity addresses the sampling distribution of a sample estimator. In this case the sample estimator is $\overline{X}$, that is, the average of a sample of X-values drawn from a population of X-values. The question is how will $\overline{X}$ match the true population average.

In learning unit 4 we will again deal with the sampling distribution. In that case our interest lies with the sample distribution of $\widehat{\beta}$ where $\widehat{\beta}$ is a sample estimate of a coefficient of a regression equation. In both cases, however, the principle of a sample distribution is similar.

Explain the meaning of

- the sampling distribution of $\overline{X}$.
- the expected value of $\overline{X}$, as well as bias.
- the standard deviation (also called standard error) of $\overline{X}$.

ANSWERS

The easiest way to explain the meaning of a sampling distribution is to use a simulation approach. The following steps outline this approach.

(1) The first step is to define precisely what characteristic of the population we wish to measure. Assume that we wish to determine the population average (or mean) of variable X of the population.

(2) In this case we need to determine whether the sample average ($\overline{X}$), based on a random sample, is a good estimator of the population average (µ). The goal of the procedure is to determine how well the sample estimator $\overline{X}$ performs.

(3) We create a known population by simply generating, say 50 000 random values of X.

(4) We then sample repeatedly from this population by random selection of, say, samples of 20 observations each.

(5) We calculate the sample mean of each sample ($\overline{X}$). We record these estimates into a histogram.

(6) The distribution of these estimates defines the sampling distribution of $\overline{X}$. Because we know the true mean (µ), we can determine how much the sample estimates of the mean ($\overline{X}$) deviate from µ.

Your lecturer has applied these steps in practice. First (step 3) observations (X) for the population were generated which conform to the normal distribution with the average μ = 100 and a standard deviation of X of σ = 15. This is easily done by using a PC and MS Excel.

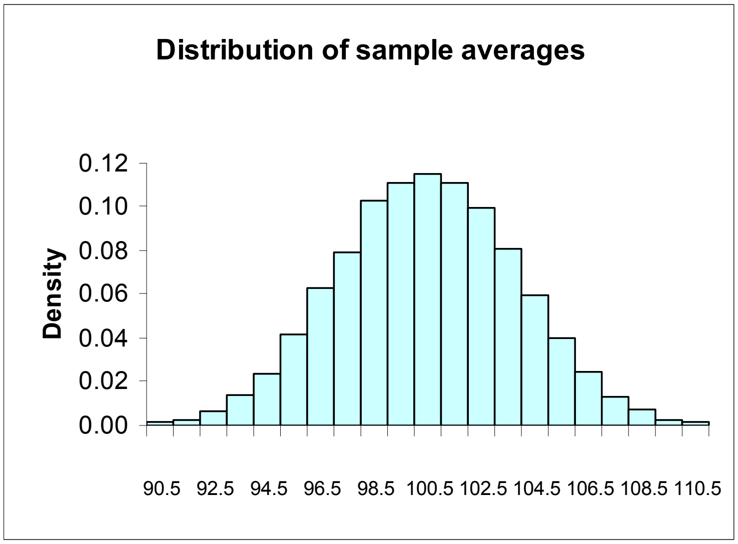
Then a large number of random samples (each of sample size 20) were selected (step 4). The sample mean of each sample ($\overline{X}$) was derived and recorded into a histogram (step 5). The histogram summarises the frequency of values of different $\overline{X}$ obtained from all samples.

With respect to the histogram, the Y-axis measures the relative occurrence of the values of $\overline{X}$. The number on the X-axis represents the upper bound, for example, 100.5 represents values of $\overline{X}$ falling between 99.5 and 100.5.

Which conclusions can we make based on this sampling distribution?

(1) The first, possibly unexpected, fact is that the outcomes of random sampling produce a well behaved distribution! The sample averages appear to cluster around the true value being estimated and the distribution is symmetric. The expected value (weighted average) of the sample means of all samples is equal (at least very close) to the true value of μ = 100. This implies that the estimator $\overline{X}$ is unbiased. Bias in the estimator occurs when the expected value of $\overline{X}$ is not equal to μ.

(2) Deviations ($\overline{X}$ – μ) do occur, which are both positive and negative. However, in most cases, these deviations are relatively small. Large deviations do occur, but the probability of this is relatively low.



(3) To further judge the accuracy of a sample value $\overline{X}$ we need information regarding the "width" of the sample distribution of $\overline{X}$. For example, in the histogram above almost all observations of $\overline{X}$ fall within the range μ − 10 to μ + 10. The SE($\overline{X}$) is such a measure (but different from the value of 10) where SE: standard error (also called standard deviation).

Statistical theory tells us that the SE($\overline{X}$) may be derived as follows:

$SE(\overline{X}) = \frac{SE(X)}{\sqrt{N}} = \frac{\sigma}{\sqrt{20}} = \frac{15}{\sqrt{20}} = 3.354$ *where N is the sample size.*

The SE($\overline{X}$) measures the "width" of the sample distribution of $\overline{X}$. Of course, the more "narrow" the sample distribution of $\overline{X}$ is, the more accurate its estimates are.

(4) The distribution of deviations $\overline{X} - \mu$ conforms to the normal distribution. This allows us to make probability statements regarding the extent of deviations $\overline{X}$ – μ. It is convenient, however, to write this in its standardised form:

$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{N}}$ *where* $\sigma / \sqrt{N}$ *is the standard error of* $\overline{X}$

σ is the SE(X) and N is the sample size.

The advantage of this form is that Z is normally distributed with an average of 0 and a standard error of 1. Because tables of the normal distribution are published in this form it is then easy to compare values of Z (obtained from the sample) to that of z (the values published in the table).

The distribution of random sampling estimates of $\overline{X}$ is highly predictable. We may assume that a single random sample estimate will conform to this behaviour.

LEARNING ACTIVITY 8

Explain the meaning of the t-distribution (with respect to sample estimator $\overline{X}$) and explain which sources of sampling variation it accounts for.

This learning activity provides some background regarding the t-distribution which will again appear in the next learning unit.

ANSWER

In learning activity 12.3.1(4) reference was made to the standardised form of $\overline{X}$ – μ, that is

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{N}} \qquad \textit{equation A}$$

Because we only have sample data, the sample will provide values for $\overline{X}$ and N. Both μ and σ are, however, unknown. The first (μ) is not really a problem due to the nature of hypothesis testing. In the next learning unit, you will learn that we simply replace μ with a fixed value, that is a value of which its "compatibility" with $\overline{X}$ is tested. The second, σ = SE(X), remains unknown.

Because the sample consists of N values of X, σ may in fact be estimated. An unbiased estimator of σ is s, where

$$s = \sqrt{\frac{\sum (X_i - \overline{X})^2}{N - 1}} \qquad \textit{equation B}$$

If we replace σ within equation A with its estimate s in equation B, then

$$t = \frac{\overline{X} - \mu}{s/\sqrt{N}} \qquad \textit{equation C}$$

Although Z is normally distributed, t is distributed like the t-distribution. The t-distribution copes with two sources of variation, that is, $\overline{X}$ *and s, which of course vary from sample to sample.*

In the next learning unit the t-value is also used to test the coefficients of a regression equation for statistical significance. You only have one sample, and this sample provides only one estimate each of $\widehat{\beta}$ *and SE(*$\widehat{\beta}$*). It is derived as*

$$t = \frac{\widehat{\beta} - \beta_0}{SE(\widehat{\beta})}$$

where β_0 *is the* H_0 *value of the coefficient being tested and* $\widehat{\beta}$ *is the sample estimate of coefficient β.*

LEARNING UNIT 4

THE CLASSICAL MODEL

ECONOMETRICS IN ACTION

Kennedy[2]: writes

> *Although many feel comfortable doing maths, they do not really understand the method of statistics. They approach statistics as a bunch of mechanical techniques applying formulas. The vast majority do not understand the method of statistics which is captured by the sampling distribution concept.*
>
> *Using a formula to estimate say statistic B (e.g. a mean) can be conceptualized as the statistician shutting his/her eyes and obtaining an estimate by reaching blindly into the sampling distribution of B to obtain a single estimate. Choosing between method B_1 and a competing method B_2 to estimate statistic B comes down to the following: Would you prefer to produce your estimate of B by reaching blindly into the sampling distribution of B_1 or by reaching blindly into the sampling distribution of B_2?*
>
> *The desirable properties of an estimator B_1 are defined in terms of its sampling distribution. For example, B_1 is unbiased if the mean of its sampling distribution equals the number B being estimated. This explains why statisticians spend so much energy figuring out sampling distribution properties.*

In econometrics, the sampling distribution is of fundamental importance. This learning unit deals mainly with the sampling distribution of $\widehat{\beta}$.

STUDY OBJECTIVES

When you have studied this learning unit you should

- understand the statistical requirements for the OLS method to perform well
- be familiar with standard notation
- understand why we may assume that the regression equation error term is normally distributed
- understand why the $\widehat{\beta}$ estimator has a sampling distribution
- understand that the performance of an estimator depends on its sampling distribution
- that is, bias, minimum variance and consistency

This chapter requires a good understanding of statistical concepts dealt with in chapter 12.

[2] Kennedy, PE. 2001. Bootstrapping Student Understanding of What is Going On in Econometrics. *Journal of Economic Education*, 32, 110–123.

(A) PRESCRIBED MATERIAL

For OLS to work well, a number of conditions must be met and these are introduced one by one. In chapters 8 to 10 we will study the implications when some of these assumptions are not met. Thus, some of the material will also be covered in more detail in later chapters.

The following sections are prescribed:

(1) The classical assumptions.
(2) The sampling distribution of $\hat{\beta}$.
(3) The Gauss-Markov theorem and the properties of OLS estimators.
(4) Standard econometric notation.

(B) SOME IMPORTANT CONCEPTS

4.1 The Classical Assumptions

It is important that you understand the meaning of the classical assumptions. Because the classical assumptions are stated in mathematical notation, you must be familiar with standard econometric notation. The classical assumptions are as follows:

Specification

- The regression model is linear in the coefficients, is correctly specified, and has an additive error term.

  $$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i$$

  Some specification problems are dealt with in chapters 6 and 7.

Error term

- The error term has a zero population mean $E(\varepsilon_i) = 0$ for all i = 1 … n.
- All explanatory variables are uncorrelated with the error term $E(r_{X_{ki}},_{\varepsilon_i}) = 0$ for k = 1 … K and i = i … n where r_{X_i,ε_i} denotes the simple correlation coefficient between variable X_{ki} and the error term, ε_i. This problem occurs in simultaneous equation models, which is not dealt with in this course.
- Observations of the error term are uncorrelated with each other $E(r_{\varepsilon_i \varepsilon_j}) = 0$ for all i, j = 1 … n except i = j. The violation of this requirement is called serial correlation. This is studied in chapter 9.
- The error term has constant variance $VAR(\varepsilon_i) = \sigma_i{}^2 = \sigma^2$ for all i = 1 … n. The violation of this requirement is called heteroskedasticity. This is studied in chapter 8.
- The error term is normally distributed. Strictly speaking, this condition is not required to perform OLS but it is required for statistical testing.

Explanatory variables (X_k's)

- No explanatory variable is a perfect linear function of any other explanatory variable(s). The violation of this requirement is called multicollinearity. This is studied in chapter 11.

You should know the meaning of each of these assumptions. We expect you to use standard notation when referring to these assumptions.

Many students learn these assumptions by heart without knowing their meaning. In the examination I often get answers that refer to, say, assumption III or whatever, without further explanation. This approach is not recommended. Rather focus on the meaning of these assumptions. And if you refer to these assumptions, then explain their meaning, using mathematical notation of course.

For standard econometric notation see section 4 (p. 112 of prescribed book)

This course uses mathematical notation. In the previous section this was used to specify the regression model and to set out the requirements for OLS, amongst other things. It is a convenient form of expressing complex concepts or conditions for which words are a rather inconvenient form of expression. Please familiarise yourself with this form of notation. Table 4.1 provides a summary of these terms. Please use them in the examination.

In addition, we refer in consistent terms throughout the textbook to the concepts listed below:

- n denotes the number of sample observations.
- K denotes the number of explanatory variables excluding the constant term. The constant term is β_0 while β_K refers to the k-th slope coefficient of the regression equation.
- We use the summation sign (∑) as a shorthand operator. Make sure you understand its meaning. It is dealt with in chapter 12.

Expected value

Some of these classical assumptions refer to "expected value". In the case of a single random variable, the meaning of *expected value* is explained in learning unit 12.

LEARNING ACTIVITY 1

Explain the meaning of $E(\hat{\beta})$.

ANSWER

In the case of $E(\hat{\beta})$ the following is implied:

- *The sample estimator $\hat{\beta}$ is a continuous random variable in the sense that it can assume a range of possible values of which the outcome depends on chance.*
- *There exists a probability distribution $P(\hat{\beta})$ for the range of values of $\hat{\beta}$. $\hat{\beta}$ could, for example, be normally distributed.*

- $E(\widehat{\beta}) = \sum \widehat{\beta}.P(\widehat{\beta})$ *in the case of a discrete variable*[3]*.* $E(\widehat{\beta})$ *is the weighted average of all possible* $\widehat{\beta}$ *values where their probabilities are given by the probability density function of* $\widehat{\beta}$ *called P(* $\widehat{\beta}$ *).*

LEARNING ACTIVITY 2

Explain the meaning of E(εᵢ) = 0 where εᵢ is normally distributed.

ANSWER

- *ε_i is a random variable in the sense that it can assume a range of values, from negative to positive.*
- *There exists a probability distribution P(ε_i) which conforms to the normal distribution.*
- $E(\varepsilon_i) = \sum \varepsilon_i P(\varepsilon_i) = 0$*, that is, the weighted average of all possible ε_i values is zero.*

Econometrics also deals with the expected relationship between variables. For example, if X_{ki} and ε_i are positively related, we would expect that if X_{ki} increases, then ε_i would also tend to increase, and if X_{ki} decreases, ε_i would decrease. To represent this interdependency between variables, the textbook uses the expected value of the correlation coefficient between the two variables, for example in $E(r_{X_{ki}}, {}_{\varepsilon_i})$*. In the case of* $E(r_{X_{ki}}, {}_{\varepsilon_i}) = 0$*, the two variables X_{ki} and ε_i are uncorrelated.*

If, for example, $E(r_{X_{ki}}, {}_{\varepsilon_i}) < 0$ *then there is an expected negative linear relationship between X_{ki} and ε_i, which means that if X_{ki} increases, we could expect ε_i to decrease, and vice versa.*

If $E(r_{X_{ki}}, {}_{\varepsilon_i}) = 0$ *applies for all k = 1 to K, and all i = 1 to n, this means that not one of the explanatory variables X_{ki} (k = 1 … K) has a linear relationship to any other one of ε_i (i = 1 … n). This is the nature of our assumption that all explanatory variables are uncorrelated with the error term: for all k = 1 … K, i = 1 … n.*

The error term is normally distributed

One of the assumptions of the classical model is that the error term is normally distributed. The error term is a random variable of which there is good reason to believe that it is normally distributed. The *bell-shaped* normal distribution is explained in chapter 12.

Consider the population regression function

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i$$

[3] *We use* $\int \widehat{\beta}.P(\widehat{\beta})\delta\widehat{\beta}$ in the case of a continuous variable.

The error term (ε_i) consists of four major components. They are:

- omitted Xs. In regression equations we include only the X's which have a major impact on Y and we exclude X's which have minor effects. A simple model, having a few X's, is often better than a more complicated model having more X's.
- measurement errors. It is often difficult to measure economic phenomena accurately. Often, economic data are affected, for example, by sampling errors.
- errors in functional form. We can never be sure of the precise functional relationship. The one we use is often only approximately true.
- purely random errors. The behaviour of economic units, and of humans, often includes a pure random component. Humans do not react in a mechanistic way and their individual behaviour is not always perfectly predictable.

Note that this applies to the error term of the population regression function.

The central limit theorem

The central limit theorem supports the assumption that the error term is normally distributed.

- The central limit theorem states that the sum of a number of independent and identically distributed variables will tend to be normally distributed.
- We previously stated that the error term has several components. They are: omitted Xs, measurement errors, errors in functional form and purely random errors. Thus each ε_i consists of four error components:

$$\epsilon_i = \epsilon_{i,\text{omitted Xs}} + \epsilon_{i\ ,\text{measurement errors}} + \epsilon_{i,\text{errors in functional form}} + \epsilon_{i,\text{random errors}}.$$

This is the situation the central limit theorem refers to. The concept of a sum of a number of independent (and identically distributed variables) applies to the sum of the components of the error term.

- Is there a simply intuitive explanation for the phenomenon that the sum of variables is normally distributed? Yes, there is. It is rather unlikely that the components of the sum will all operate in the same direction for a particular observation of ε_i. It is more likely that the different components of ε_i will tend to cancel each other, because they are independent.
- The requirement of identically distributed variables prevents one component of ε_i from dominating, which would happen if one source of error were much larger than the others.

4.2 The sampling distribution of $\widehat{\beta}$

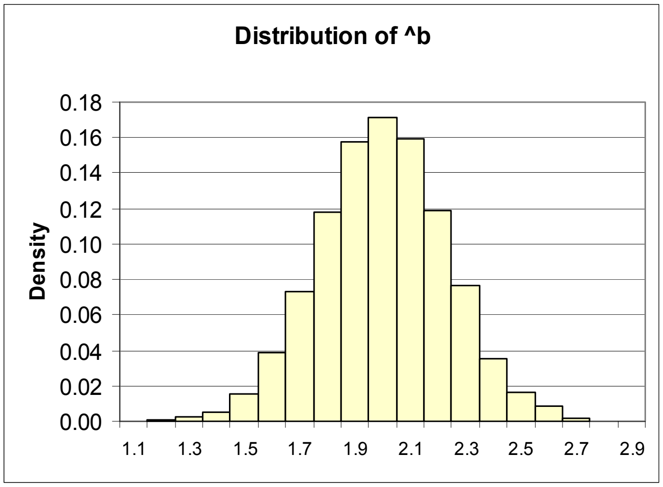
The estimated regression coefficient $\widehat{\beta}$ can be viewed as a random variable with its own probability distribution $P(\widehat{\beta})$. This very important concept is summarised below.

LEARNING ACTIVITY 3

Explain the meaning of the sampling distribution of $\widehat{\beta}$. Pay attention to the purpose of $\widehat{\beta}$, its probability density function (feel free to use a chart), why sampling variation occurs, its statistical distribution and how the sampling distribution may be used to measure the performance of $\hat{\beta}$. (15)

ANSWER

- *The purpose of sample estimator $\hat{\beta}$ (a SRF) is to estimate the regression population parameters as in $Y = \beta_0 + \beta_1 X$ (the PRF). Since a random sample is a subset of the population, it is natural for sampling variation $\hat{\beta} - \beta$ to occur, simply because the sample is not always a good representation of the population.*



- *The sampling distribution of an estimator shows what sampling variation we may expect from sample estimates. The sampling distribution of $\hat{\beta}$ is illustrated on the left by its probability density function. The horizontal axis measures $\hat{\beta}$ while the vertical axis measures probability density, such that the area under the curve is one. The sampling distribution shows that the outcomes of random samples are well behaved. The $\hat{\beta}$'s appear to cluster around the true value being estimated and the distribution appears symmetric. Indeed, large deviations do occur, but the probability of large deviations is relatively small.*
- *Repeated sampling from the population is one possible way to derive the probability density function of a sample estimator. One could, for example, draw 300 samples from the population and derive $\hat{\beta}$ for each sample. One could then display the outcomes of the estimates of $\hat{\beta}$ by means of a histogram, which demonstrates that certain values of $\hat{\beta}$ occur more frequently than others. The lecturer used this method to derive the graph above.*
- *Note that the standard error of $\hat{\beta}$ may also be derived. Of course, the SE($\hat{\beta}$) measures the "width" of the sampling distribution of $\hat{\beta}$.*
- *The sampling distribution of $\hat{\beta}$ may also be derived theoretically (see the next section). The outcome is that*

$$\widehat{\beta}_1 = \beta_1 + w_1\varepsilon_1 + w_2\varepsilon_2 + ... + w_n\varepsilon_n \qquad \textit{(equation C)}$$

which may also be written as

$$\widehat{\beta}_1 - \beta_1 = w_1\varepsilon_1 + w_2\varepsilon_2 + ... + w_n\varepsilon_n \qquad \textit{(equation D)}$$

- *Equation C shows that $\hat{\beta}_1$ depends on just three things: the true β_1 which is fixed, a set of factors (w_i) which depend on the X_i's and which show some variation across different samples and finally the ε_i's from which most variation arises. Thus the variation in $\hat{\beta}_1$ depends mostly on the values of the set of (true) error terms of the observations contained in the sample.*
- *Because we can assume that the error terms are normally distributed (according to the central limit theorem), it follows that $\hat{\beta} - \beta$ (see equation D) is normally distributed as well. This is based on a statistical theorem which we have to assume.*
- *There is another source of variation to be dealt with, however. Just as $\widehat{\beta}$ varies across samples, the SE($\widehat{\beta}$) may also be expected to vary across samples. The t-value, $t = \dfrac{\hat{\beta} - \beta}{SE(\hat{\beta})}$ takes account of this additional variation, where t is distributed according to the t-distribution.*

The sampling distribution of $\widehat{\beta}$ is an important indicator of the performance of $\widehat{\beta}$.

- *The estimator is unbiased if the expected value (weighted average) of all possible estimates is equal to the true value being estimated.*
- *The estimator is minimum variance if no other estimator has a smaller SE($\widehat{\beta}$). The smaller the SE($\widehat{\beta}$), the more accurate the estimates of $\widehat{\beta}$ will be.*
- *The estimator is consistent if increasing the sample size (1) causes $\widehat{\beta}$ to remain unbiased and (2) causes the SE($\widehat{\beta}$) to decrease.*

Theoretical derivation of the sample variation $\hat{\beta} - \beta$

In the prescribed book, the sampling distribution is dealt with in section 2 using repeated sampling. This section demonstrates an alternative method of deriving the sampling distribution, that is, in a theoretical manner. In the case of regression analysis, we are primarily interested in estimating the slope coefficients of the regression equation. Thus, in the example below we deal primarily with the sampling distribution of estimator $\widehat{\beta}_1$ of the regression equation (PRF): $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$.

- A sample of n observations of (Y_i, X_i) is selected from the population. Assume we use OLS to estimate the slope coefficient. The formula for the OLS estimator is:

$$\widehat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} \text{ where } \begin{aligned} x_i &= X_i - \overline{X} \ and \\ y_i &= Y_i - \overline{Y} \end{aligned} \qquad \textit{(equation 4.2.1)}$$

- Hidden within every Y_i observation are the unknowns: β_1 and ε_i according to $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. Thus, the selection of the sample implies simultaneously selecting, amongst others, n error terms. These error terms will vary from sample to sample in the case of repeated sampling.
- If $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ is substituted into equation 4.2.1, we ultimately get the very elegant expression:

  $$\widehat{\beta}_1 - \beta_1 = w_1\varepsilon_1 + w_2\varepsilon_2 + ... + w_n\varepsilon_n \qquad \textit{(equation 4.2.2)}$$

  where $w_i = \frac{x_i}{\sum x_i^2}$ for i = 1 to n and ε_i are the true error terms associated with sample observations i = 1 ton.

Equation 4.2.2 is derived in appendix 4.1 at the end of this learning unit. It is not derived in the book. The textbook uses a different intuitive method, namely that of repeated sampling, to explain the concept of the sampling distribution of $\hat{\beta}_1$. Make sure you understand at least one of these procedures. We previously used repeated sampling (Chapter 17) to explain the sampling distribution in a general sense. However, with respect to the sampling distribution of $\hat{\beta}_1$, equation 4.2.2 provides an exact and powerful expression to explain sampling variation.

You may refer to equation 4.2.2 in the examination although its derivation is not examination material. This equation may be used to demonstrate some characteristics of the sampling distribution most elegantly.

- The sample variation $\widehat{\beta}_1 - \beta_1 = w_1\varepsilon_1 + w_2\varepsilon_2 + ... + w_n\varepsilon_n$ originates due to the composition of the sample. Sampling variation occurs because each sample has a different set of ε_i's and w_i's. The sample variation $\widehat{\beta}_1 - \beta_1$ is simply a "weighted" average of n population error terms contained in the sample with the factors $w_i = \frac{x_i}{\sum x_i^2}$. Note that some w_i factors will be negative fractions while others are positive, and that $\Sigma w_i = 0$ (because $\Sigma x_i = 0$).
- Although different samples will produce different factors, the effect of these factors on sample variation is likely to be small. The variation of the error terms in $\widehat{\beta}_1 - \beta_1 = w_1\varepsilon_1 + w_2\varepsilon_2 + ... + w_n\varepsilon_n$ is likely to be much more dominant.
- $\widehat{\beta}_1 - \beta_1$ will be normally distributed. In the previous section we noted that each ε_i is likely to be normally distributed because each error term reflects the sum of four independent effects. Because $\widehat{\beta}_1 - \beta_1$ is a linear combination of n error terms, each of which is normally distributed, statistical theory tells us that $\widehat{\beta}_1 - \beta_1$ will then be normally distributed as well.

LEARNING ACTIVITY 4

You wish to estimate Y = β_0 + β_1X and observe the following data:

	X	Y	x = $X - \overline{X}$	y = $Y - \overline{Y}$	x.y	x^2
	2	7	-4	8.4	-33.6	16
	4	5	-2	6.4	-12.8	4
	6	-4	0	-2.6	0.0	0
	8	-4	2	-2.6	-5.2	4
	10	-11	4	-9.6	-38.4	16
Average	6	-1.4				
Sum					-90.0	40

Assume that the true relationship (PRF) is: $Y_i = 10 - 2X_i$.

(a) Calculate $\hat{\beta}_1$ by OLS and then calculate $\hat{\beta}_1 - \beta_1$.

(b) Calculate the true error terms (ε_i) for each observation.

(c) Confirm by calculation that

$$\hat{\beta} - \beta = \sum w_i \varepsilon_i \text{ where} w_i = \frac{x_i}{\sum x_i^2}.$$

Explain the practical relevance of expression (c) since we cannot observe the true β_1 nor the true ε_i's.

ANSWERS

$$\hat{\beta}_1 = \frac{\sum xy}{\sum x^2} = \frac{-90}{40} = -2.25$$

(a) *thus* $\hat{\beta}_1 - \beta_1 = -2.25 - (-2) = -0.25$

(b) *The error terms in the table below are based on the true coefficients. For example, $Y_{1,true}$ = 10 – 2(2) = 6. Thus ε_1 = 7 – 6 = 1.*

X_i	***Y_i***	$Y_{true,i} =$ $10 - 2X_i$	$\varepsilon_i =$ $Y_i - \hat{Y}_{true,i}$	$w_i = \frac{x_i}{\sum x_i^2}$	$w_i.\varepsilon_i$
2	*7*	*6*	*1*	*-0.10*	*-0.10*
4	*5*	*2*	*3*	*-0.05*	*-0.15*
6	*-4*	*-2*	*-2*	*0.00*	*0.00*
8	*-4*	*-6*	*2*	*0.05*	*0.10*
10	*-11*	*-10*	*-1*	*0.10*	*-0.10*
Sum			*3*	*0.00*	*-0.25*

$$\hat{\beta}_1 - \beta_1 = \sum w_i \varepsilon_i \text{ where } w_i = \frac{x_i}{\sum x_i^2}$$

(c) *See the table.*

(d) *True, the population coefficients β₀ and β₁ are unknown, which means that the εᵢ's are unknown as well. However, the expression shows theoretically how the sampling error $\hat{\beta}_1 - \beta_1$ depends on the true error terms. It shows how the selection of different samples (which imply different εᵢs and wᵢs) gives rise to the different sampling errors of $\hat{\beta}_1 - \beta_1$. It may also be used to derive the sampling distribution of $\hat{\beta}_1 - \beta_1$.*

4.3 Gauss-Markov theorem and properties of OLS estimators

Statisticians judge the quality of a sample estimator by looking at its sampling distribution. Sample estimators are good if they are

- unbiased
- minimum variance
- consistent

Make sure these criteria are well understood.

How good is the OLS estimator by these standards? The Gauss-Markov theorem states that if the classical assumptions hold, then the OLS estimator is a *best estimator*, because the OLS estimators are

- unbiased
- minimum variance
- consistent
- normally distributed (if the SE is known)

4.4 Standard econometric notation

You should be familiar with the standard notation as it is a shorthand method of reference.

(C) TRUE/FALSE QUESTIONS (F) = false (T) = true

(1) Student X confirms that all seven classical assumptions hold and proceeds to estimate the coefficients of a regression equation by OLS. Student X is assured that his OLS estimates are necessarily accurate. (F)

The Gauss-Markov theorem states that if all seven classical assumptions hold, then the OLS estimator is BLUE, that is, it is the best linear, minimum variance and unbiased estimator. However, owing to random variation, the OLS estimates may still be inaccurate, although large deviations are unlikely to occur.

(2) The t-test is valid only if the error term is normally distributed. (T)

The t-test rests upon the assumption that the error term is normally distributed.

(3) The central limit theorem and the Gauss-Markov theorem are essentially the same. (F)

The central limit theorem deals with the distribution of error terms, while the Gauss-Markov theorem deals with the characteristics of the OLS estimator.

(4) The degrees of freedom df = n – K – 1. If n = 5, then OLS can derive estimates when K = 4. (F)

The application of OLS requires that df > 0.

(5) A minimum variance estimator is necessarily unbiased. (F)

A minimum variance estimator is not necessarily unbiased, and an unbiased estimator is not necessarily a minimum variance one.

(D) EXAMINATION: PARAGRAPH QUESTIONS

(1) Briefly explain the meaning and purpose of each of the seven classical assumptions using standard notation.

(2) What is the rationale of including a stochastic error term in a regression equation? Explain what factors contribute to the error term. (8)

(3) What relevance does the central limit theorem have for the distribution of the error term? (4)

(4) Explain what is the meaning of the sampling distribution of $\hat{\beta}$. Also explain: (15)

Which factor(s), in principle, give(s) rise to a variation in $\hat{\beta}$

The meaning of its probability density function P($\hat{\beta}$).

Why P($\hat{\beta}$) is likely to be normally distributed.

The meaning of the expected value of $\hat{\beta}$. When is an estimator unbiased?

The meaning of the $SE(\hat{\beta})$ and what is a minimum variance estimator? When is an estimator consistent?

(5) Explain briefly whether or not an unbiased estimator is necessarily better than a minimum-variance estimator. Demonstrate graphically. (5)

(6) With regard to the OLS estimators, explain the meaning and relevance of (6)

The Gauss-Markov theorem
A BLUE estimator

(E) EXAMINATION: PRACTICAL QUESTIONS

Most of this chapter deals with theoretical aspects and most exam questions will necessarily address these. However, make sure you can interpret the formulas in the formula sheet and that you can apply them.

APPENDIX 4.1 (not for examination purposes)

Derivation of equation 4.3.2: $\hat{\beta}_1 - \beta_1 = w_1\varepsilon_1 + w_2\varepsilon_2 + \ldots + w_n\varepsilon_n$ where $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Box 1:

$$\boxed{\begin{aligned} \hat{\beta}_1 &= \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i(\beta_1 x_i + \varepsilon_i - \bar{\varepsilon})}{\sum x_i^2} \\ &= \frac{\beta_1 \sum x_i^2 + \sum x_i(\varepsilon_i - \bar{\varepsilon})}{\sum x_i^2} \\ &= \beta_1 + \frac{\sum x_i \varepsilon_i}{\sum x_i^2} = \beta_1 + \sum w_i \varepsilon_i \end{aligned}}$$

where $y_i = \beta_1 x_i + (\varepsilon_i - \bar{\varepsilon})$:

Box 2: Proof that $y_i = \beta_1 x_i + (\varepsilon_i - \bar{\varepsilon})$ used in Box 1

$$\boxed{\begin{aligned} y_i &= Y_i - \bar{Y} = \beta_0 + \beta_1 X_i + \varepsilon_i - \beta_0 - \beta_1 \bar{X} - \bar{\varepsilon} \\ &= (\beta_1 X_i - \beta_1 \bar{X}) + (\varepsilon_i - \bar{\varepsilon}) \\ &= \beta_1 (X_i - \bar{X}) + (\varepsilon_i - \bar{\varepsilon}) = \beta_1 x_i + (\varepsilon_i - \bar{\varepsilon}) \end{aligned}}$$

Box 3: Proof that $\sum x_i.\bar{\varepsilon} = 0$ used in Box 2

$$\boxed{\sum x_i.\bar{\varepsilon} = \bar{\varepsilon}\sum x_i = \bar{\varepsilon}\sum (X_i - \bar{X}) = 0}$$

LEARNING UNIT 5

BASIC STATISTICS AND HYPOTHESIS TESTING

ECONOMETRICS IN ACTION

Deirdre McCloskey teaches the diverse fields of economics, history, and English at the University of Illinois, and economics, philosophy, and art and cultural studies at Erasmus University, Rotterdam.

She (previously he) is a colourful and controversial figure. Deirdre McCloskey has made a name for herself by critically examining the logic and rhetoric of economic arguments.[4]

She studied all the empirical articles published in the American Economic Review in the 1980s. She reports that 96% of them confused statistical significance and substantive significance. She explains:

The problem is that a number fitted from the world's experiments can be important economically without being noise-free. And it can be wonderfully noise-free without being important.

McCloskey concludes:

> *Most of what appears in the best journals of economics is unscientific rubbish.*

What is the difference between **statistical significance** and **substantive significance**?

When McCloskey refers to statistical significance (of estimated coefficients), she refers to estimates being "noise free". The "noise" represents the random variation present in estimates. The term substantive significance refers to the economic importance of an X-variable. In this module we use the term "economic significance" of an X-variable.

STUDY OBJECTIVES

When you have studied this learning unit you should

- understand the goal, method and risks of hypothesis testing
- know how to apply t-tests and F-tests

(A) PRESCRIBED MATERIAL

This learning unit deals with the statistical tests used in regression analysis. To fully understand this chapter you need to have studied chapter 12 (Statistical principles) of the prescribed book which deals with the statistical foundations.

[4] McCloskey, D. 2002. *The Secret Sins of Economics*. Prickly Paradigm Press.

The following sections are prescribed

What is hypothesis testing?

(1) The t-test
(2) Examples of t-tests
(3) Limitations of the t-test
(4) Appendix: The F-test

(B) SOME IMPORTANT CONCEPTS

5.1 What is hypothesis testing?

(a) Which steps are applied in hypothesis testing?

Hypothesis testing involves the following steps:

- setting the null and alternative hypotheses
- choosing the level of significance (the critical value)
- performing the test
- taking the decision: accepting or rejecting the null hypothesis

In the textbook, these concepts are explained with reference, first, to econometric testing, and then to the different possibilities faced by somebody accused of murder. In this guide, a third example is used, namely the example of students who write an examination, as it provides a convenient vehicle to illustrate the basic elements of hypothesis testing. The examination is the test and the 50% pass mark is the critical value which determines the fail or pass. Lastly, the fail or pass decisions could each be in error.

(b) How does one set up the null and the alternative hypotheses?

The goal of hypothesis testing is to select one of two rather simplistic alternatives, called the null (H_0) and alternative (H_a) hypotheses. H_0 is generally a statement of what the researcher does not expect to occur, and H_a is the hypothesis which is expected. Hypothesis testing always starts off by setting these alternatives first.

In the case of the examination example, these are as follows:

- H_0: The student does not know the work.
- H_a: The student knows the work.

Let's take an econometric example. Assume a researcher estimates a demand function:

$$Q = \beta_0 + \beta_P.\text{Price} + \beta_{Inc}.\text{Income}$$

where Q: Quantity demanded of the good, Price: Price of the good.

We would expect that $\beta_P < 0$. Thus the respective hypotheses are:

- H_0: $\beta_P \geq 0$
- H_a: $\beta_P < 0$

In econometrics the goal of hypotheses testing is to decide which of H_0 and H_a is most compatible with the empirical data.

(c) What is the critical value and how does it affect errors in testing?

In the examination example, the critical value is 50%. An examination mark below 50% is typically a fail and above 50% is a pass. In econometric testing the critical value is looked up in statistical tables. This provides a benchmark value which determines which hypothesis is rejected or accepted.

The critical value affects the probability of both types of errors which occur in testing. Virtually all testing is imperfect. Differences between the test result and the true state of affairs may occur. The possible outcomes are summarised in table 5.1 with respect to the examination example.

Table 5.1 Possible errors in testing		State of the truth	
		Student does not know the work **H_o is true**	Student knows the work **H_a is true**
Test result and decision	Fail the student **H_o is not rejected**	Correct decision	**Type II** error Incorrectly failing a good student
	Pass the student **H_o is rejected**	**Type I** error Incorrectly passing a bad student	Correct decision

There are two types of errors. A type I error occurs

- when a student deserves to fail (H_0 is true) but
- incorrectly passes the examination (H_0 is incorrectly rejected)

Some reasons why this could occur, are for example:

The student spotted just the right questions; the lecturer marked the examination script late at night and missed some serious mistakes made by the student, or added up the marks incorrectly in the student's favour.

A type II error occurs

- when a student knows his/her work (H_a is true)
- but fails the examination (H_a is incorrectly rejected).

This could be due to any one or more of the following reasons:

The variability of human nature – the examination day is simply a bad day, the student has had a bad night's sleep, etc.; the variability of external conditions the student crashed his/her car on the way to the examination location, arrived late for the examination, etc.; due to random chance the student was insufficiently prepared for the examination questions asked, while he/she knew the other work well; the student misinterpreted some of the questions.

The selection of the critical value (pass mark) is not totally objective. We could use a different critical value, say, a more lenient 45% pass mark or a stricter 55% pass mark. In the case of a 45% pass mark:

- The good thing would be that the type II error would decrease. A lower proportion of students who deserve to pass, would be incorrectly failed.
- The bad thing would be that the type I error would increase. A larger proportion of students who do not know the work, would pass.

In similar fashion a 55% pass mark

- would lead to a decrease in the proportion of bad students who are passed (type I error) but
- at the cost of an increase in the proportion of good students who are failed (type II error).

In general, the consequence of adjusting the critical value is that one type of error can only be decreased at the cost of increasing the other error type.

In econometrics, the same type of errors may occur. The focus in econometrics is on the type I error, also called the level of significance, which determines the critical value.

Let's refer to our previous example:

- H_0: $\beta_P \geq 0$
- H_a: $\beta_P < 0$.

A type I error occurs when H_0 is true, but is incorrectly rejected.

(d) How is a regression coefficient tested by means of a t-test?

Assume that $\widehat{\beta}$ = 14, $SE(\widehat{\beta})$ = 7, degrees of freedom (n-K-1) = 20, and a 5% level of significance applies. Further assume H_a: β > 0 and H_0: β ≤ 0.

- We derive the t-value: $t = \dfrac{\widehat{\beta} - \beta_{H_o}}{SE(\widehat{\beta})}$. Usually β_{H0} = 0. Thus t = 2.
- We look up the critical t-value in statistical tables: $t_{5\%,\text{ one-sided, }20}$ = 1.725.
- We then take a decision. We reject H_0 (accept H_a) if two conditions are met. First, the empirical sign of the coefficient (+14) must be correct. This condition is met because H_a: β > 0 and $\widehat{\beta}$ = 14. Secondly, the absolute value of the observed t-value must exceed the critical t-value. This condition is also met because 2 > 1.725.

Thus, we reject H_0 meaning that H_a: $\beta > 0$ is more compatible with our regression outcome.

(e) What is the meaning of a 5% level of significance in a t-test and why is it linked to the critical value?

The level of significance is subjectively selected by the researcher (usually 10%, 5% or 1%). It also determines the critical value in testing. Thus the choice of the level of significance may affect the outcome of the test.

Estimates of the coefficients of a regression equation are not "noise free", that is, they are subject to random variation. In the previous learning unit we explained why sampling error $\hat{\beta} - \beta$ occurs in sample estimates. In principle the purpose of hypothesis testing is to account for sample variation, that is, to determine whether an estimate $\widehat{\beta}$ is "sufficiently" different from its null hypothesis value which typically includes the value $\beta = 0$.

Let's assume our previous example:

- H_0: $\beta_P \geq 0$
- H_a: $\beta_P < 0$

A type I error occurs when H_0 is true, but is incorrectly rejected, that is, H_a is accepted. For H_a to be accepted, a value of $\beta_P < 0$ must have occurred in spite of the fact that H_0: $\beta_P \geq 0$ is true. Let us assume that the value of H_0: $\beta_P = 0$ applies (it meets the condition $\beta_P \geq 0$ and it is the "nearest threat" to H_a). How is it possible that an empirical value of $\hat{\beta}_P < 0$ could have occurred if in truth $\beta_P = 0$?

The answer, of course, is that $\hat{\beta}_P$ is subject to sampling error. Sampling error causes a random variation in $\hat{\beta}_P$. But there is more than that. The extent of the variation and its probability are also perfectly predictable. Because $t = \dfrac{\widehat{\beta}_P - 0}{SE(\widehat{\beta}_P)}$ adheres to the t-distribution, we may use t-tables to lookup the critical value of t which matches a given probability. For example, the critical t-value, $t_{5\%,\text{one-sided},20} = 1.725$. Even if $\beta_P = 0$ is true, values of $t < -1.725$ may occur, but their probability of occurrence is 5%. Because $t = -1.725$ is a critical value which defines the "border" value at which H_a is accepted, this implies that we will reject H_0, even when it is true.

In summary, the level of significance and the critical value are linked. They determine the probability of a type I error.

Which is best: a 5% or a 1% level of significance? It all depends on the risk the researcher is willing to take. In econometrics we usually choose a 5% level of significance, but 1% and 10% are also used. There is a trade-off. We can reduce the probability of a type I error (incorrectly rejecting a true H_0) by using a 1% rather than a 5% level of significance. But in doing so, we increase the probability of a type II error (incorrectly not rejecting H_0 when H_a is true.

We cannot control the type II error. This probability depends on the true value of H$_a$, which is of course unknown. Thus no strong statements about the type II error can be made. All we can say is that the probability of a type II error increases as we reduce the probability of a type I error.

5.2 The t-test

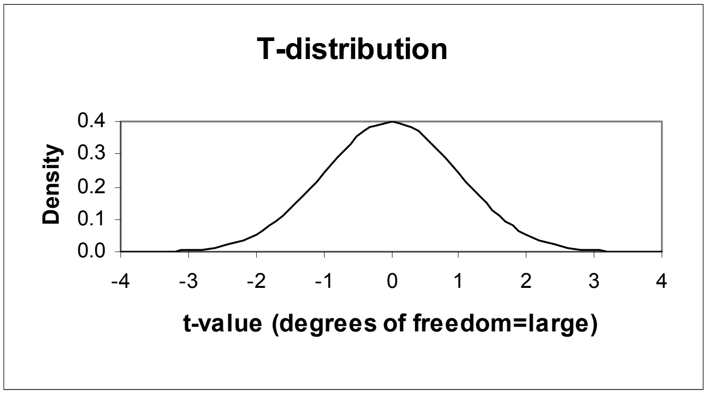
(a) Why do we use the t-test to test coefficients of a regression equation?

The t-test is usually applied to the slope coefficients of a regression equation

$$t = \frac{\widehat{\beta}_k - \beta_{H_o}}{SE(\widehat{\beta}_k)} \ .$$

In particular, the t-distribution is used rather than the normal distribution. The formula above implies two sources of variation, that is, both $\widehat{\beta}_k$ and $SE(\widehat{\beta}_k)$. Both vary due to sampling error. The t-distribution accounts for both sources of variation.

The t-distribution is very similar to the normal distribution but also accounts for the degrees of freedom: n-K-1 where n: sample size and K: number of X-variables included in the regression equation. The t-distribution is somewhat wide and more flat than the normal distribution.



Make sure you understand when and how to perform the following variations of the t-test:

Case	Expected sign of β	H$_a$	H$_0$
One-sided	Negative	β < 0	β ≥ 0
	Positive	β > 0	β ≤ 0
Two-sided	Unsure of sign	β ≠ 0	β = 0

5.3 Examples of t-tests

LEARNING ACTIVITY 1

Example of testing the coefficients of a regression equation using the t-test

Regression equation ONE, which explains beer consumption (QB), has been estimated as follows:

$$\begin{array}{lcccccccccc} QB = & +1200 & – & 15(PB) & + 5(POTH) & – & 0.02(INC) & + 6(TIME) \\ & [200] & & [4] & [3] & & [0.01] & [3] \end{array}$$

R^2 = 0.98; n = 35 (data: 1971–2005); and where numbers in brackets […] denote the respective standard errors of their coefficients.

The variables have the following meaning:

QB: Domestic consumption of beer (volume index, 1971 = 100)

PB: Real price of beer (nominal price of beer deflated by CPI, index 1971 = 100)

POTH: Real price of other alcoholic beverages (nominal price deflated by CPI, index 1971 = 100)

INC: Real per capita income (Rand per person per year at 2005 prices).

TIME: Assumes values of 1, 2, 3, 4,… where 1 = 1971, 2 = 1972, 3 = 1973, … , 35 = 2005.

Set appropriate hypotheses for the coefficients and then test the estimated coefficients for statistical significance at the 5% level. Hint: Please present your answers in table format. In the case of TIME, assume you are unsure of its expected sign on the basis of theory. (10)

ANSWER

A table provides a compact way to present the results. The numbers in brackets (example ½) denote the marks allocated.

X variable	***Expected value of coefficient***	***H_0 & H_a***	***t-value***	***Decision***	***Marks***
	N-k-1 = 35 – 4 – 1 = 30 (½)	*Critical t-value 1-sided (½)*		*$t_{30.5\%,1-s}$ = 1.697 (½)*	*(1½)*
PB	*Expect negative (½)*	*H_0: β > = 0* *H_a: β < 0 (½)*	*-3.75 (½)*	*Reject H_0* *Abs(t) > 1.697 (½)*	*(2)*
POTH	*Expect positive (½)*	*H_0: β < = 0* *H_a: β > 0 (½)*	*+1.67 (½)*	*Not reject H_o* *1.67 < 1.697 (½)*	*(2)*

INC	*Expect positive (½)*	*H_0: β < = 0* *H_a: β > 0 (½)*	*-2* *(½)*	*Not reject H_o* *Unexpected sign (½)*	*(2)*
TIME	*Unsure of sign*	*H_0: β = 0* *H_a: β ≠ 0 (½)*	*+2* *(½)*	*Do not reject H_o* *2 < 2.048 (½)*	*(1½)*
		Critical t-value, 2-sided		*$t_{30.5\%,2\text{-}s}$ = 2.048*	*(1)*

Some notes

(1) The expected value of a coefficient is based purely on theoretical considerations. Do not be misled by the empirical value of the coefficient. For example, the expected coefficient of INC is positive, although its empirical estimate is negative (-0.02). In this case the coefficient of INC has an unexpected sign.

(2) The H_a specification of a coefficient always agrees with its (theoretical) expected value.

(3) The t-value is derived as $t = \frac{\widehat{\beta} - 0}{SE(\widehat{\beta})}$. Do not omit its sign.

(4) If a coefficient has an unexpected sign, then the decision is always not to reject H_0.

(5) The two-sided t-test is performed when one is unsure of the sign of the coefficient.

5.4 Limitations of the t-test

Please take note of all the limitations as mentioned in the textbook.

What is the difference between statistical significance and economic impact?

The t-test does not test the importance of a variable. If an estimated coefficient is statistically significant, this implies that it is unlikely that the true value of the coefficient is zero, taking account of random variation.

The importance of an X-variable may be measured by its economic impact.

Economic impact = Slope coefficient of X_k multiplied by average value of variable X_k.

LEARNING ACTIVITY 2

The following regression equation has been estimated (equation 5.4):

YIELD= -1.0 + 0.004(WATER) + 0.003(FERT)

where

YIELD**:** Yield of crop (t/ha: ton per hectare)

WATER: Rain & irrigation within growing season (mm); Average(WATER) = 500 mm

FERT**:** Application of fertiliser (kg/ha); Average(FERT) = 800 kg/ha

Prove that the economic impact of a variable

(1) confirms that FERT has a greater economic impact than WATER although its coefficient is smaller (equation 5.4).

(2) is not affected by its unit of measurement. Use the case where WATER is measured in cm rather than mm.

ANSWER

- *The economic impacts are derived as follows.*

Unit of measurement	*Coefficient* *Yield (t/ha) per unit of X*	*Economic impact =* *Coefficient of X . Average(X)*
FERT		
kg/ha	*0.003 t/ha per kg/ha of fertiliser*	*0.003 x 800 kg of fert. = 2.4 t/ha*
WATER		
mm	*0.004 t/ha per mm rain*	*0.004 x 500 mm of water = 2 t/ha*
cm	*0.04 t/ha per cm rain*	*0.04 x 50 cm of rain = 2 t/ha*

Judging from the coefficients of equation 5.4, it may appear that WATER has a greater impact on YIELD than FERT because the coefficient of WATER (0.004) is greater than the coefficient of FERT (0.003). This is, however, not true. The impact of FERT (2.4 t/ha) exceeds that of WATER (2 t/ha).

The unit of measurement of the X-variable affects the value of its coefficient.

- *The effect of an increase in 1mm of rain on yield is dYIELD / dWATER = 0.004 t/ha per 1 mm of WATER.*
- *If WATER is measured in cm (mm / 10) then the new coefficient dYIELD / dWATER is 10 x the previous coefficient: 0.04 t/ha per 1 cm of RAIN.*

However, the economic impact of WATER remains the same (2 t/ha), irrespective of the unit of measurement of WATER.

5.5 The f-test

In econometrics, the F-test is used to test the overall significance of a regression equation. See section 5.6.

Of the two tests, the t-test is used more often since it provides information on each of the slope coefficients.

LEARNING ACTIVITY 3

Test the overall significance of regression equation ONE (see section 5.6). The empirical F-value = 500 (as determined by the regression).

- specify H_0 and H_a
- look up the critical F-value and
- draw the appropriate conclusion

ANSWER

Set hypotheses:

H_0: $\beta_{PB} = \beta_{POTH} = \beta_{INC} = \beta_{TIME} = 0$

H_a: *H_0 is not true*

Critical F-value:

$F_{K,n\text{-}K\text{-}1\ 5\%} = F_{4,30,5\%} = 2.69$

Decision:

Reject H_0 because the actual F-value (500) > critical F-value.
Thus the equation is overall significant at the 5% level of significance.

(C) TRUE/FALSE QUESTIONS (F) = FALSE (T) = TRUE

(1) Student A fails the examination. In this particular case there may be the possibility of a type II error, but not a type I error. (T)

See table 5.1 in the study guide. If person A fails the examination, then there are only two possibilities. Either person A is failed correctly, or he/she may be failed incorrectly, in which case a type II error is made. A type I error can only occur when a person is passed.

(2) Student B passes the examination. This particular decision carries the risk of a type I error. (T)

See table 5.1 in the study guide. If the examination is passed, then there are only two possibilities. Person B is passed either correctly or incorrectly, in the last case a type I error is made.

(3) If the pass mark is increased to 65%, then the probability of a type II error is increased. (T)

A stricter critical value will increase the risk of failing capable persons (type II error), although the risk of a type I error (passing the incapable) is decreased.

(4) An improved examination system may decrease the occurrence of both types of errors. (T)

In the extreme case a perfect examination would have no errors at all!

(5) A 1% level of significance is better than a 5% level of significance. (F)

The advantage of a stricter critical value (1% level of significance) is that the risk of a type I error is reduced – at the cost, however, of increasing the risk of a type II error. Which of the two error types is considered to be more important, depends on the situation and ultimately on subjective judgement.

Assume we are testing prospective pilots. Because the consequences of a type I error (putting bad pilots in charge of aeroplanes) could be extremely costly (crashed aeroplanes and hundreds of people killed), it makes sense to try and minimise the type I error although we know it increases the probability of a type II error (failing good pilots).

(6) The t-value can be negative. (T)

Yes, the sign of the coefficient determines the sign of the t-value.

(7) The following hypothesis is tested:

H_0: $\beta \geq 0$

H_a: $\beta < 0$

The test result is as follows (n = 20, K = 2)

$$\hat{\beta} = +25.838 \ and \ SE(\hat{\beta}) = 9.582 .$$

We can reject the null hypothesis at the 99% confidence level. (F)

The sign of the coefficient is incorrect. It is positive and does not agree to H_a. Thus we cannot reject the null hypothesis. Formally the null hypothesis can be rejected if two conditions are met, viz

- the coefficient has the correct sign, as implied by the alternative hypothesis
- the absolute value of the calculated t-statistic exceeds the critical t-value

In this case the first condition is not met. In fact, when the sign of the coefficient does not agree to expectations, then we do not even have to test for the second condition!

(8) A variable which has a large economic impact is unlikely to be statistically insignificant. (F)

The statistical significance of a variable depends on the size of the standard error relative to that of its coefficient. The impact (or economic significance) of a variable X on variable Y is typically calculated as the slope coefficient multiplied by the average of X.

(9) The following joint hypothesis is tested:

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

H_a: H_0 is not true

The test results are as follows (n=30, K=4)

$$\sum e_i^{\ 2} = 250 \, and \sum \left(Y_i - \overline{Y}\right)^2 = 450$$

We can reject H_0 at the 1% level of significance. (T)

The F-value is calculated as

$$ESS/K \Big/ RSS/(n-K-1) = 200/4 \Big/ 250/(30-4-1) = 5$$

where ESS = TSS − RSS = 450 − 250 = 200

Since 5 exceeds the critical $F_{4,\ 25,\ 1\%}$ = 4.18 the null hypothesis can be rejected.

(D) PRACTICAL EXERCISES

(1) What is the meaning of a 5% level of significance when we are testing the case of

H_0: β = 0?

H_a: β ≠ 0?

In this case:

- We assume that H_0: β = 0 is true.
- We use the distribution of

$$t = \frac{\hat{\beta} \ - 0}{SE(\hat{\beta})}$$

to derive critical values such that the probability of obtaining a t-value outside this range, due to random chance, is 5%.

- Since this is a two-sided test, and because the t-distribution is symmetrical, $t_{critical}$ is selected so that the probability that t > $t_{critical}$ is 2.5%. The probability that t < -$t_{critical}$ is also 2.5%. Adding the two gives 5%.
- The meaning of a 5% level of significance is that there is a probability of a type I error: incorrectly rejecting H_0 when H_0 is in fact correct.

(2) The South African annual production of apricots is as follows:

Year	Production 1000 tons
1990	48
1991	56
1992	57
1993	54
1994	58
1995	67
1996	79

2.1 Estimate the regression equation: Q = a + bT where Q: production and T: year by OLS. Use your spreadsheet to confirm that:

$\hat{Q}_i$ = -8196.86 + 4.142857(T_i) where T_i: 1990, 1991, ... and interpret the results.

The slope coefficient (b) implies that the production of apricots increased on average by 4 143 tons per year.

The constant term (a) lumps together all the factors which affect the production of apricots, excluding time. Note that the interpretation: the production of apricots for T = 0, that is, for the year 1900, is economically inappropriate (although technically correct) since T = 0 falls beyond the range of observations T = 1990 to 1996.

2.2 Derive the standard error of the slope coefficient

SE($\hat{b}$) $= 0.993859$ (Use your spreadsheet and OLS to confirm this result)

2.3 Derive and interpret R^2

R^2 = 0.776 which indicates a good fit. Since R^2 = ExplainedSS / TotalSS, 77.6% of the variation in production is explained by time.

2.4 Test whether the slope coefficient is significantly positive.

State the null and alternative hypotheses:

H_0: $b \leq 0$

H_a: $b > 0$

Calculate the t-statistic

$$t = \frac{4.142857}{0.993859} = 4.1685$$

Look up the critical t-value (df = n – K – 1 = 5, use a 5% level of significance)

$t_{5,\ one\text{-}sided,5\%}$ = 2.015

Conclusion

Since the actual t-value exceeds the critical t-value, and the sign of the coefficient is correct, we can reject H_0. Thus the slope coefficient is significantly greater than zero.

(E) EXAMINATION: PARAGRAPH QUESTIONS

(1) Explain the goal and method of hypothesis testing and assess its risks. In your answer refer to the null and alternative hypotheses, the critical value, the meaning of type I and type II errors and the level of significance. (12)

(2) Explain (12)

- the meaning of a type I and type II error in hypothesis testing.
- the impact of using a relatively higher or lower critical value on the occurrence of the two types of errors.
- whether a 5% level of significance is necessarily better than a 1% level of significance.
- the difference between the statistical significance and the importance of variables.

(F) EXAMINATION: PRACTICAL QUESTIONS

You should know

- when to use one-sided and two sided t-tests and how to apply them

In the examination you may be tested on your ability to

- set appropriate H_0 and H_a hypotheses based on theoretical considerations
- calculate t-values or F-values
- look up critical t-values or F-values in statistical tables (the tables will be provided in the examination)
- draw appropriate conclusions

ECONOMICS IN ACTION (LOOKING BACK)

Deirdre McCloskey mentioned at the beginning of this learning unit, refers to the difference between the statistical significance of the coefficient of a variable and the economic impact of a variable.

McCloskey uses the term "noise" to refer to the accuracy of estimating coefficients in the presence of random variation. Coefficients of a regression equation cannot be measured 100% accurately due to **sampling error**. Sampling error depends, amongst others, on the size of the error terms. If the error terms are relatively small then the $SE(\hat{\beta})$ will also be small and the estimates of $\hat{\beta}$ are comparatively "noise free".

Substantive significance refers to the **economic impact** or importance of a variable. One way to measure it is to derive $\hat{\beta}_k.\overline{X}_k$. The economic impact also allows us to compare the economic effects of different variables. See learning activity 5.4.

PART III

SPECIFICATION

LEARNING UNIT 6

CHOOSING THE INDEPENDENT VARIABLES

ECONOMETRICS IN ACTION

Hood and Koopmans are two well known econometricians from the early period of the development of econometrics. They[5] wrote the following:

> *In the application of statistical methods to economics two broad problems of economic analysis must be faced. The first is that the scope for experimentation is limited. Broadly speaking, economic history can only be observed as it is lived, uninfluenced for purposes of scientific inquiry. The second is that analysis must seek to answer questions concerning the effects of specific policies of governments, private firms, or individuals. From the studies assembled here it appears that, under these circumstances, the application of statistical methods to a given set of observations must lean heavily on preconceptions as to the nature and persistence of behaviour relationships.*

Are preconceptions regarding economic behaviour really that important in econometric specification? Is econometrics not meant to be scientific and objective?

STUDY OBJECTIVES

When you have studied this learning unit you should understand

- the meaning of specification error
- the impact of an omitted variable
- the impact of an irrelevant variable
- what is wrong with specification searches

(A) PRESCRIBED MATERIAL

The next two chapters deal with the problem of specification. Specifying a regression equation involves two issues. This chapter deals with selecting the appropriate independent variables. The next chapter deals with the appropriate functional form.

The following sections are prescribed:

(1) Omitted variables
(2) Irrelevant variables
(3) Specification searches

[5] Hood. WC & Koopmans, TC (eds). *Preface to Studies in Econometric Method*. John Wiley & Sons. Cowles Commission Monograph no. 14

We recommend that you study the following sections in parallel with the prescribed sections:

(4) An illustration of the misuse of specification criteria
(5) An example of choosing independent variables

Section 7 Appendix: Additional specification criteria are not prescribed.

(B) SOME IMPORTANT CONCEPTS

The problem of specification

We suggest you revise section 1 of this study guide which deals with the role of economic theory in specification. Preconceptions (or assumptions) do play an important role in specification. We refer to the example in section 1 which deals with the monetarist and the post-Keynesian approaches to monetary theory. Econometrics is not purely objective because specification will always remain subjective. Ultimately, economic theory dictates the causality, the choice of variables included in the regression equation, and the functional form.

Econometrics cannot tell the researcher how to specify an equation. It can also never prove that a model is correct. Econometrics can only estimate a given equation.

This chapter deals with the problem of specifying the appropriate variables in an equation. Since specification is really an economic issue, this chapter looks at incorrect specification from a statistical point of view. Two problems are addressed:

- excluding a variable which should have been included
- including a variable which should have been excluded

Both theoretical and empirical considerations should be used in specification. Four criteria can be used to support the inclusion/exclusion of a variable:

- *Theory:* Is the variable's place in the equation unambiguous and theoretically sound? This course does not deal with economic theory but draws on what you have already learned in micro- and macroeconomics.
- *T-test:* Is the variable's estimated coefficient significantly different from zero? This is dealt with in chapter 5.
- $\overline{R}^2$: Does the overall fit of the equation (adjusted for degrees of freedom) improve when the variable is added to the equation? This is dealt with in chapter 2.
- *Bias:* Do the signs of the other variables change much when the variable is added to the equation?

6.1 Omitted variables

The approach followed in section 1 of the textbook (p. 178) is to assume a correct specification which includes more than one X, and to derive the consequences when one of these is omitted from the equation. In general this leads to biased estimates of the remaining (included) coefficients of the remaining (included) variables.

The general expression to determine the direction of bias of an included coefficient is

$$E\left(\hat{\beta}_{incl} - \beta_{incl}\right) \quad = \beta_{omit} \cdot f\,(r_{X_{incl},X_{omit}}) \qquad \textit{(equation 6)}$$

where β_{incl} refers to the coefficient of any one of the coefficients of the remaining X_{incl} variables and β_{omit} refers to the true coefficient of the omitted variable. The correlation coefficient refers to the correlation between the included variable X_{incl} and the omitted variable: X_{omit}. The function, f(r) indicates that it is really a function of r that affects bias. For practical purposes, however, all that counts is the sign of r.

This formula applies only to the case where the omitted variable is not the constant term. In the case of the constant term, $f(r_{X\ omit,\ X\ incl}) = 0$. Secondly, the formula only predicts the direction of bias in terms of β_{omit} and $r(X_{incl}, X_{omit})$, that is, whether these two terms are positive, negative or zero.

You should know equation 6. It will not appear in the formula sheet. You must be able to determine the direction of bias on any of the remaining coefficients when any variable is omitted from a regression equation.

LEARNING ACTIVITY 1

Please confirm or refute statements 1 to 5. Statements 1 to 4 require the application of equation 6.1.

A correctly specified equation for household consumption of good Q is:

$$Q_i = \beta_0 + \beta_P P_i + \beta_{INC} INC_i + \beta_{PZ} PZ_i + \varepsilon_i \qquad \textit{(equation 6.1)}$$

where

Q_i: Quantity demanded of good Q by household i

P_i: price of good Q

PZ_i: Price of good Z (a substitute for good Q)

INC_i: Per capita income of household i

$r_{P,INC} > 0$, $r_{P,PZ} < 0$, $r_{INC,PZ} = 0$

Statements to be evaluated		*$Q_i = \beta_0 + \beta_P P_i + \beta_{INC} INC_i + \beta_{PZ} PZ_i + \varepsilon_i$ where the expected values are: $\beta_P < 0$, $\beta_{INC} > 0$, $\beta_{PZ} > 0$.* *$r_{P,INC} > 0$, $r_{P,PZ} < 0$, $r_{INC,PZ} = 0$* *Statements 1–4* *use* $E(\hat{\beta}_{incl} - \beta_{incl}) = \beta_{omit} \cdot f\,(r_{X_{incl},X_{omit}})$
1	If INC is omitted from equation 6.1, then coefficient β_P is likely to be overestimated. (T)	*X_{incl} is P; $\beta_{omit} = \beta_{INC} > 0$; $r_{P,INC} > 0$* $E(\hat{\beta}_P - \beta_P) = \ \beta_{INC} \cdot f\,(r_{P,INC})$ = (+).(+) = (+) *Statement 1 is correct.*

2	If INC is omitted from equation 6.1, then coefficient β_{PZ} will not be affected (no bias). (T)	$X_{incl} = PZ; \beta_{omit} = \beta_{INC} > 0; r_{P,PZ} < 0$ $E(\hat{\beta}_{PZ} - \beta_{PZ}) = \beta_{INC} \cdot f(r_{INC,PZ})$ = (+).(0) = (0) *Statement 2 is correct.*
3	If PZ is omitted from equation 6.1, then coefficient β_P will be underestimated (negative bias). (T)	X_{incl} *is P;* $\beta_{omit} = \beta_{PZ} > 0; r_{P,PZ} < 0$ $E(\hat{\beta}_P - \beta_P) = \beta_{PZ} . f(r_{P,PZ})$ = (+).(-) = (-) *Statement 3 is correct.*
4	If PZ is omitted from equation 6.1, then coefficient β_{INC} will not be affected (no bias). (T)	$X_{incl} = INC; \beta_{omit} = \beta_{PZ} > 0; r_{INC,PZ} = 0$ $E(\hat{\beta}_{INC} - \beta_{INC}) = \beta_{PZ} . f(r_{INC,PZ})$ = (+).(0) = (0) *Statement 4 is correct.*
5	If INC is omitted from the equation, then R^2 is likely to decrease. (T)	*Since the addition of a variable to a regression equation will never decrease R^2 (but will most likely increase it), it follows that when any variable is omitted, the fit will probably be worse and R^2 is likely to decrease.*

6.2 Irrelevant variables

Irrelevant variables do not cause bias, but decrease the precision of estimators. Assume a model is specified as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

and that X_1 is irrelevant, that is, $\beta_1=0$. The expected bias of the remaining coefficient X_3 is

$$E(\hat{\beta}_3 - \beta_3) = \beta_1 . f(r_{X_3,X_1}) .$$

It follows that since $\beta_1 = 0$, there is no systematic bias in estimating β_3. Because $\beta_1 = 0$, there will also be no bias in estimating β_2.

Standard error of the estimated coefficient

The expression for the standard error, in the case of a two-variable model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

is specified as (also see the formula sheet which will be provided in the examination):

$$SE(\hat{\beta}_1) = \sqrt{\frac{\sum e_i^2 / (n-3)}{\left(\sum x_1^2\right)\left(1 - r_{12}^2\right)}}$$

Assume that X_2 is an irrelevant variable. If $r_{12} = 0$ then the SE is not affected. However, any value of $r_{12} \neq 0$ reduces the denominator term and increases the $SE(\hat{\beta}_1)$. Thus the inclusion of irrelevant variables reduces the precision of $\hat{\beta}_1$ if $r_{12} \neq 0$.

6.3 Specification searches

One of the weaknesses of econometrics is that a researcher can manipulate a data set to produce almost any desired result or, stated differently, "if the data are tortured long enough, it will certainly confess". Because of advances in computational ability this practice (called data mining) can easily be misused. Some degree of data mining is, however, almost always present in practical research.

Section 4, (p. 191): A specification search describes the most common data-mining techniques which attempt to estimate the best form from a number of combinations. These techniques include stepwise regression and sequential specification searches.

However, the practice of data mining has undesirable consequences. For one thing, it is likely to cause exaggerated claims of significance, since data mining increases the probability of a type I error. A significance level of, say, 5% becomes meaningless and is actually a much higher figure of, say, 15%. Data mining thus increases the probability of finding coefficients which are significantly different from zero, when in reality they are zero.

The textbook offers an intuitive explanation. Every regression run contains a risk of, say, 5% that a regression coefficient which appears significant occurred by chance rather than by truth. Consequently, multiple regression runs compound this error.

Estimation and hypothesis testing procedures are valid only when *a priori* considerations, instead of exploratory data-mining techniques, are used. In fact, every new regression run should in theory be run on new data!

In practice it is best to focus strongly on theoretical considerations when selecting variables (or functional forms), and to keep the number of regression runs as low as possible.

LEARNING ACTIVITY 2

Answer problem 2 at the end of chapter 6 (pp. 200 to 201 in the prescribed textbook). Some hints:

- Clearly state the H_0 and H_a in (a).

- To answer (b), the economic impact of each of the variables will have to be derived, amongst other things. Assume that an average American car costs \$20 000 and weighs 4 000 pounds (= 40 units of W).

ANSWER

The model which explains car prices is:

$$\hat{P}_i = 3.0 + 0.28W_i + 1.2T_i + 5.8C_i + 0.19L_i$$

where

P_i : list price of USA car i (1996) in \$1000
W_i: weight of ith car(hundreds of pounds)
T_i: transmission where automatic = 1 and manual = 0
C_i: cruise control dummy (device which maintains constant speed). $C_i = 1$ if cruise control is present and $C_i = 0$ when it is absent.
L_i: engine size (litres)

(a) Testing the coefficients

Coefficient	***β_W*** ***Weight***	***β_T*** ***Trans-mission***	***β_C*** ***Cruise control***	***β_L*** ***Engine size***
Expected sign	***Positive***	***Positive***	***Positive***	***Positive***
H_0	$\beta_W \leq 0$	$\beta_T \leq 0$	$\beta_C \leq 0$	$\beta_L \leq 0$
H_a	$\beta_W > 0$	$\beta_T > 0$	$\beta_C > 0$	$\beta_L > 0$
t-value	*4.0*	*3.0*	*2.0*	*0.95*
Decision	*Reject H_o*	*Reject H_o*	*Reject H_o*	*Do not reject H_o*

The critical t-value; $t_{5\%,\text{one-sided},30\text{ degrees of freedom}} = 1.697$

(b) The following econometric problems appear to exist:

- The engine size variable (L) appears redundant, since its t-value denotes an insignificant coefficient. The effect of the engine size is probably already included in W.
- The coefficient of C appears far too large. The presence of cruise control adds \$5 800 to the price of the car which is, say, 25% of its price! C appears to act as a proxy for other luxury options, or for the general quality of the car.

That the coefficient of C appears far too large can be deduced by deriving the economic impact of each variable on price. In the case of a 4 000 pound car which has automatic transmission and cruise control and a four-litre engine, the expected price will be:

$3.0 + 0.28W_i + 1.2T_i + 5.8C_i + 0.2L$
$= 3.0 + 0.28(40) + 1.2 + 5.8 + 0.2(4)$
$= 3.0 + 11.2 + 1.2 + 5.8 + 0.8$
$= 22$ which represents $22000.

(c) *Model T from which variable L has been dropped is:*

- $\hat{P}_i = 18.0 + 0.29W_i + 1.2T_i + 5.9Ci$

  The new model should be judged on the basis of the four standard criteria.

- *Theory:* The effect of the size of the car is probably already included in variable W. The empirical results confirm that the impact of engine size is relatively small.
- *T-values of coefficients:* In the new model, the t-scores denote that all the respective coefficients are significantly different from zero. In fact, in the case of W and T, the t-scores have even improved slightly.
- $\overline{R}^2$ has not changed. Thus L contributes nothing in terms of a better fit.
- *Bias:* Note that the coefficients of the included variables show very little change compared with the old model.

  Thus, the new model appears better than the old one.

(C) TRUE/FALSE QUESTIONS (F) = false (T) = true

(1) Consider YIELD = -1.0 + 0.004(WATER) + 0.003(FERT) used in learning activity 5.4 (learning unit 5). If FERT is incorrectly omitted then $\hat{\beta}_{WATER}$ will be upward biased to compensate for the effect of fertilizer. Assume that $r_{WATER,FERT} > 0$. (T)

$E(\hat{\beta}_{WATER} - \beta_{WATER}) = \beta_{FERT} . f(r_{WATER,FERT})$. The expected value of $\beta_{FERT} > 0$ and $r_{WATER,FERT} > 0$. Thus $\hat{\beta}_{WATER}$ will be upward biased.

(2) The most important criterion for the selection of a variable in an equation is theory. (T)

(D) EXAMINATION: PARAGRAPH QUESTIONS

(1) Discuss which criteria in general should be used to select the independent variables in an equation. (8)

(2) Discuss the problems inherent in data mining and specification searches. (6)

(3) Discuss the problem of selecting the variables of an equation. Pay attention to (10)

- the way in which an omitted variable (both incorrectly and correctly omitted) may lead to specification bias (6)
- the impact of an irrelevant variable (4)

(E) EXAMINATION: PRACTICAL QUESTIONS

Practical situations may be provided in the examination which require of you to infer the impact of incorrectly excluding a variable which should have been included; and/or incorrectly including a variable which should have been excluded. You should **know** and be able to apply the **formula for bias**. Note that **this formula is not included** in the formulas sheet.

You should also be able to apply the four criteria which can be used to support the inclusion/ exclusion of a variable in an equation.

LEARNING UNIT 7

CHOOSING A FUNCTIONAL FORM

ECONOMETRICS IN ACTION

Some researchers[6] prefer a linear specification:

> *A potentially important issue is the appropriate functional form for the equations. A standard approach is to transform the variables into logarithms. This transformation allows the coefficients to be interpreted directly as elasticities and can also take account of non-linearities in the variables. However, it does have disadvantages. Specifically, using a logarithmic form imposes the restriction that elasticity is constant along the demand curve, which is unfortunate since it is reasonable to expect that demand will be more price-sensitive at higher prices (and lower quantities). For this reason, we prefer to use a linear specification.*

while others[7] favour a non-linear specification:

> *Most econometric models assume a linear relationship among variables and this limitation can be one reason for the poor forecasting performance of these models.*

It's time to consider the various functional forms. When should we use logarithmic transformations? When do we use the other forms? How can dummy variables be used?

(A) PRESCRIBED MATERIAL

The following sections are prescribed:

(1) The use and interpretation of the constant term
(2) Alternative functional forms
(3) Lagged independent variables
(4) Using dummy variables
(5) Slope dummy variables
(6) Problems with incorrect functional forms

[6] Paton, D, Siegel, DS & Williams, LV. A Time Series Analysis of the Demand for Gambling in the United Kingdom*. NUBS Working Paper,* No. 2001.II. Nottingham: Nottingham University Business School, 2001a

[7] Boero, G & Cavalli, E. 1996. Forecasting the exchange rate: A comparison between econometrics and neural network models. *AFIR,* Vol II, pp. 981–996.

STUDY OBJECTIVES

When you have studied this learning unit you should

- understand the meaning and use of the constant term
- understand the basic functional forms dealt with in this chapter, their characteristics and when to use them
- understand how to use dummy variables, that is, both intercept dummies and slope dummies
- acknowledge the problems that may arise when using functional forms in certain ways

(B) SOME IMPORTANT CONCEPTS

7.1 The use and interpretation of the constant term

The constant term (for example in $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$) includes at least 3 components:

- True β_0
- Mean effect of minor (excluded) Xs
- The mean of ε (if not zero)

The interpretation that the constant term is the value of the dependent variable when all the Xs are zero, is theoretically correct but should in practice be used with care, because the value of the Xs = 0 often lies outside the range of observation of the Xs.

The constant term is always included unless there is a sound theoretical reason for not doing so. Note the rationale for including the constant term. What happens when the constant term is suppressed? When the constant term is suppressed, the estimated regression equation necessarily passes through the origin. Make sure you understand the problems which arise when the constant term is suppressed.

Although the constant term may be tested for statistical significance, this is seldom done. Testing the constant term opens up the possibility of rejecting the constant term and this is precisely what should be avoided.

7.2 Alternative functional forms

(a) Slope and elasticity

In the case of the linear form $Y = \beta_0 + \beta_1 X$, the slope is constant. The slope is not constant for non-linear forms.	$Slope = \frac{\Delta Y}{\Delta X} = \beta_1$
The elasticity is another measure which may be used to describe a functional form Y = f(X). Elasticity is the **percentage increase in Y** divided by **the percentage increase in X.**	$e = \frac{\Delta Y / Y}{\Delta X \,/\, X} = \frac{\Delta Y}{\Delta X} \bullet \frac{X}{Y}$

(b) Functions, differences and derivatives

Because the slope and elasticity of a function are often expressed in mathematical terms, we provide some notes on functions, differences and derivatives.
You must understand

- functional notation, for example the meaning of $Y = f(X)$, $Y = f(X, Z)$ or $Y = f(X_1, X_2, X_3)$
- the concept of the slope of function $Y = f(X)$, denoted by $\Delta Y/\Delta X$
- the concept of elasticity

We recommend that you use mathematical notation when dealing with the functional forms. In this module knowledge of calculus is not required, although it will help if you do understand calculus.

Functions

- In a single variable function $Y = f(X)$ the one variable X affects variable Y. Examples are $Y = 3 + 2X$, $Y = 2 + 3X^2$ or $Y = 40 + \log(X)$.
- Econometrics mostly uses multivariate functions. Assume the quantity of meat demanded (Q_m) is affected by the price of meat (P_m), the price of chicken (P_c) and per capita income (Y/N). Then the general form of the demand function may be written as

$$Q_m = f(P_m, P_c, Y/N).$$

- Of course, you would estimate its specific form, for example

$$Q_m = 13.5 - 0.78(P_m) + 1.2(P_c) + 3.1(\log(Y/N)).$$

Differences

- The delta operator (Δ) denotes a difference. If variable X increases from say 4.1 to 4.2 then $\Delta X = 0.1$
- The slope of a function $Y = f(X)$ may be expressed in difference form, that is, as $\Delta Y/\Delta X$ if ΔX is sufficiently small.

Derivatives

- Studenmund consistently uses "delta" notation (Δ) to denote the slope ($\Delta Y/\Delta X$).
- For the mathematically minded, we may also use calculus notation, that is, the derivative (d) or partial derivative (∂).
- The derivative of $Y = f(X)$ at the point X is denoted as dY/dX. In mathematical terms:

$$\frac{dY}{dX} = \underset{\Delta X \to 0}{Lim} \left\{ \frac{f(X + \Delta X)}{\Delta X} \right\}$$

- The partial derivative (∂Y/∂X) is used in the case of a multivariate function, for example Y = f(X,Z,Q). It denotes the slope of a function (ΔY/ΔX), that is, with Z and Q held constant.

  The difference form of the slope (ΔY/ΔX) may be used to approximate both dY/dX and ∂Y/∂X if ΔX is sufficiently small. Studenmund uses ΔY/ΔX notation for all these cases. We will do the same.

(c) Log transformations

A log transformation is one of the most widely used transformations in Y = f(X), used either on the Y side of the regression equation, or on the X side (for one or more Xs), or on both sides. We expect of you to understand the meaning of logs. See pp. 226 to 227 in the prescribed book for a brief explanation.

The textbook and this document use natural logs (base e = 2.718282) and not logarithms to base 10. The number e is a mathematical constant and is different from e, the error term. In all the examples in this document log(X) implies ln(X) where ln(X) refers to the natural logarithm.

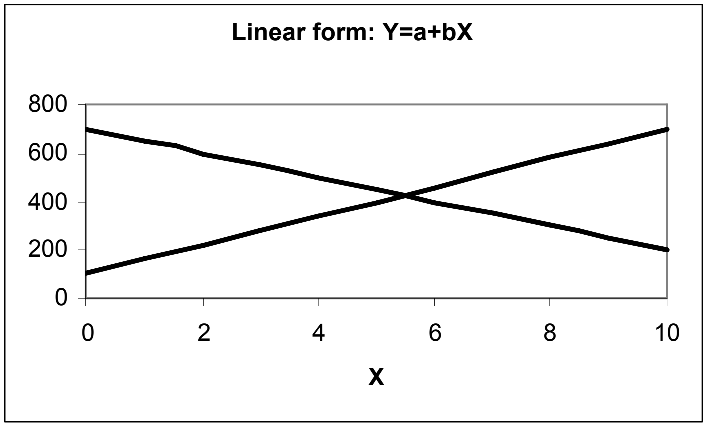
Because $e^{2.721295}$ = 15.2 then log(15.2) = 2.721295 and exp(2.721295) = 15.2	Within Excel the following functions apply: = ln(15.2) = 2.721295 = exp(2.721295) = 15.2

Table 1 (p. 234 of prescribed book) summarises six functional forms which are commonly used. You must know them all. In the examination we expect of you to know

- their respective equations
- their graphical representations
- their characteristic, for example in the case of log(Y) = a + bX, then (ΔY/Y)/ΔX is constant
- their expressions for the slope and elasticity (which may be easily derived from their characteristic)
- their use and/or advantages in practical situations

7.2.1 Linear form

The linear form $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_KX_K\epsilon$ is the default form. It assumes a linear relation between Y and X_k, that is, $\Delta Y/\Delta X_k = \beta_k$ for k = 1 to K. Graphically, this form is represented by a straight line.



It is the default form unless we know more of the relationship between Y and X. In economics, many relationships between variables are known to be non-linear on theoretical grounds, hence several non-linear forms are used. These are discussed below.

7.2.2 Double-log form: log Y = $\beta_0 + \beta_1 \log(X_1) + \beta_2 \log(X_2) + \log(\varepsilon)$

This form is also called the Cobb-Douglas specification, in which case Y measures output and X_1 and X_2 measure factors of production, usually capital and labour. This form is the equivalent of $Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \varepsilon$ and is linearised by taking logs both sides. The error term, ε in this last form, varies around the value of one.

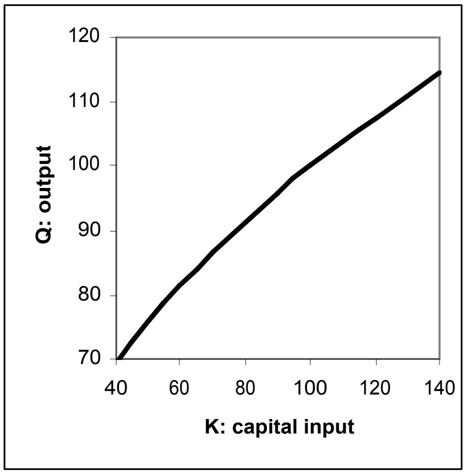
Its main feature is that it has a constant elasticity[8], that is

$$\frac{\%\,change\,in\,Y}{\%\,change\,in\,X} = \frac{\Delta Y / Y}{\Delta X / X} = \beta_k$$

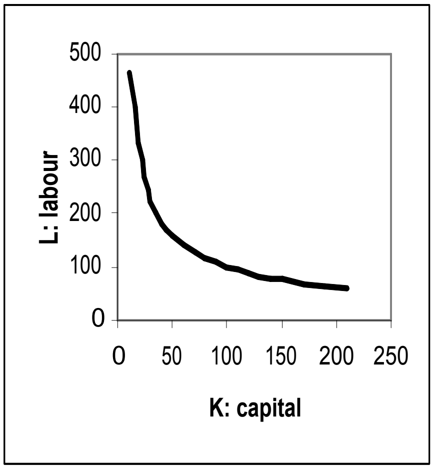
Besides its use as a simple production function (for one of K or L held constant), it may also be used to depict an isoquant (for fixed Y), that is, the substitution possibilities between factors of production.

[8] If Y = a.X^b then, using calculus, ΔY/ΔX = a.b.X^(b-1) = (Y/X^b).b.X^(b-1) = b.Y/X. Then (ΔY/Y)/(ΔX/X) = b.

Two applications of the case $Q = K^{0.4}L^{0.6}$ are demonstrated below.



Production function: Q = f(K) with L constant



Isoquant: Relationship between K and L with Q = 100

With respect to Q = f(K) with L constant:

- The slope ΔQ/ΔK is not constant but decreasing. (Only if the coefficient of K is less than one.)
- The ratio (ΔQ/Q)/(ΔK/K) is constant and equal to the coefficient of K = 0.4. This can be interpreted as the percentage increase in Q per 1% increase in K.
- This form cannot model a more realistic production function which first displays an increasing average product and then a decreasing one.

LEARNING ACTIVITY 1

The table below provides volume indexes of production (Q), and of its capital (K) and labour (L) inputs. Estimate a Cobb-Douglas type of production function and interpret its coefficients

Q	K	L
100.0	100	100
102.6	103	102
104.2	105	103
105.2	106	104
108.0	110	105

ANSWER

The Cobb-Douglas production function is $Q = aK^bL^c$. To estimate this function, we linearise this form first: log(Q) = log(a) + b.log(K) + c.log(L)

log(Q)	*log(K)*	*log(L)*
4.6052	*4.6052*	*4.6052*
4.6308	*4.6347*	*4.6250*
4.6463	*4.6540*	*4.6347*
4.6559	*4.6634*	*4.6444*
4.6821	*4.7005*	*4.6540*

Its estimate is (use the OLS method and your spreadsheet)

$$\log(Q) = 0.016272 + 0.609334[\log(K)] + 0.387131[\log(L)]$$

and its original form is thus

$$Q = 1.0164 \times K^{0.609334} \times L^{0.387131}$$

The slope coefficients are elasticities of production with respect to their respective inputs.

$$\frac{\Delta Q / Q}{\Delta K / K} = 0.609 \ \ and \ \ \frac{\Delta Q / Q}{\Delta L / L} = 0.387$$

If, for example, capital (K) increases by 1%, then output (Q) will increase by 0.609%.

7.2.3 Semilog form

(a) Semi-log form: log(Y) = log(a) + log(b).X

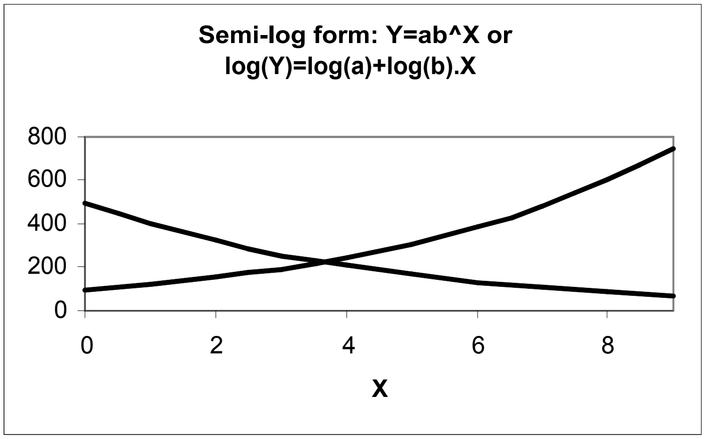
This form[9] is derived from $Y = ab^X$. If logs are taken both sides, we obtain log(Y) = log(a) + log(b).X

This form is used when Y increases at a geometric rate with respect to X, that is, when (ΔY/Y)/ΔX is constant[10].

The graph demonstrates two cases of this semi-log form.

[9] We use the power function in this case. For example 2 to the power 3, that is, 2 x 2 x 2 is denoted as 2^3 or as 23.

[10] This is easy to prove using calculus. If Y = a.b^X then ΔY/ΔX = Y.log(b). Thus, (ΔY/ΔX).(1/Y) = log(b) and (ΔY/Y)/ΔX = log(b).



Negative

slope Y = 100(0.80)^X

(b) The meaning of a geometric increase

Many Y = f(X) phenomena in economics change at a geometric rate where (ΔY/Y)/ΔX is constant. In most of these X represents time.

- Populations typically grow at a geometric rate per year. Population growth is almost always expressed at a specific rate, say 2.7% per year.
- Many time series of real spending (at constant prices) or of the volume of production grow at a geometric rate over time (e.g. increasing by an average of 4.5% per year), rather than at a linear rate.
- Price levels tend to change at a geometric rather than a linear rate with respect to time. The average inflation rate is expressed as, say, 6.5% per year.

LEARNING ACTIVITY 2

Consider the following time-series data:

Year (X)	Y
1	100.00
2	125.00
3	156.25

Calculate for the years X = 2 and X = 3:

(a) The differences: ΔX and ΔY
(b) the slope ΔY/ΔX
(c) the change in Y: ΔY/Y and (ΔY/Y)/ΔX
(d) the elasticity (ΔY/Y)/(ΔX/X).

ANSWER

X	*Y*	*(a)* *ΔX*	*ΔY*	*(b)* *ΔY/ΔX*	*(c)* *ΔY/Y which in this case = (ΔY/Y)/ΔX*	*(d)* *E = (ΔY/ΔX).(X/Y)*
1	*100.00*					
2	*125.00*	*1*	*25.00*	*25*	*0.25 (= 25/100)*	*0.25 = 25 x (1/100)*
3	*156.25*	*1*	*31.25*	*31.25*	*0.3125 (= 31.25/125)*	*0.50 = 31.25 x (2/125)*

The table above demonstrates three characteristics of this function. The (1) slope and (2) the elasticity increase as X increases, and (3) Y increases geometrically.

Y increases geometrically (ΔY/Y is constant) when X increases by 1 (ΔX = 1). Because (ΔY/Y)/ΔX = 0.25, this means that Y increases by a constant percentage (25%) per unit increase in X.

In the table above, when calculation ΔY/Y and e = (ΔY/ΔX).(X/Y), we have used the previous observation's X and Y values, and not those of the current observation. However, the current observation's values of (X,Y) may also be used, which will change the values of ΔY/Y and e = (ΔY/ΔX).(X/Y). The difference arises because, in the table above, ΔX = 1 is relatively large. In calculus, where we work with infinitesimal small changes of ΔX and ΔY, there would be no such difference.

LEARNING ACTIVITY 3

You are required to build a model that explains population size in terms of time. The population (POP) is expected to increase, on average, at a constant percentage per year (T). Explain which functional form is appropriate in this case.

ANSWER

The appropriate form is POP = $a.b^T$ where T: year, for example, T = 90, 91, 92, ... If, say, the percentage increase per year is 2.7%, then POP = $a(1.027)^T$. This form is equivalent to the form: log(POP) = log(a) + T.log(b).

LEARNING ACTIVITY 7.2.3.3

A regression has been estimated as follows:

log(Y) = 2.721295 – 0.077962(T)

where T = 0, 1, 2 ... represents time and log(10) = 2.302585.

Interpret this result.

ANSWER

First convert the log form to the form $Y = ab^T$

$$Y = \exp(2.721295) \times [\exp(-0.077962)]^T$$

$$Y = 15.2 \times 0.925^T$$

Log(10) = 2.302585 confirms that natural logs were used because $\log_{10}(10) = 1$.

Interpretation:

Constant term: If T = 0 then Y=15.2

Slope: $Y_t = Y_{t-1} \times 0.925$, that is, Y decreases by 7.5% per unit increase in T.

LEARNING ACTIVITY 4

The South African production of good wine in bulk is as follows:

Year	T	Q	log(Q)
1991	1	396	5.9814
1992	2	427	6.0568
1993	3	395	5.9789
1994	4	421	6.0426
1995	5	498	6.2106
1996	6	577	6.3578
1997	7	547	6.3044
1998	8	544	6.2989
1999	9	596	6.3902
2000	10	552	6.3135

Where
T: year;
Q: production of good wine in bulk measured in 1 000 million litres log(Q) is the natural logarithm of Q.

(a) Confirm that the production of good wine increased on average by 4.9% per year over the 1991 to 2000.

(b) Derive a forecast of the production of good wine for the year 2005, based on the previous result.

ANSWERS

To derive an average annual growth rate, we need to use the functional form $Q = ab^T$ where Q:production and T:year (1, 2 ...). Its linearised form: log(Q) = log(a) + log(b). T has to be estimated, which implies that log(Q) – the Y-variable – must be regressed against T – the X-variable. We use OLS to estimate the coefficients log(a) and log(b).

The estimated coefficients using OLS are:

$$log(Q) = 5.931654 + 0.047615(T) \qquad R^2 = 0.791$$

this designates the original form

$$Q = 376.7773(1.048766)^T$$

because, for example, antilog(0.047615) is derived as

$$e^{0.047615} = 1.048766$$

This confirms that the production of good wine increased on average by 4.9% per year over the sample period.

Forecast:

Log(Q_{2005}) = 5.931654 + 0.047615(15) = 6.6459. Thus Q = $e^{6.6459}$ = 769.6.

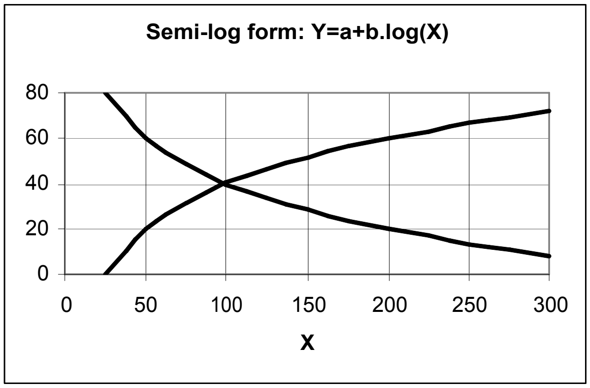
The characteristic of a function and its method of estimation

The functional form Q = ab^T (or its linearised form: log(Q) = log(a) + log(b).T) has the characteristic that (ΔQ/Q)/ΔT = log(b). This may appear like a method to estimate log(b), and consequently that of b, but it is not. The characteristic only applies to points on the curve $(\hat{Q},T)$ after the coefficients have been estimated. It does not apply to observed (Q,T) data. Perhaps it is better to say $\frac{\Delta\hat{Q}}{\hat{Q}}.\frac{1}{\Delta T} = \log(b)$. However, because we are implying points on the curve when dealing with the characteristic of a functional form, the "hat" is usually omitted.

To estimate form log(Q) = log(a) + log(b). T, we use OLS. We first transform Q to log(Q) and then use (log(Q), T) as input data. OLS then provides estimates of log(a) and log(b).

(c) Semi-log form: Y = β_0 + β_1 log (X)

This form is often used in the case of cross-sectional data. Consider, for example, the relationship between consumption (Y) and income (X) of households. Typically, consumption increases with income, but at a decreasing rate (Engel curves). See for example the graphs below.



Positive sloping curve:

$$Y = -92.9 + 28.9\log(X)$$

Negative sloping curve:

$$Y = 172.9 - 28.9\log(X)$$

Consider the positive sloping curve. The relationship between Y and X is such that X increases geometrically while Y increases linearly.

- Assume the starting point X = 50 and Y = 20.
- If X doubles to 100 (X increases geometrically), then Y increases by 20 to Y = 40.
- If X doubles again to 200, then Y increases by another 20 to Y = 60.

The condition which describes this behaviour is that $\Delta Y/(\Delta X/X)$ is constant.[11]

It is perfectly in order have a mixed specification such as

$$Y = \beta_0 + \beta_1\log(X_1) + \beta_2 X_2 + ... + \beta_K X_K$$

where the log is applied to only some of the Xs. In this case the relationship between Y and X_1 is such that:

$$\Delta Y/(\Delta X_1/X_1) = \beta_1$$

While the relationship between Y and X_2 is linear:

$$\Delta Y/\Delta X_2 = \beta_2.$$

LEARNING ACTIVITY 5

The data below refers to a sample of households (cross-sectional data).

(1) Estimate an appropriate form: CONS = f(INC). Assume that ΔCONS/ΔINC decreases as INC increases. Use your spreadsheet to perform the regression.

(2). Calculate ^CONS for INC = 3000, 6000 and 12000. Use these values to verify that Δ^CONS/(ΔINC/INC) is constant.

INC	CONS	where INC: disposable household income and CONS: household consumption on food & beverages (both measured as Rand per month).
3000	1200	
3500	1100	
4000	1500	
8000	2000	
9000	1950	
10000	2300	
15000	2600	

$$\Delta Y/\Delta X_2 = \beta_2$$

[11] This is easy to confirm using calculus. If Y = a + b.log(X), then assuming natural logs, $\Delta Y/\Delta X = b/X$. Thus $\Delta Y/(\Delta X/X) = b$.

ANSWERS

(1) The appropriate form is

CONS = a + b.log(INC)

Its estimate (using OLS and a PC) is

^CONS = -6058.639 + 897.0534 x log(INC)

where log is the natural logarithm.

(2) See the table below

INC	***Log(INC)***	***^CONS***	***Δ^CONS***	***ΔINC/INC***
3000	*8.0064*	*1123.50*		
6000	*8.6995*	*1745.29*	*621.79*	*1.000*
12000	*9.3927*	*2367.08*	*621.79*	*1.000*

It is evident that as INC doubles, ^CONS increases by a constant 621.79. Another way to state this is that Δ^CONS/(ΔINC/INC) is constant. The last column in the table uses the previous observation's INC.

(d) How to derive formulas for the slope and elasticity

The formulas that describe the slope and elasticity of the functional forms that use logs will not be given in the examination formula sheet, but may be easily derived.

Functional form	**Characteristic**	**Slope ΔY/ΔX**
Linear: Y = a + bX	ΔY/ΔX = b	b
Semi-logY: Y = ab^X or log(Y) = log(a) + log(b).X	(ΔY/Y)/ΔX = log(b)	log(b).Y
Semi-logX: Y = a + b.log(X)	ΔY/(ΔX/X) = b	b/X
Double log: Y = aX^b or log(Y) = log(a) + b.log(X)	(ΔY/Y)/(ΔX/X) = b	b(Y/X)

The second column provides an expression for the characteristic of each of the functional forms. This expression may be easily derived from that of the linear form. First note that the non-linear forms are based on the linear form by replacing Y by log(Y), and/or X by log(X). The corresponding characteristic of the non-linear form is derived from that of the linear form by replacing:

- ΔY by (ΔY/Y) if log(Y) is used in the form instead of Y
- ΔX by (ΔX/X) if log(X) is used in the form instead of X

Once the characteristic of the form has been derived, the expressions for slope and elasticity may be easily deduced from this. For example, in the case of form semilog Y, log(b) = (ΔY/Y)/ΔX

- the slope ΔY/ΔX = log(b).Y
- Proof

- Rewrite (ΔY/Y)/ΔX = log(b) as
- $\frac{\Delta Y}{Y}.\frac{1}{\Delta X} = \log(b)$. Now multiply both sides of this equation by Y which gives
- $\frac{\Delta Y}{\Delta X} = Y.\log(b)$

- the elasticity (ΔY/Y)/(ΔX/X) = log(b).X

7.2.4 Polynomial forms

The two most frequently used polynomial forms are the quadratic and the cubic forms. The quadratic form:

$$Y = aX^2 + bX + c$$

has one turning point which is either a minimum or a maximum point. Two examples of $Y = aX^2 + bX + c$ are demonstrated below.

Chart 1: Function has maximum point: Case a < 0

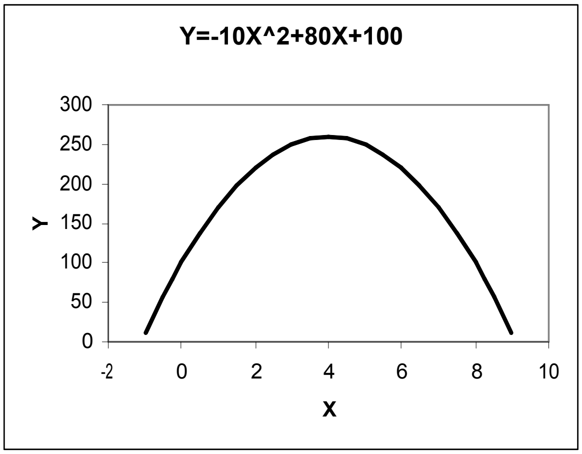
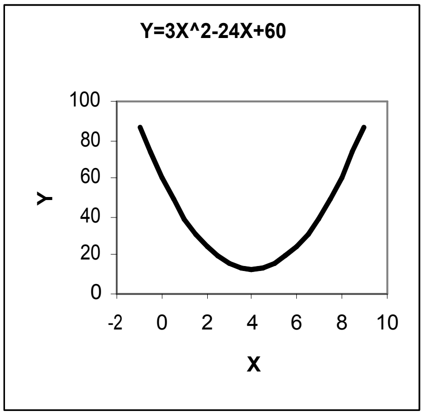


Chart 2: Function has minimum point: Case a > 0



Some practical examples of these cases are provided below.

(1) Consider the relationship between salary (Y) and work experience (X), assuming *ceteris paribus* conditions with respect to all other factors. Given a particular job, it is unlikely that salary will increase linearly with years of working experience. The situation

displayed in chart 1 could be more realistic where salary increases with experience (and age) at a decreasing rate. Even the eventual decline could be realistic in the sense that work performance drops as the age of a worker increases.

(2) Consider the effect of the age of a house (X) on its price (Y), again assuming conditions of *ceteris paribus*. Generally we would expect that the price of a house decreases as age increases. The decreasing leg of chart 2 could be a realistic representation of this price/age relationship if the age of the house increases.

The cubic function

$$Y = aX^3 + bX^2 + cX + d$$

has two turning points. One could use this function to describe a production function. However, since this form may sometimes produce unwanted results (see Studenmund), it should be used with care.

7.2.5 Inverse form

This form is used when the impact of X_1 is expected to decrease as X increases ($\beta_1 > 0$) or when Y is expected to increase and approach an asymptote value ($\beta_1 < 0$). In both cases the slope flattens when X increases. See figure 7.5: Inverse functions in the textbook.

A typical use of this form is in the Phillips curve. This curve relates the inflation rate (or the rate of increase of money wages) to the unemployment rate. As the inflation rate decreases, the unemployment rate increases, assuming a trade-off between the goals of low inflation and low unemployment.

7.3 Lagged independent variables

Lagged independent variables are important in econometric models. Lagged variables are used in time series models to model a time lag between cause and effect. The principle is simple to understand. Consider the consumption equation

$$C_t = \beta_0 + \beta_1 P_t + \beta_2 C_{t-1} + \varepsilon_t$$

This equation assumes a lag structure, that is, some of its impacts are not instantaneous, but spread over more than one period. In this case, consumption in the current period (C_t) is affected by the price (P_t) in the current period and by the level of consumption (C_{t-1}) in the previous period. In this case C_{t-1} is a one period lagged variable.

7.4 Using dummy variables

Dummy variables are commonly used as X-variables in regression equations to represent qualitative differences, for example, the condition of gender (male or female) or meeting a particular condition (having at least a master's degree).

(A) INTERCEPT DUMMIES

The simplest case is the intercept dummy variable as in

$$Y = \beta_0 + \beta_1 X + \beta_2 G + \varepsilon \qquad \textit{(equation 7.4)}$$

where G is the dummy variable which assumes values of 0 or 1 depending on gender. If, for example, the observation applies to a man then G = 1, and if the observation applies to a woman, then G = 0. In equation 7.4, β_2 then represents the difference in intercept explained by G (gender). In fact, equation 7.4 represents two different equations, one for men and one for women. This may be proved as follows:

If G = 0 then $Y = \beta_0 + \beta_1 X + \varepsilon$

If G = 1 then $Y = \beta_0 + \beta_1 X + \beta_2 + \varepsilon$ or $Y = (\beta_0 + \beta_2) + \beta_1 X + \varepsilon$

The intercept term changes from β_0 to $\beta_0 + \beta_2$ when G changes from 0 to 1.

The slope coefficient ($\beta_1 = \Delta Y/\Delta X$) remains the same irrespective of the value of G.

LEARNING ACTIVITY 6

Consider a model which explains the expenditure on recreation and sport (Y) by young unmarried professionals. The data are provided below.

Person	Monthly income INC	Spending Y	Age A	Gender G	Education level E
1	10 000	400	27	male	BA
2	16 000	200	24	female	MCom
3	8 500	500	25	male	BSc
4	12 600	0	29	male	Matric
5	24 500	600	30	female	MBA
6	6 700	200	22	male	Matric
7	7 000	200	23	female	BCom
8	14 600	300	34	male	Matric
9	9 000	250	28	female	BA

Compile an equation which expresses Y as a function of income (INC), age (A), gender (G) and level of education (E: graduates or non-graduates).

- Specify the equation (called equation 1) and state the expected value of all its coefficients.
- Define the meaning of all dummy variable(s).
- Assume a non-linear specification with respect to age. Assume that Y initially increases with age, and then decreases.
- Assume that men spend more than women, *ceteris paribus*.
- Assume that graduates spend more than non-graduates, *ceteris paribus.*

Assume an alternative specification which differs from model 1 in the sense that the meaning of the dummy variable for gender has been reversed (if previously G = 1 denoted men, now G = 1 denotes women). Explain what impact this has on the interpretation of the coefficient of G.

ANSWERS

(1) A simple specification is:

$$Y = \beta_0 + \beta_{INC}INC + \beta_A A + \beta_{A2}A^2 + \beta_G G + \beta_E E + \varepsilon \qquad \text{(equation 1)}$$

The dummy variable G denotes gender with say, G = 1 for men and G = 0 for women. The dummy variable E denotes level of education where E = 1 for graduates and E = 0 for non-graduates.

Variable	*Expected value of coefficient and reason*
Income	*$\beta_{INC} > 0$ since Y can be expected to increase with an increase in INC. One could also use log(INC), meaning that when INC increases, Y increases at a decreasing rate.*
Age	*A quadratic form is used. Since C must first increase with age, and then decrease, the expected value of coefficient $\beta_{A2} < 0$.*
Gender	*$\beta_G > 0$ since men spend more than women.*
Education	*$\beta_E > 0$ since graduates spend more than non-graduates.*

(2) Assume gender is specified as G = 0 for men and G = 1 for women. Because men spend more than women, the expected value of coefficient β_G will now be negative.

(B) USING DUMMY VARIABLES TO DENOTE MORE THAN TWO LEVELS

The previous section assumed using only 0 and 1 values of one dummy variable which denotes the presence or absence of one qualitative factor. For example, based on

$$Y = \beta_0 + \beta_{INC}INC + \beta_A A + \beta_{A2}A^2 + \beta_G G + \beta_E E + \varepsilon \qquad \text{(equation 1)}$$

the value of G determines the intercept which has two possible levels.

If G = 0 then $Y = \beta_0 + (\beta_{INC}INC + \beta_A A + \beta_{A2}A^2 + \beta_E E + \varepsilon)$ while

if G = 1 then $Y = (\beta_0 + \beta_G) + (\beta_{INC}INC + \beta_A A + \beta_{A2}A^2 + \beta_E E + \varepsilon)$.

Dummy variables can also be used to represent more than two levels of qualitative factors.

LEARNING ACTIVITY 7

Based on equation 1 explain how you would incorporate not only two levels of education (non-graduate and graduate) but four levels (non-graduate, B-degree, M-degree and D-degree). To keep things simple, only include the Y-variable and the variable dealing with the level of education.

- Explain why model A is insufficient

$$Y = \beta_0 + \beta_{INC}INC + \beta_E E + \varepsilon$$

where E = 0 for non-graduates), E = 1 for a B-degree, E = 2 for an M-degree and E = 3 for a D-degree.

- Explain why model B

$$Y = \beta_0 + \beta_{INC}INC + \beta_{EN}EN + \beta_{EB}EB + \beta_{EM}EM + \beta_{ED}ED + \varepsilon$$

is also insufficient where EN = 1 for non-graduates, EB = 1 for B-degrees, EM = 1 for M-degrees and ED = 1 for D-degrees and all these dummy variables = 0 if otherwise.

- Explain why model C (using all the variables of model B except EN)

$$Y = \beta_0 + \beta_{INC}INC + \beta_{EB}EB + \beta_{EM}EM + \beta_{ED}ED + \varepsilon$$

is correct, as well as the meaning of its coefficients.

ANSWER

It may be tempting to use the model A

$$Y = \beta_0 + \beta_{INC}INC + \beta_E E + \varepsilon$$

where E = 0 (non-graduates), E = 1 (B-degrees), E = 2 (M-degrees) and E = 3 (D-degrees). This specification, however, assumes a constant difference in Y between each successive level of education, which might be unrealistic. For example, the impact on Y between a non-graduate and a B-graduate may be much larger than between a B-graduate and an M-graduate.

Model B

$$Y = \beta_0 + \beta_{INC}INC + \beta_{EN}EN + \beta_{EB}EB + \beta_{EM}EM + \beta_{ED}ED + \varepsilon$$

is insufficient on grounds of multicollinearity. This matter is discussed fully in chapter 8. Let's use a table to demonstrate the problem. Each person's level of education could be coded as follows:

Condition	***EN***	***EB***	***EM***	***ED***
Non-graduate	*1*	*0*	*0*	*0*
B-degree	*0*	*1*	*0*	*0*
M-degree	*0*	*0*	*1*	*0*
D-degree	*0*	*0*	*0*	*1*

For every observation within the dataset, EN + EB + EM + ED = 1. This implies that if any three variables are known, then the remaining variable is also known. If, for example, it is known that EB = EM = ED = 0, then the observation must necessarily deal with a non-graduate. Thus EN is redundant and the solution is to omit variable EN. In fact, any one of (EN, EB, EM, ED) may be omitted, although the interpretation of these coefficients will then change.

Model C is a proper model to use.

$$Y = \beta_0 + \beta_{INC}INC + \beta_{EB}EB + \beta_{EM}EM + \beta_{ED}ED + \varepsilon$$

The interpretation of the coefficients is as follows.

(a) β_0 *includes the (constant term) effect on Y of non-graduates*
(b) β_{EB} *measures the difference in intercept between non-graduates and B-graduates*
(c) β_{EM} *measures the difference in intercept between non-graduates and M-graduates and*
(d) β_{ED} *measures the difference in intercept between non-graduates and D-graduates*

7.5 Slope dummy variables

Dummy variables can also be used to represent differences in slope. This is done by adding an X-variable called an interaction variable, as in

$$Y_i = \beta_0 + \beta_{INC}INC_i + \beta_G G_i + \beta_X(G_i.INC_i) + \varepsilon_i \qquad \text{(equation 2)}$$

where G.INC is the "interaction variable", which is the product of two X-variables.

LEARNING ACTIVITY 8

Demonstrate that equation 7.5 copes with both the effects of an intercept and slope change when G changes from 0 to 1.

ANSWER

If $G_i = 0$*, then equation 7.5 changes to*

$$Y_i = \beta_0 + \beta_{INC}INC_i + \varepsilon_i$$

if $G_i = 1$ then equation 7.5 changes to

$$Y_i = \beta_0 + \beta_{INC}INC_i + \beta_G + \beta_X(INC_i) + \varepsilon_i$$

which may also be written as

$$Y_i = (\beta_0 + \beta_G) + (\beta_{INC} + \beta_X)INC_i + \varepsilon_i$$

The effect is that when G_i changes from 0 to 1 then the

- intercept changes from β_0 to $\beta_0 + \beta_G$ and the
- slope changes from β_{INC} to $\beta_{INC} + \beta_X$.

Thus equation 2 accommodates both the effect of an intercept dummy and of a slope dummy.

LEARNING ACTIVITY 9

(a) Develop model 7.5.2, based on model 7.4.1, but where the slope ΔY/ΔINC varies with the level of education (E). Compare the slope and the intercept of model 7.5.2 when E changes from 0 to 1.

$$Y = \beta_0 + \beta_{INC}INC + \beta_A A + \beta_{A2}A^2 + \beta_G G + \beta_E E + \varepsilon \qquad \textit{(equation 1)}$$

where E = 0 for non-graduates and E = 1 for graduates.

(b) Create the dataset needed to estimate equation 7.5.2.

(c) Use your spreadsheet to estimate the coefficients of equation 2 by OLS.

ANSWERS

(a) The required equation in which ΔY/ΔINC varies with the level of education (E) is:

$$Y = \beta_0 + \beta_{INC}INC + \beta_A A + \beta_{A2}A^2 + \beta_G G + \beta_E E + \beta_X(INC.E) + \varepsilon \qquad \textit{(equation 2)}$$

As previously, INC.E is the interaction variable.

When E = 0 then

$$Y = \beta_0 + \beta_{INC}INC + \beta_A A + \beta_{A2}A^2 + \beta_G G + \varepsilon$$

When E = 1 then

$$Y = \beta_0 + \beta_{INC}INC + \beta_A A + \beta_{A2}A^2 + \beta_G G + \beta_E + \beta_X(INC) + \varepsilon$$

Comparison of slope and intercept:

	E = 0	***E = 1***
Intercept	β_0	$\beta_0 + \beta_E$
Slope ΔY/ΔINC	β_{INC}	$\beta_{INC} + \beta_X$

(b) The full dataset needed for the regression is provided below.

Y	*INC*	*A*	*A^2*	*G*	*E*	*INC.E*
400	*10000*	*27*	*729*	*1*	*1*	*10000*
200	*16000*	*24*	*576*	*0*	*1*	*16000*
500	*8500*	*25*	*625*	*1*	*1*	*8500*
0	*12600*	*29*	*841*	*1*	*0*	*0*
600	*24500*	*30*	*900*	*0*	*1*	*24500*
200	*6700*	*22*	*484*	*1*	*0*	*0*
200	*7000*	*23*	*529*	*0*	*1*	*7000*
300	*14600*	*34*	*1156*	*1*	*0*	*0*
250	*9000*	*28*	*784*	*0*	*1*	*9000*

(c) The regression coefficients are provided below in table format.

Variable	***Coefficient***	***Standard Error***	***t Stat***
Intercept	*3924.07*	*2229.64*	*1.760*
INC	*-0.05*	*0.03*	*-1.537*
A	*-298.75*	*168.54*	*-1.773*
A^2	*6.11*	*3.08*	*1.988*
G	*246.06*	*98.41*	*2.500*
E	*-204.87*	*327.05*	*-0.626*
INC.E	*0.06*	*0.03*	*2.117*

Note that these results do not appear very promising. Although the overall fit is quite good ($R^2 = 0.92$), some of the coefficients have unexpected values and some of them are insignificant. But then the sample size is very small!

7.6 Problems with functional forms

(a) Comparing R^2 of different functional forms

R^2 , which is ESS/TSS, measures the goodness of fit. The question arises whether R^2 can be used to compare two regressions which are of different functional forms, in order to decide which form is best. The answer is that the R^2s are not comparable. The reason is that the sums of squares (ESS and TSS) are affected by their unit of measurement, and thus by the log transformations. See the textbook, section 7.6.

(b) Extrapolating outside the sample range

A common error is to use regression results for forecasting outside the sample range. This becomes especially hazardous when using an erroneous functional form, which can lead to large forecasting errors. See the textbook, Figure 7.8.

(C) TRUE/FALSE QUESTIONS **(F) = false (T) = true**

(1) It is best always to include a constant term in a regression equation, even if it turns out to be statistically insignificant. (T)

(2) For OLS to work, a form must be linear in the coefficients, or it must be convertible to a form which is linear in the coefficients. (T)

An equation is linear in the coefficients if the coefficients appear in their simplest form, are not raised to any power (except one), are multiplied or divided by other coefficients, or appear within a function.

(3) For OLS to work, a form must be linear in the variables. (F)

The form $Y = a + bX + cX^2$, for example, is not linear in the variables, yet OLS can be used to estimate such a form. For OLS to work, it does require a form which is linear in the coefficients. In the form $Y = a + bX + cX^2$, all of (a, b, c) appear in their simplest form.

(4) $e^{2.302585} = 10$ where e = 2.71828. (T)

Make sure you know how to use your calculator or spreadsheet to derive this result. In Excel, for example, you use the function = EXP(2.302585).

(5) The average annual growth rate of real GDP over the period 2001–2006 is 4.4% per year. (T)

Year	Real GDP
2001	947 373
2002	982 121
2003	1 012 763
2004	1 062 027
2005	1 115 135
2006	1 175 216

Use OLS to derive $GDP = a.b^{YEAR}$ which must first be linearised to form log(GDP) = log(a) + log(b).YEAR. The estimate (use a PC and Excel) of this form is: log(GDP) = -72.353764 + 0.043032 (using natural logs = ln and rounding to 6 decimals). Values of YEAR: 2001, 2002, …) were used.

The growth factor = exp(0.043032) = 1.044 (rounded to 3 decimals) which denotes a growth rate of 4.4% per year.

(6) A production function has been estimated as Q = $100K^{0.65}L^{0,45}$
where K: capital input and L: labour input.

a If K increases by 1%, then Q will increase by 0.65%. (T)

Because $\frac{\Delta Q/Q}{\Delta K/K} = 0.65$ it follows that

$\frac{\Delta Q}{Q} = 0.65\, x \frac{\Delta K}{K} = 0.65\, x\, 0.01 = 0.0065$ which is an increase of 0.65%.

b If K increases by 0.45%, then Q will increase by 1%. (F)

The coefficient of L is 0.45, that of K is 0.65. Thus if K increases by 1%, then Q will increase by 0.45%.

c If both K and L increase by 1%, then Q will increase by 1%. (F)

If both K and L increase by 1%, then Q will increase by 0.45% + 0.65% = 1.10%.

(7) Assume we need a form Y=f(X) where Y: level of salary and X: age. Assume that Y increases as X increases up to the age of 55, but that Y decreases after the age of 55.

a A quadratic form will do. (T)

b One can also use a cubic form. (T)

c It is best to use an inverse form. (F)

d t is advisable to use a semilog form of type log(X). (F)

e It is best to use a double log-form. (F)

(8) When one uses a dummy variable (G) as in

$C_t = \beta_0 + \beta_1 Y_t + \beta_2 G_t + \varepsilon_t$

a then G_t assumes values of either 0 or 1 (T)

b it is required that $\beta_2 > 0$ (F)

c when G changes from 0 to 1, C will change by a constant amount of β_2 irrespective of the value of Y. (T)

d the slope ΔC/ΔY remains constant, irrespective of the value of G or Y. (T)

(E) EXAMINATION: PARAGRAPH QUESTIONS

(1) Explain the meaning of the constant term. Why is it almost always included in a regression equation, but almost never tested for statistical significance?

(2) When would the following functional forms be used? Provide the algebraic expression of each, as well as a key characteristic of each form.

- the linear form
- the double-log (exponential) form
- the semilog forms, both the log(Y) and the log(X) types
- the polynomial form
- the inverse form

(3) Explain fully why it is unwise to

- compare the R^2 values of different functional forms
- use functional forms outside the range of observation

(F) EXAMINATION: PRACTICAL QUESTIONS

You should be able to apply each of the six functional forms in practical situations, that is

- know when to use each of them
- know their characteristics
- be able to estimate their coefficients making use of appropriate transformations
- be able to interpret their coefficients

You should be able to specify, use and interpret

- slope dummies
- intercept dummies

Section C (true/false) and the activities in this learning unit, as in all chapters, provide good coverage of examination material.

PART IV

DEALING WITH ECONOMETRIC PROBLEMS

LEARNING UNIT 8

MULTICOLLINEARITY

ECONOMETRICS IN ACTION

If you search for the meaning of "multicollinearity" on the web (and how easy it has become to do such searches!), you will come across different explanations, for example:

- Multicollinearity occurs because two (or more) variables are related – they measure essentially the same thing.
- Multicollinearity is when variables are highly correlated (0.90 and above).
- Multicollinearity is a matter of paucity of information in the data. We don’t have enough independent variation in the Xs.
- You will also find different solutions for dealing with the problem.
- Multicollinearity is not a “disease”. It is not a violation of the model assumptions.
- Don’t over-invest in fancy statistical techniques to overcome the paucity of data. Invest in collecting more data.
- Multicollinearity exposes the redundancy of variables and the need to remove variables from the analysis.

Let us together explore the meaning of multicollinearity.

STUDY OBJECTIVES

The next three chapters deal with conventional econometric problems. Let's provide the broad context first. This course deals with regression analysis, and more specifically, the method of OLS. The purpose of OLS is to estimate the coefficients of a regression equation as accurately as possible. In learning unit 6: Choosing the independent variables, we already studied one type of problem, that is, of omitting a true X-variable. In this particular case, it had the undesirable consequence of biased estimates of the remaining coefficients.

Chapters 8 to 10 all have a common structure although they deal with different problems.

(1) First, the nature of the problem is defined.
(2) The consequences of the problem are explored.
(3) Methods of detecting the problem are dealt with.
(4) Lastly, remedies of dealing with the problem are addressed.

The first econometric problem is multicollinearity.

(A) PRESCRIBED MATERIAL

The following sections are prescribed:

(1) Perfect versus imperfect multicollinearity
(2) The consequences of multicollinearity

(3) The detection of multicollinearity
(4) Remedies for multicollinearity
(5) An example of Why Multicollinearity Often Is Best Left Unadjusted

Section 7 (p. 290): The SAT interactive regression learning exercise is not prescribed. It does, however, provide a useful learning-by-doing exercise. You are encouraged to perform this excercise, as it will improve your practical feel for econometrics.

(B) SOME IMPORTANT CONCEPTS

8.1 Perfect versus imperfect multicollinearity (The nature of the problem)

(a) Perfect collinearity

Assume the regression equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \; .$$

We have perfect collinearity if two X-variables, say X_1 and X_2, are perfectly linearly related. Linearly related means that

$$X_{2i} = a + bX_{1i} \qquad \textit{(equation 1)}$$

with a and b constants, for example a = 10 and b = 2.

How do we detect perfect collinearity? The graph of X_1 against X_2 will be a straight line. Alternatively we may measure the correlation coefficient between X_1 and X_2. If $r^2_{X1,X2} = 1$ then we have perfect collinearity.

(b) Perfect multicollinearity

Perfect multicollinearity occurs when one independent variable can be completely explained by a linear function of one or more other independent variables. One can, for example, have a relationship such as

$$X_{3i} = a + bX_{2i} + cX_{1i} \qquad \textit{(equation 2)}$$

where a, b and c are constants. Multicollinearity is more difficult to detect since more Xs are involved. The presence of such a linear relationship can be detected by regressing one X on the remaining Xs as in equation 8.1.2. Perfect multicollinearity is indicated when the R^2 of this fit is equal to 1.

Strictly speaking, in the case of two Xs, the term "collinearity" applies. Multicollinearity applies when more than two Xs are involved. However, both instances are covered by the general term "multicollinearity".

(c) Imperfect multicollinearity

In practice, we deal more often with imperfect multicollinearity. Imperfect collinearity occurs when

$X_{2i} = a + bX_{1i} + u_i$ *(equation 3)*

where the u_i's are error terms. If the u_is are relatively small, then $r^2_{X1,X2}$ will be relatively high, say 0.9.

Imperfect multicollinearity in the case of three Xs, occurs when

$X_{3i} = a + bX1_{1i} + cX_{2i} + u_i$ *(equation 4)*

The smaller the error term, the greater the multicollinearity will be. Multicollinearity may, of course, also involve more than three Xs.

In practice, multicollinearity is often the linear relationship between two explanatory variables. This can often be detected on theoretical grounds. When two Xs are used which virtually measure the same characteristic, then one should be especially cautious of the presence of multicollinearity. There are many examples of variables which are likely to be highly correlated. Consider the following examples: income and wealth of households, height and weight of persons, size of the left foot and the size of the right foot of individuals, and the value and the volume of sales of firms.

In the case of imperfect multicollinearity the OLS technique will work but the standard errors of the estimated coefficients will be high. Under conditions of perfect multicollinearity the OLS technique will not work. Depending on the software package, some will abort OLS, and give a message "numerically singular matrix". Others might proceed with estimation but with the standard errors of the estimated coefficients set equal to zero.

LEARNING ACTIVITY 1

Does collinearity exist between X_1 and X_2?

ANSWER

X_1	X_2
0	*4*
1	*2*
2	*0*
3	*-2*

Since $X_2 = 4 - 2X_1$, there is perfect collinearity between X_1 and X_2.

Also, $r_{12} = -1$.

In Excel, use the = Correl(range X_1, range X_2) function.

LEARNING ACTIVITY 2

Assume you are having a need to include gender in a regression equation. Amongst the X-variables, you define two dummy variables, MALE and FEMALE, which are both included in the regression equation. Explain whether this scheme gives rise to problems.

Name	**MALE**	**FEMALE**
ZITHULELE	1	0
RAJENDRA	0	1
MBELU	1	0
PATIENCE	0	1

ANSWER

Yes, there is a problem. The two variables MALE and FEMALE as used above are perfectly collinear since

$$MALE + FEMALE = 1$$

for all observations. Thus a perfect linear relationship exists between the X's. If one value, say that of MALE, is known, then the variable FEMALE is automatically known. The solution is to drop one of these variables from the regression equation.

LEARNING ACTIVITY 3

A regression equation uses the following quadratic form:

$$Y = a + bX + cX^2 + dZ + \varepsilon$$

Since X and X^2 are related, does this not imply perfect multicollinearity?

ANSWER

*The answer is no. Perfect multicollinearity requires that a perfect **linear** relationship exist between Xs, which is of course not the case here. The r^2 between X and X^2 is, however, likely to be high.*

LEARNING ACTIVITY 4

Student Y wants to accommodate the seasonal variation in quarterly data (seasonally unadjusted) in a model which explains expenditure for the period 1990Q1 to 2006Q4. A seasonal variation is the variation in expenditure due to the time of the year, that is, depending on the quarter in which it occurs. For example, when real expenditure in the first quarter is typically R50 million below the quarterly average of the year, in the second quarter R25 million below, in the third quarter R15 million above and in the fourth quarter R60 million above the quarterly average, this constitutes a seasonal effect. Student Y proposes to use four dummy variables D_1, D_2, D_3 and D_4 to account for the seasonal effect as follows:

Quarter	D_1	D_2	D_3	D_4
1990.1	1	0	0	0
1990.2	0	1	0	0
1990.3	0	0	1	0
1990.4	0	0	0	1
1991.1	1	0	0	0
1991.2	0	1	0	0
Et cetera				

Explain why this method will not work. Propose a method which will work.

ANSWER

This method will not work because of multicollinearity. Since $D_1 + D_2 + D_3 + D_4 = 1$ for all observations, a perfect linear relationship exists between these four variables.

The solution is to omit one of the four, for example D_4. In this case $D_1 + D_2 + D_3$ is not equal to a constant any more. Although $D_1 + D_2 + D_3 = 1$ for quarter 1 to 3, it is equal to zero for quarter 4.

8.2 Consequences of multicollinearity

The consequences of multicollinearity are as follows:

- Estimates of $\hat{\beta}$ remain unbiased.
- The $SE(\hat{\beta})$ will increase. In the case of $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$

$$SE(\hat{\beta}_1) = \sqrt{\frac{\sum e_i^{\,2} /(n-3)}{\sum\left(X_{1_i} - \overline{X}_1\right)^2\left(1 - r_{12}^{\,2}\right)}}.$$

  In the case of considerable multicollinearity between X_1 and X_2, r_{12} will be high. Thus the denominator term will decrease, so that the standard error (SE) will increase. In fact, in extreme cases, the SE is infinite for example under conditions of perfect multicollinearity, when $r_{12} = 1$.

- Because the $SE(\hat{\beta})$ increases, the t-values will decrease.
- Estimates will be very sensitive to changes in specification.
- The overall fit (R^2) is unaffected.
- The greater the multicollinearity**,** the greater the consequences will be.

8.3 Detection of multicollinearity

If the nature of and the consequences of multicollinearity are understood, its detection follows logically. The signs to watch out for are as follows:

- The regression equation has a high R^2 but no significant t-values.
- High r^2 values between pairs of Xs.
- Because the r^2 value only detects collinearity between pairs of the Xs, it may be deficient as a measure of multicollinearity. Multicollinearity may also involve more Xs. In the case of three Xs, this does not necessarily imply a high r^2 between (X_1, X_2) or between (X_1, X_3).
- High variance inflation factors (VIFs).

  This is a more comprehensive measure than r^2 and it copes with the previous problem of more Xs.

  Assume we want to estimate

  $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon : \qquad \textit{(equation 5)}$$

  To calculate the VIF for say β_1, its X-variable, X_i is expressed as a linear function of all the other Xs:

  $$X_1 = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 + v \qquad \textit{(equation 6)}$$

  where v is an error term. To calculate the VIF, OLS is applied to equation 8.3.2, which is called the auxiliary equation. In this case our only interest is its goodness of fit (R^2 of equation 2). In general the VIF is then calculated as: $VIF(\beta_i) = \frac{1}{1 - R_i^{\,2}}$ where R_i^2 is the unadjusted R^2 of X_i regressed on the other X-variables of equation 5. The VIF denotes the degree of multicollinearity. There is no table of critical values but a value above 5 is generally regarded as being indicative of serious multicollinearity.

LEARNING ACTIVITY 5

Evaluate: Equation 5 has an R^2 of 0.8. Thus its VIF = 5.

ANSWER

Incorrect. The R^2 refers to that of the auxiliary equation, and not to that of the original equation.

LEARNING ACTIVITY 6

Evaluate: For equation 5 one may derive three VIFs.

Correct. Since there are three Xs, there are also three VIFs based on the three auxiliary equations:

$$X_1 = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 + v_1$$

$X_2 = \alpha_1 + \alpha_1 X_1 + \alpha_3 X_3 + v_2$

$X_3 = \alpha_1 + \alpha_1 X_1 + \alpha_2 X_2 + v_3$

8.4 Remedies for multicollinearity

Multicollinearity is a "normal" phenomenon, as it often occurs in data. Thus, the problem of multicollinearity is one of degree. The proper remedy depends on the severity of the consequences. Let us consider the two extreme cases.

- In the case of perfect multicollinearity, it is obvious that something must be done, since the OLS method does not work. The most common remedy is dropping one of the multicollinear variables, or to combine the two collinear X-variables into one X-variable.
- If multicollinearity is present, but the t-values are nevertheless significant or appear sufficiently reliable, then the proper action may be simply to do nothing!

In practice it may well be that the proper remedy falls between these two extremes. Some techniques which can be used are the following:

- The method of first differences, which has both advantages and disadvantages - see the textbook.
- Increasing the sample size is a technique which may sometimes be used. Since a larger sample size reduces the $SE(\hat{\beta})s$, this may offset the adverse effect of multicollinearity in an indirect way.

(C) TRUE/FALSE QUESTIONS (F) = FALSE (T) = TRUE

Assume the following regression equation in the questions 1 and 2 below.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i \qquad \textit{(equation 1)}$$

(1) In equation (1) a high r^2 between Y and X_1 does no harm. (T)

It may in fact be desirable!

(2) A high r^2 between X_1 and X_2 in equation 1 is harmful. (T)

Correct. This is the problem of multicollinearity.

(3) The problem of multicollinearity can and must be eliminated completely. (F)

Incorrect. Since multicollinearity is a normal phenomenon of economic data, the real problem is one of its severity.

(4) In the presence of multicollinearity, a change in specification of the regression equation, may lead to large differences in the estimates of the coefficients. (T)

Correct. Since the SEs are large, large fluctuations in the estimates of the coefficients may occur when the specification changes.

(5) Multicollinearity tends to reduce the value of R^2 of the regression equation. (F)

Multicollinearity is a problem of dependencies between the Xs and typically manifests itself in high R^2 and low t-values.

(6) Multicollinearity can lead to the problem of unexpected signs of coefficients. (T)

(E) EXAMINATION: PARAGRAPH QUESTIONS

Explain

1 the nature of the multicollinearity problem: the difference between perfect and imperfect multicollinearity (4)
2 the consequences of multicollinearity (5)
3 the detection of multicollinearity; explain when to use r^2 and when to use VIFs (5)
4 remedies for multicollinearity (6)

(F) EXAMINATION: PRACTICAL QUESTIONS

You should be able to detect and remedy multicollinearity problems in practical situations. You might, for example, be given a regression result of which some of the coefficients are insignificant. It is, however, left to you to identify the source of the problem, which could be multicollinearity.

The SAT[12] interactive regression learning exercise (section 7) provides good practice in the identification of multicollinearity. The "best" regression run is 6 (p. 302). Hints for the SAT interactive regression exercise are given in pp. 319 to 320 in the prescribed book.

[12] The SAT is an acronym for the Scholastic Aptitude Test, which is commonly used to gain entrance to American universities. It measures a combination of inborn intelligence plus acquired knowledge, and is a good indicator of scholastic performance.

LEARNING UNIT 9

SERIAL CORRELATION

ECONOMETRICS IN ACTION

What is serial correlation? A Google search provides an overview of the problem.

- *Data is often of a “cyclical” nature. When errors associated with observations of different time periods are related to each other, we refer to the errors as being serially correlated.*
- *Autocorrelation means that the values of the error term in one period influences the error term in another period. That probably sounds pretty arcane and technical.*
  Serial correlation appears to occur often.
- *The returns to hedge funds and other alternative investments are often highly serially correlated.*
- *Technical analysts use serial correlation to determine how well the past price of a security predicts the future price.*
  Serial correlation can cause misleading results.
- *With positive serial correlation, the OLS estimates of the standard errors will be smaller than the true standard errors. This will lead to the conclusion that the estimates are more precise than they really are.*
- *Estimates of statistical significance will be vastly exaggerated.*

In this learning unit, we hope to teach you

- *to respect serial correlation, but not to fear it. It's a beast, but one that can be tamed.*

STUDY OBJECTIVES

When you have studied this learning unit you should understand

- the nature of serial correlation
- the consequences of serial correlation
- how to detect serial correlation
- how to fix serial correlation

(A) PRESCRIBED MATERIAL

The second conventional econometric problem is about serial correlation, also called autocorrelation.

All the sections are prescribed:

(1) Pure versus impure serial correlation
(2) The consequences of serial correlation
(3) The Durbin-Watson d test
(4) Remedies for serial correlation

(B) SOME IMPORTANT CONCEPTS

9.1 Pure versus impure serial correlation (The nature of the problem)

(a) Pure serial correlation

You should be able to explain pure serial correlation using mathematical notation, that is,

$E(r_{\varepsilon_i \varepsilon_j}) = 0$ for all i, j = 1 … n except i = j

as well as first-order serial correlation

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

where ε_t is the true error term of observation t, ρ is the first-order correlation coefficient of which the value always falls between -1 and 1 and u_t denotes a non-serially correlated error term. Both pure positive and pure negative first-order serial correlation occur depending on the value of ρ. Because pure positive first-order serial correlation occurs most frequently in practice, problems in the examination will only deal with positive first-order serial correlation.

(b) Impure serial correlation

Since the error term depends on the specification of the equation, a specification error can lead to the error term behaving as if pure first-order serial correlation is present. Consult the textbook for some examples.

- Since the error term is forced to absorb the effect of the omitted variables, omitted variables may lead to characteristics as if the error term is serially correlated.
- The use of an incorrect functional form may also lead to behaviour indistinguishable from a serially correlated error term. See figures 4 and 5 (pp. 329 to 330) in the textbook.

9.2 Consequences of serial correlation

Pure serial correlation

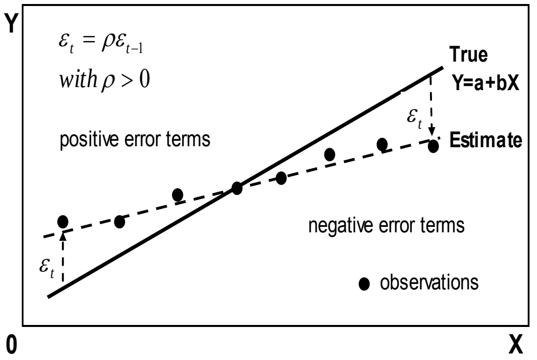
- does not cause bias in the coefficient estimates.
- increases the standard errors of $\hat{\beta}$.
- causes OLS to underestimate the standard errors of $\hat{\beta}$. This implies that the t-values of coefficients are likely to be overestimated.

(a) An intuitive explanation of the consequences of serial correlation

Assume a simple linear equation $Y_i = \beta_0 + \beta_1X_i + \varepsilon_i$. If the error terms are randomly scattered around the true regression equation, then the estimates of the coefficients are likely to be reasonably accurate. In the case of serial correlation, this is not the case. With serial correlation, there is a non-random pattern in the error terms, which causes problems of estimation.

The principle can best be demonstrated graphically. Assume that positive first-order serial correlation is present and that the first error term is positive. Intuitively, by inspection of the extreme case in graph 9.2 below, it can be said that

- the large positive value of the first error term propagates itself in a first-order serial correlation pattern
- this leads to the problem that the slope coefficient is underestimated



GRAPH 1: Extreme example of the effect of positive serially correlated error terms on a slope estimate

This particular result hinges on the assumption that the first error term is both large and positive. This, of course, is not always necessarily true. The first error term may also be both large and negative, in which case the slope is likely to be overestimated. Or, it is also possible that the initial error term is small, in which case the estimated slope will be more or less correct.

Can anything in general be deduced from this?

- The estimate of the slope, in the presence of serial correlation, is subject to more variation than would otherwise be the case. If the initial error terms are positive and the last error terms are negative (as above) then the slope is underestimated. If the initial error terms are negative and the last error terms are positive then the slope is overestimated. We cannot predict which case occurs in practice since the true error terms are unknown. But there is increased variation in estimating the slope. The increased variation is equivalent to saying that the standard errors of the slope coefficients are increased.
- It is likely that the standard errors of the slope coefficients will be underestimated. (see graph 1). If a straight line is fitted through the observed points, then the error terms are underestimated, since the true slope is estimated incorrectly. It is not uncommon to find that the standard errors of the estimated coefficients are underestimated by a factor of two! This leads to exaggerated claims of accuracy of the coefficient estimates.

The previous example is rather extreme, and serves only to explain the consequences of serial correlation. Note that the example displays an incomplete cycle of error terms, starting with a large positive value and ending with a large negative one. In practice we frequently observe at least a full cycle of error terms (for example, in the case of positive serial correlation an error term which starts positive, turns negative, and then turns positive again).

If less than a full cycle of error terms is observed, then the detection of serial correlation would be difficult, if not impossible. You will note that the critical values for the Durbin-Watson test, which tests for serial correlation, require a minimum of 15 observations. The intuitive reason for this requirement is to allow for at least a full cycle of the error terms.

9.3 The Durbin-Watson d test (Detecting serial correlation)

A special test has been devised to test for the presence of serial correlation. The test uses the residuals which arise from the usual OLS estimate. The assumptions which underlie this test are as follows:

- The regression model includes an intercept term.
- The serial correlation is first-order by nature.
- The regression model does not include a lagged dependent variable as an independent variable.

The Durbin-Watson d statistic (DW-d) is calculated as

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^{\,2}} \qquad \textit{(equation 1)}$$

where T denotes the number of observations and e_t denotes the residual term. Formula 1 is included in the examination formula sheet. Make sure you can apply it.

In the case of extreme positive serial correlation, $e_t = e_{t-1}$. This will cause the numerator of equation 1 to be zero, thus d = 0. In the case of extreme negative serial correlation, $e_t = -e_{t-1}$. This will cause d ≈ 4[13]. In the case of no serial correlation, d = 2. See figure 9.6 in Studenmund for the full range of range of values of the DW-d statistic, their meaning and for an explanation of the inconclusive range.

The following steps apply when using the DW-d statistic.

- Run an OLS regression to obtain residual terms.
- State the null and alternative hypotheses. In most cases we test for the presence of positive serial correlation.
- Derive the DW-d statistic.
- Compare the observed DW-d statistic with the critical values in the tables.
- Draw the appropriate conclusion.

LEARNING ACTIVITY 1

Assume a time-series regression. Test for the presence of positive serial correlation when DW-d = 1.26, the number of observations is 36 and the number of slope coefficients is 5. Use a 5% level of significance.

[13] Equation 1 will be: $\sum(2e_t)^2/\sum e_t^2 = \sum 4e_t^2/\sum e_t^2 = (T - 1)4/T \approx 4$.

ANSWER

The first step is always to state the hypotheses. In this case a one-sided test of positive serial correlation applies.

H_o: $\rho \leq 0$ and H_a: $\rho > 0$

The critical values are (n = 36, K = 5, see table B4 in Studenmund):

$D_L = 1.18$ and $D_U = 1.80$.

Since the observed DW-d of 1.26 falls within D_L and D_U, the test is inconclusive. A DW-d value of more than 1.80 is required to confirm the absence of positive serial correlation. A DW-d value of less than 1.18 would confirm the presence of positive serial correlation.

9.4 Remedies for serial correlation

Assuming that a correct diagnosis of serial correlation has been made, and in view of the adverse consequences, what can be done about the problem?

First, one has to ascertain the nature of the serial correlation problem. If it is of the impure type, then the obvious solution is to improve the specification. This may be done, for example, by identifying and adding the omitted variable.

In the case of pure serial correlation, the method of generalised least squares (GLS) can be used. The principle of this method is to transform the equation, so that the classical requirement in respect of the error terms is met. We expect you to be able to explain this method in the examination, which implies that you must understand and be able to present equations 15 to 20 in the prescribed book.

We start with the equation to be estimated

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \qquad \textit{(equation 2)}$$

where the error terms have first-order serial correlation

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

The following transformation "purifies" equation 9.4.1 of the serially correlated error term

$$Y_t - \rho Y_{t-1}$$

since it leads to

$$Y_t - \rho Y_{t-1} = \beta_0(1-\rho) + \beta_1(X_t - \rho Y_{t-1}) + u_t \qquad \textit{(equation 3)}$$

but also written as

$$Y_t^* = \beta_0^* + \beta_1 X_t^* + u_t \qquad \textit{(equation 4)}$$

where $Y_t^* = Y_t - \rho Y_{t-1}$ and $X_t^* = X_t - \rho X_{t-1}$.

Note that in equations 3 and 4 the error term u_t is not serially correlated any more.

How do we estimate equation 3? Because equation 3 is not linear in the coefficients we cannot apply OLS to it. There are three approaches to this problem.

- The first is to estimate ρ separately and then use it to transform equation 3 to 4 which is OLS friendly. This is the generalised least squares (GLS) method.
- A variation of this is the Cochrane-Orcutt method which applies GLS in an iterative way.
- The AR(1) method estimates all the coefficients in equation 3 simultaneously but does not use OLS.

GLS-method

GLS is not difficult to understand. It involves the following steps:

Step	Procedure
1	Find an initial estimate of ρ called $\hat{\rho}$.
2	Transform: $Y_t^* = Y_t - \rho Y_{t-1}$ and $X_t^* = X_t - \rho X_{t-1}$.
3	Run a normal OLS regression on $Y_t^* = \beta_0^* + \beta_1 X_t^* + u_t$
4	Transform the constant term Since $\beta_0^* = \beta_0(1-\rho)$ then $\beta_0 = {\beta_0^*}/{(1-\rho)}$. The slope does not have to be transformed.

How do we estimate an initial ρ? The method most used is to simply run OLS to initially estimate the residual terms e_t. We then run another OLS on these residuals to estimate $\hat{\rho}$ in $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$ where the constant term is suppressed and u_t is a non-serially correlated error term. The Cochrane-Orcutt method in the next section also uses this method.

Cochrane-Orcutt method

This is an iterative procedure particularly useful for small samples. Iterative means that the GLS procedure is run a number of times.

Assume that it is required that

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i \qquad \textit{(equation 5)}$$

be estimated and that first-order serial correlation is present

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t .$$

Step	Procedure
1	Run OLS to initially estimate the residual terms e_t.
2	Run an OLS regression on the residuals to estimate $\hat{\rho}$ in $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$. The constant term is suppressed. The term u_t is a non-serially correlated error term.
3	Use $\hat{\rho}$ as obtained in step 2 in a GLS procedure. This provides estimates of all the coefficients of equation 5 as well as a new set of residuals, e_t.

Steps 2 and 3 are run repeatedly until $\hat{\rho}$ (and the estimated coefficients) converge. Each run of step 2 uses updated residuals and provides updated estimates of $\hat{\rho}$. Step 3 uses the updated $\hat{\rho}$ to provide an updated GLS estimate of equation 2.

AR(1) method

An alternative to the Cochrane-Orcutt method is the AR(1) method of applying GLS. This is an iterative method that is most popular in practice, but not dealt with in this course. It estimates the coefficients of equation 9.4.2 simultaneously. This, however, requires a more advanced method than OLS. The advantage of using an econometric package is evident in this case. The AR(1) method is performed effortlessly in EViews by simply adding an AR(1) term to the specification.

LEARNING ACTIVITY 2

This learning activity takes you through all the steps of GLS. Its results confirm that GLS improves on the OLS estimates.

This learning activity uses a technique often applied in econometrics, that is, of simulation. It starts off by assuming some true model, building into it some problem (in this case serial correlation) and then testing whether some technique (GLS) overcomes the problem.

A true model Y = 15.5 + 0.8X is assumed. Serial correlation of the form $\varepsilon_t = 0.65\varepsilon_{t-1} + u_t$ is then incorporated into it. The observed data is provided at the right (which includes the effect of the serially correlated error terms).

Both OLS and GLS are then applied to this data.

Year	X	Y
85	12	23.1
86	28	32.6
87	34	38.5
88	35	41.4
89	36	44.5
90	45	54.8
91	55	73.5
92	64	73.8
93	65	78.3
94	76	86.5
95	78	94.0
96	87	93.7
97	90	89.3
98	92	81.2
99	112	106.5

True equation:

$Y_t = 15.5 + 0.8(X_t) + \varepsilon_t$

where

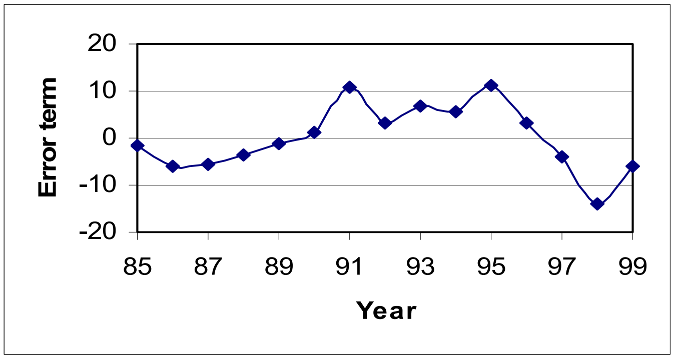
$\varepsilon_t = 0.65\varepsilon_{t-1} + u_t$

Note: Use a PC and OLS to perform the calculations.

(1) Estimate the coefficients of Y = a + bX by OLS

(2) Derive the residual terms of OLS. Plot the residual terms. Can you ascertain the nature of the serial correlation graphically?

(1) The OLS estimate is $\hat{Y} = 14.12295 + 0.8799(X)$. Use your PC to verify this result.



(2) The graph indicates positive serial correlation. There are "runs" in the error terns, meaning that the terms are first negative, slowly turn to positive, and then turn negative again. Such a pattern is not random but denotes interdependence between consecutive error terms, which of course reflects $\varepsilon t = 0.65\varepsilon_{t-1} + u_{t,}$ which has been built in.

(3) Calculate the DW-d statistic.

Y	$\hat{Y}_i$	**e_i**	**$(e_i - e_{i-1})^2$**	**e_i^2**
23.10	*24.68*	*-1.58*		*2.50*
32.60	*38.76*	*-6.16*	*20.97*	*37.96*
38.50	*44.04*	*-5.54*	*0.38*	*30.70*
41.40	*44.92*	*-3.52*	*4.08*	*12.39*
44.50	*45.80*	*-1.30*	*4.93*	*1.69*
54.80	*53.72*	*1.08*	*5.67*	*1.17*
73.50	*62.52*	*10.98*	*98.02*	*120.58*
73.80	*70.44*	*3.36*	*58.05*	*11.30*
78.30	*71.32*	*6.98*	*13.10*	*48.74*
86.50	*81.00*	*5.50*	*2.19*	*30.28*
94.00	*82.76*	*11.24*	*32.95*	*126.40*
93.70	*90.68*	*3.02*	*67.56*	*9.14*
89.30	*93.32*	*-4.02*	*49.56*	*16.13*
81.20	*95.08*	*-13.88*	*97.22*	*192.56*
106.50	*112.68*	*-6.18*	*59.31*	*38.13*
Sum			*513.99*	*679.67*

DW-d = 513.99/679.67 = 0.7562

(4) Test the DW-d statistic for statistical significance at the 5% level.

The null and alternative hypotheses are:

H_0: $\rho \leq 0$ and H_a: $\rho > 0$

The critical values are (N = 15, K = 1,5%):

$D_L = 1.08$ $D_U = 1.36$

Since 0.75 < D_L we can reject H_0. Thus positive serial correlation is present.

(5) Apply the GLS method to estimate the coefficients of $Y = \beta_0 + \beta_1 X$ using $\hat{\rho} = 0.6219$.

The transformed X and Y* are:*

*X**	*Y**
20.537	*18.235*
16.587	*18.227*
13.856	*17.458*
14.234	*18.754*
22.612	*27.126*
27.016	*39.421*
29.797	*28.092*
25.200	*32.405*
35.578	*37.807*
30.737	*40.208*
38.494	*35.244*
35.897	*31.030*
36.031	*25.666*
54.787	*56.004*

The GLS estimates are:

$$Y^* = 6.5055 + 0.833655\left(X^*\right).$$

Of course, the constant term must be transformed:

$$\hat{\beta}_0 = \frac{\hat{\beta}_0{}^*}{(1-\hat{\rho})} = \frac{6.5055}{\left(1-0.6219\right)} = 17.2$$

(6) Compare the slope and the SE(slope) of OLS with that of GLS. Does this confirm the theoretical expectations?

Estimation method	**Slope**	**SE of slope**
OLS	0.98	0.068
GLS	0.83	0.147

The theoretical expectations of serial correlation are

(1) estimates of the coefficients remain unbiased

(2) the standard errors of $\hat{\beta}$ increase

(3) OLS underestimates the standard errors of $\hat{\beta}$.

We cannot really draw any conclusions regarding the unbiasedness of $\hat{\beta}$. But we can observe the net effect of (2) and (3). This confirms that OLS underestimates the SE(slope) by roughly a half! The OLS estimates are misleading!

(C) TRUE/FALSE QUESTIONS **(F) = FALSE (T) = TRUE**

(1) Pure serial correlation is present when the error term of a time-series regression equation is systematically affected by the previous observation's error term. (T)

Note that this applies only in the case of a time series regression. In the case of cross-sectional data, the term "previous observation's error term" is meaningless, because the order of observation is irrelevant.

(2) It is good practice to test all time-series data variables for serial correlation before they are used in a regression. (F)

The error term applies only to a regression equation and not to a data series. We cannot test a variable for serial correlation outside the context of a regression model.

(3) Of all the types of serial correlation, first-order positive serial correlation is the most commonly observed in practice. (T)

(4) Graph 9.2 attempts to explain why serial correlation leads to biased estimates of the slope coefficient. (F)

Incorrect. It explains why $\hat{\beta}$'s are subject to increased variation and why the SE($\hat{\beta}$) is underestimated.

(5) Equation Y = a + bX is estimated and first-order positive serial correlation is indicated. This implies that the reported SE of the slope coefficient, say 0.16, is in truth much smaller, say 0.08. (F)

Since serial correlation is likely to cause underestimation of the standard errors of the coefficients, the true SE is likely to be higher than 0.16, say 0.30.

(6) An equation is corrected for serial correlation by using the GLS procedure. However, the SE of the slope coefficients turns out to be larger than before. Thus we can conclude that the GLS procedure has not been successful. (F)

Since serial correlation is likely to cause underestimation of the standard errors of the coefficients, the GLS estimates of the standard errors are likely to be larger, but more accurate.

Note that no transformation is involved in the case of estimating the slope coefficients by GLS. Thus the reported SE of the slope coefficients is comparable between OLS and GLS.

(7) GLS only uses a set of N-1 observations where N is the total number of observations in the original data set. (T)

(8) The residuals resulting from a successful GLS procedure are not serially correlated. (T)

(D) EXAMINATION: PARAGRAPH QUESTIONS

(1) Explain

(a) the nature of serial correlation: the difference between pure and impure serial correlation (5)
(b) the consequences of serial correlation (5)
(c) how serial correlation may be detected by the use of the DW-d statistic (5)

(2) Discuss remedies for the problem of serial correlation. Pay attention to (10)

(a) the theoretical foundations of the GLS method (derive an appropriate algebraic transformation which shows why the method works)
(b) the application of the GLS-method
(c) the application of the Cochrane-Orcutt method

(E) EXAMINATION: PRACTICAL QUESTIONS

Make sure you know how to apply the Durbin-Watson d test to test for positive serial correlation

You should also be able to

- Detect a serial correlation problem in practical situations
- know how to remedy it

To hone your practical skills

- answer question 2 on p. 344 (per capita consumption of beef). Compare your answers with those provided in p. 355.

ECONOMETRICS IN ACTION (looking back)

We hope you have now learnt

- to respect serial correlation, but not to fear it. It's a beast, but one that can be tamed!

What is serial correlation?

- Please confirm that it is a model phenomenon. The error term of a regression equation ($Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$) in period t is affected by the error term in period t-1 ($\varepsilon_t = \rho\varepsilon_{t-1} + u_t$).
- There is no reason that serial correlation sounds pretty arcane and technical anymore!

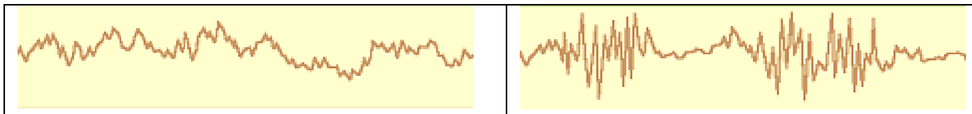
LEARNING UNIT 10

HETEROSKEDASTICITY

ECONOMETRICS IN ACTION

Well, I can hardly pronounce "heteroskedasticity"; can anyone tell me what it is?

- The classic example of heteroskedasticity is that of income versus food consumption. As one's income increases, the variability of food consumption will increase. A poorer person will spend a rather constant amount on food; a wealthier person may occasionally eat an expensive meal. Those with higher incomes display a greater variability of food consumption.
- Heteroskedasticity is best explained graphically against time:



Homoskedasticity

Heteroskedasticity

- The consequences are similar, but not quite the same as for serial correlation. When OLS is applied to heteroskedastic models the estimated variance is a biased estimator of the true variance. That is, it either overestimates or under-estimates the true variance and, in general, it is not possible to determine the nature of the bias. The standard errors may therefore be either understated or overstated.
- Two economists, Robert Engle and Clive Granger, shared the 2003 Nobel Prize for Economics for their paper on "Autoregressive conditional heteroskedasticity".

In this chapter, we will stick to the simple version of heteroskedasticity.

STUDY OBJECTIVES

When you have studied this learning unit you should understand

- the nature of heteroskedasticity
- its consequences
- how to detect it
- how to fix it

Note: Heteroskedasticity is also written as “Heteroskedasticity”. Both spellings may be used.

(A) PRESCRIBED MATERIAL

Heteroskedasticity is the third conventional econometric problem to be dealt with.

The following sections are prescribed:

10.1	Pure versus impure heteroskedasticity
10.2	The consequences of heteroskedasticity.
10.3	Testing for heteroskedasticity.
	The Park test and the White test are prescribed.
	The Goldfeld-Quandt test is **not** prescribed.
10.4	Remedies for heteroskedasticity.

Section 10.5: A more complete example is useful to reinforce the ideas learnt in the previous sections

B SOME IMPORTANT CONCEPTS

10.1 Pure versus impure heteroskedasticity

(a) Pure heteroskedasticity

We expect you to be able to describe the problem of pure heteroskedasticity using mathematical notation. Heteroskedasticity occurs when the classical assumption that the error term has constant variance

$$Var(\varepsilon_i) = \sigma_i^2 = \sigma^2 \text{ for all I = 1 … n}$$

is not met. Make sure you understand the graphical explanation provided in figure 10.1 in the textbook. The problem often occurs when estimating regression equations which use cross-sectional data.

To model the problem of heteroskedasticity in the typical regression equation

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \varepsilon_i$$

the variance of the error term is assumed to be

$$Var(\varepsilon_i) = \sigma^2 Z_i^2$$

where Z_i is the proportionality factor. See figures 10.2 and 10.3 in the textbook for a graphical representation of this type of heteroskedasticity. This form may be an oversimplification of the real situation. It is nevertheless used because of its simplicity, in the same way as, in the case of serial correlation, the first-order type is mostly used. In practice the level of income is often used as a proxy for the Z variable.

(b) Impure heteroskedasticity

You should be able to distinguish between pure and impure heteroskedasticity. Impure heteroskedasticity is caused by an error in specification. Since the error term also absorbs any error in specification, it may well cause heteroskedasticity in the error terms. The proper remedy in the case of impure heteroskedasticity would be to include the omitted variable, which could diminish the extent of the error term quite substantially by absorbing a large part of the previously unexplained variation.

See section 10.1 in the textbook for an example of impure heteroskedasticity.

10.2 Consequences of heteroskedasticity

The consequences of heteroskedasticity are almost identical to those of serial correlation.

(1) Pure heteroskedasticity in a regression equation does not cause bias in the coefficient estimates. Thus $E(\hat{\beta}) = \beta$.

(2) Pure heteroskedasticity increases the $SE(\hat{\beta})$ meaning that the OLS coefficients are no longer minimum variance.

(3) Pure heteroskedasticity causes biased OLS estimates of $SE(\hat{\beta})$. Although the bias is mostly negative in practice (underestimating the SE), positive bias (overestimating the SE), may also occur. The direction of bias cannot be determined beforehand.

10.3 Testing for heteroskedasticity

There is no universal test for heteroskedasticity. The reason is that heteroskedasticity may take different forms and it is difficult to determine which of these applies.

(a) Graphical method

Heteroskedastic error terms can often be detected by inspection of the residuals. See figure 10.4 in the textbook for a graphical presentation.

(b) Park test

The Park test tests for heteroskedasticity in its most simple form:

$$Var(\varepsilon_i) = \sigma^2 Z_i^{\ 2}$$

where ε_i is the error term of the equation being tested, σ^2 is the "base" variance of the homoscedastic error term and Zi is the proportionality factor.

The Park test consists of three steps:

Step	Procedure
1	Obtain the residuals of the estimated regression equation: $e_i = Y_i - \hat{Y}_i$ where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_k X_k$.
2	Use these residuals in a double-log regression against Z: $\ln(e_i^{\ 2}) = \alpha_0 + \alpha_1 \ln(Z_i) + u_i$ where Z is the proportionality factor. In practice, finding a proper Z is not an easy learning activity. See the textbook.

3	Test the significance of the slope coefficient H_0: $\alpha_1 = 0$ H_a: $\alpha_1 \neq 0$ by using a two-tailed test. If $\alpha_1 \neq 0$, then heteroskedasticity is indicated.

LEARNING ACTIVITY 1

Answer question 5(a) to (b) on p. 416 in the in prescribed book (the regression equation explains average consumption as a function of average income). Use a 5% level of significance for testing.

(a) The estimated equation is:

$$\hat{C} = 1273.2 + 0.72(INC)$$

(291.9) (0.044)

Where

C: average consumption

INC: average income

and the standard errors are in brackets below their coefficients. $\overline{R}^2 = 0.97$

(b) The estimated error terms are as follows:

1086.27
-511.72
-452.42
-342.39
-190.36
57.03
232.73
4.01
16.89

The (auxiliary) equation is estimated as

$$\ln(e_i^{\,2}) = 29.54 - 2.34\ln(Z_i) + u_i$$

and the SE(slope) = 0.94

To test for heteroskedasticity, we test the significance of the slope coefficient

H_0: $\alpha_1 = 0$ and H_a: $\alpha_1 \neq 0$

by using a two-tailed test. Since the absolute value of

$$t = \frac{-2.34}{0.94} = -2.48$$

exceeds the critical $t_{5\%,\ df=7, K=1,\ two\text{-}sided}$ = 2.365, H_0 can be rejected and the presence of heteroskedasticity can be assumed. Note that σ_i decreases as INC_i increases.

(c) White test

The White test for heteroskedasticity is more commonly used than the Park test. Although the Park test appears simpler because it uses only one proportionality factor Z_i, in practice the choice of this variable is not always obvious. The White test overcomes this problem by estimating a linear auxiliary equation of the following form:

e^2 = f(constant term, X_ks, X_k^2s, X_kX_l s)

where

e^2: square of residual term

X_ks: all explanatory variables

X_k^2s: the squares of the Xs and

X_kX_l : all cross-product terms of the Xs (which are sometimes omitted).

The White test for heteroscedasticity is performed as follows:

Step	Procedure
1	As for the Park test, obtain the OLS residuals of the estimated regression equation.
2	Estimate the auxiliary equation: e^2 = f(constant term, X_ks, X_k^2s, X_kX_l s). The test is based on a goodness of fit of this auxiliary equation. Since the large number of explanatory variables could be a problem, the cross-product terms are sometimes excluded from the auxiliary equation.
3	State the alternatives being tested: H_0: Error terms are homoskedastic H_a: Error terms are heteroskedastic. The test statistic is N.R^2 which is distributed like Chi-square$_k$ where k: the number of slope coefficients included in the auxiliary equation, R^2 is the fit of the auxiliary equation and N: the number of observations. If N.R^2 > Chi-square$_k$ then reject H_0 (heteroskedasticity is indicated) If N.R^2 ≤ Chi-square$_k$ then do not reject H_0 The better the fit of the auxiliary equation, the larger R^2 will be, and the more likely it will be that heteroskedasticity is indicated.

The other test for heteroskedasticity, the Goldfeld-Quandt test, is **not** prescribed.

LEARNING ACTIVITY 1

Test the residuals in table 2 on p85 of the prescribed book for heteroskedasticity at the 5% level of significance by using the White test (Woody's regression).

ANSWER

State the null and alternative hypotheses:

H_0: Error terms are homoskedastic

H_a: Error terms are heteroskedastic

The auxiliary equation is (lower-case characters which appear at the right-hand side of the equation denote coefficients):

$$e^2 = a + bP + cN + dI + eP^2 + fN^2 + gI^2 + h(P.N) + i(P.I) + j(N.I) + u$$

($N = 33$, $R^2 = 0.122$ and $K = 9$)

The auxiliary equation yields the test statistic: $N.R^2 = 33 \times 0.122 = 4.02$

The critical Chi-square$_{9,5\%}$ = 16.92 (see table 8 in the prescribed book). Because 4.02 < 16.92, we cannot reject H_0, which means that the null hypotheses cannot be rejected. We may thus assume that the error terms are homoscedastic.

Note also that the residual plot (p. 85 in the prescribed book) does not lend support to heteroske-dastic error terms in the sense that the variation in the residuals appears "normal".

10.4 Remedies for heteroskedasticity

(a) Impure heteroskedasticity

Since an incorrect specification, such as an omitted variable, may cause impure heteroskedasticity, this possible cause should be corrected first.

In some cases, the symptoms of heteroskedasticity may be due to measurement errors of the dependent variable. Pay close attention to all possible sources of measurement error, for example, the reputation of the institution which publishes the data, its date of publication and the primary source of the data.

(b) Pure heteroskedasticity

There are three methods of avoiding or overcoming the negative effects of heteroskedasticity, namely, weighted least squares, heteroskedasticity-corrected standard errors and redefining the variables.

(c) Weighted least squares

This method provides a method for getting rid of pure heteroskedasticity of the form

$$Var(\varepsilon_i) = \sigma^2 Z_i^{\ 2}.$$

The principle of this method is to divide the equation by the proportionality factor in order to convert the error terms to a constant variance type. Studenmund explains this method in section 10.4.1. Make sure you understand this and are able to explain it in the examination, using the respective equations.

(d) Heteroskedasticity-corrected standard errors

Since heteroskedasticity does not cause problems in the estimation of the β's but only in the estimation of their SE's, the approach of this method is to improve the estimates of the SE's. The details of this approach are beyond the scope of this course.

(e) Redefining the variables

In some cases it is possible to redefine the variables of the equation in order to get rid of heteroskedastic error terms. This method is often used in the case of cross-sectional data.

The textbook quotes an example of a model of total city expenditure (C), which is explained by total city income (Y):

$$C_i = \beta_0 + \beta_1 Y_i + \varepsilon_i$$

Because cities vary considerably in size, both C and Y occur over a very wide range. Heteroskedasticity is likely since the large cities are also likely to have much larger error terms than the smaller cities.

A possible solution to this problem is to use per capita income and per capita consumption data, rather than the absolute levels of C and Y. See the textbook.

LEARNING ACTIVITY 2

This learning activity continues where learning activity 1 ended. Answer part (c) of question 5 in the prescribed book (p. 417; average consumption as a function of average income). Estimate the WLS equation and compare its results to the OLS estimate (learning activity 1).

ANSWER

The original equation is

$$C_i = \beta_0 + \beta_1 (INC_i) + \varepsilon_i \qquad \textit{(equation 1)}$$

The WLS version (with INC as the Z factor) is

$$\frac{C_i}{INC_i} = \frac{\beta_0}{INC_i} + \beta_1 + \frac{\varepsilon_i}{INC_i} \qquad \textit{(equation 2)}$$

The coefficients are "switched". The constant term in (1) is estimated as a slope coefficient in (2), and the slope in (1) as the constant term in (2).

Equation 10.4.2 is estimated as

$$\left(\frac{\hat{C}_i}{INC_i}\right) = 2400.1\left(\frac{1}{INC_i}\right) + 0.40 \qquad \textit{(equation 3)}$$

(216.4) (0.044)

where the SE's are in brackets below their coefficients..

Comparison of estimates

If equation 3 is multiplied both sides by INC, it gives the form:

$$\hat{C}_i = 2400.1 + 0.40(INC_i) \qquad \textit{(equation 4)}$$

which facilitates a comparison to the OLS estimate (learning activity 1). Note that the standard errors of equation 3 apply unchanged to equation 4

Method	***Constant term***	***Slope: Coefficient of INC***	***SE(slope)***
OLS	*1273.2*	*0.72*	*0.044*
WLS	*2400.1*	*0.40*	*0.143*

The SE(slope) is much higher for WLS than for OLS. Since we know that OLS estimates of the SE(slope) are suspect, we may assume that the WLS estimates are better.

(C) TRUE/FALSE QUESTIONS (F) = FALSE (T) = TRUE

(1) It is good practice to inspect a data series originating from cross-sectional data for heteroskedasticity. (F)

Heteroskedasticity pertains to the distribution of the error term and the error term only occurs within an equation. Thus heteroskedasticity cannot occur in a data series itself. It is, however, true that equations which deal with cross-sectional data often display the problem of heteroskedasticity.

(2) An omitted variable can cause the problem of heteroskedasticity. (T)

(3) If the proportionality factor Z_i applies to a regression equation which displays heteroskedasticity, then the expected value of SE(ε_i) = $Z_i.\sigma$ and the SE(ε_i/Z_i) = σ.

$$SE(\varepsilon_i) = \sqrt{\sigma^2 Z_i^{\ 2}} = \sigma.Z_i$$

(T)

(4) In practice, the White test is more popular than the Park test because it assumes a less restrictive form of heteroskedasticity. (T)

(5) If heteroskedasticity is present (of the appropriate form), then the Park test will confirm that $\alpha_1 > 0$. (F)

The Park test tests for H_0: $\alpha_1 = 0$ versus H_a: $\alpha_1 \neq 0$. The sign of α_1 depends on the nature of the heteroskedasticity. We can have either an increasing or decreasing σ_i if Z_i increases.

(6) In the case of WLS it is possible to derive estimates of the slope coefficients of the original equation without any transformation being required. In the case of the constant term, a transformation is required. (T)

(D) EXAMINATION: PARAGRAPH QUESTIONS

(1) Explain the the nature of heteroskedasticity: the difference between (6)

- pure and impure heteroskedasticity
- the consequences of heteroskedasticity

(2) Explain how pure heteroscedasticity may be detected by using the Park test and/or the White test. (7)

(3) Explain how pure heteroskedasticity may be remedied by using the WLS method. Set out the theoretical foundations of WLS. (5)

(4) Explain a situation in which heteroskedasticity can be remedied by redefining the variables. (3)

(E) EXAMINATION: PRACTICAL QUESTIONS

You should be able to

- explain in theory how the Park test and/or the White test may be applied to test for the presence of heteroskedasticity
- remedy a heteroskedasticity problem in practical situations.

LEARNING UNIT 11

RUNNING YOUR OWN REGRESSION PROJECT

ECONOMETRICS IN ACTION

To conclude, let's take note of some views on econometrics.

Econometrics appears to be an important field in economics because the following Nobel Prizes have been awarded in econometrics.[14]

- Jan Tinbergen and Ragnar Frisch, 1969, for having developed and applied dynamic models for the analysis of economic processes.
- Lawrence Klein, 1980 for his computer modelling work.
- Trygve Haavelmo, 1989. His main contribution was his 1944 article "The Probability Approach to Econometrics".
- Daniel McFadden and James Heckman shared the award in 2000 for their work in microeconometrics.
- Robert Engle and Clive Granger received the award in 2003 for work on analysing economic time series.

One can, however, overestimate the value of econometrics.[15]

Two beliefs of econometricians were examined:

(1) Econometric methods provide more accurate short-term forecasts than do other methods; and

(2) more complex econometric methods yield more accurate forecasts.

- A survey of 21 experts in econometrics found that 95% agreed with the first statement and 72% agreed with the second. A review of the published empirical evidence, however, yielded little support for either of the two statements.
- It appears that most econometricians believe that their method of analysis is superior. However, these beliefs are not supported by empirical evidence. Perhaps it is natural for econometricians to avoid evidence that may disconfirm their beliefs.

[14] Wikipedia, the free encyclopedia (web based).

[15] Armstrong, Scott. 1978. Forecasting with Econometric Methods: Folklore versus Fact. *Journal of Business*, 51 (4).

STUDY OBJECTIVES

This chapter is a revision of the previous chapters. After you have studied this learning unit you should

- have a better overview of the practice of econometrics
- better understand the regression user's checklist
- better understand all the econometric problems, their consequences, their detection and how to correct them

(A) PRESCRIBED MATERIAL

The following sections are prescribed:

(1) choosing your topic
(2) collecting your data
(3) advanced data sources
(4) practical advice for your project
(5) writing your research report
(6) a regression user's checklist and guide
(7) the housing price interactive exercise is useful to gain a better feel of econometrics

(B) SOME IMPORTANT CONCEPTS

There are no new concepts in this chapter as far as theory is concerned. This chapter has a practical orientation. The checklist provided in table 2 on p378 provides a useful framework when reviewing regression results. Table 3 summarises the major problems which may occur in econometrics, their consequences, their detection and remedial actions that can be taken. In practice more than one problem might well occur.

Some useful comments are made on the use of data. Although these may be viewed as common sense, the issues of averaging, changes in the quality of products, nominal and real values and deflators are important from a practical perspective.

(C) TRUE/FALSE QUESTIONS (F) = FALSE (T) = TRUE

(1) For an equation to be estimated, it is required that the degrees of freedom: n-K-1 > 0. (T)

The higher the degrees of freedom, the better, since the accuracy of estimates will improve. In the case of n-K-1 = 0, OLS does not work, since the standard errors cannot be derived. This denotes an exact fit, since the number of equations is exactly equal to the number of unknown coefficients.

(2) It is always better for K to be as large as possible. (F)

The "best" model is generally one which is as simple as possible, but still captures the essential elements of the situation being studied.

(3) Generally speaking, the larger the number of models that have been explored empirically, the more reliable the final model will be. (F)

A good model should be based on both theoretical and empirical considerations. Note that the practice of data-mining has undesirable consequences.

Section 8 (House price model): Some aspects regarding specification are dealt with below.

(4) A hedonic model seeks to explain price by the interaction between supply and demand. (F)

A hedonic model is not a price model based on supply and demand. It only includes variables which directly reflect the quality (and by implication the price) of the product.

(5) There are theoretical reasons why the sign of SP (pool dummy) could be either positive or negative. (T)

A positive sign is justified on the grounds of the additional facility provided by a pool. A negative sign could, however, also be indicated since a pool requires maintenance, and not all people would consider a pool to be an asset.

(6) There could be problems using BE and BA as independent variables together with S. (T)

It is highly likely that the number of bedrooms (BE) and the number of bathrooms (BA) are positively related to the size of the house (S). This could cause the problem of multicollinearity.

(7) The rationale for including both A and A^2 as independent variables is that the relation between Price (P) and age (A) is not linear, but that P increases as A increases, but at a decreasing rate. (F)

The use of both A and A^2 implies the use of a quadratic function, which indeed implies a non-linear relationship between P and A. However, it is rather the decreasing part of this curve which applies in this case, because price can be expected to decrease with an increase in age.

(8) Including an N x S interaction variable implies the use of a slope dummy variable. (T)

Correct. A slope dummy is represented by an interaction variable. The presence of both variables causes an additional price-increasing effect beyond that which may be explained by their intercept dummies alone.

(9) The use of the transformed variable 5-N is preferable to that of N. (T)

The expected sign of N is negative, because N = 1 designates the best neighbourhood and N = 4 the worst. To change the expected sign of neighbourhood, it may be transformed to 5-N, in which case a 1 indicates the worst neighbourhood and a 4 the best. This transformation makes it easier to interpret the interaction variable.

(D) EXAMINATION: PARAGRAPH QUESTIONS

None

(E) EXAMINATION: PRACTICAL QUESTIONS

In the examination we expect you to be able to

- review regression results (table 11.2 in Studenmund is useful in this regard)
- identity any of the econometric problems (table 11.3 in Studenmund is useful in this case)
- deal with basic issues relating to data (for example, we expect of you to understand the meaning of nominal and real values, and of a deflator)

We recommend that you work through the housing price interactive exercise.

APPENDIX 1: FORMULA SHEET

OLS estimates of $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$

$$\widehat{\beta}_1 = \frac{\sum x_{1i} y_i}{\sum x_{1i}^{\,2}} \text{ where } \begin{array}{l} x_{1i} = X_{1i} - \overline{X}_1 \\ y_i = Y_i - \overline{Y} \end{array}$$

$$SE(\hat{\beta}_1) = \sqrt{\frac{\left(\dfrac{\sum e_i^{\,2}}{n-2}\right)}{\sum x_{1i}^{\,2}}}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}_1 \text{ where } \overline{Y} = \sum Y_i \big/ n$$

TSS (Total sum of squares) = ESS (explained) + RSS (residual)

$$\sum y_i^{\,2} = \sum \left(\hat{Y}_i - \overline{Y}\right)^2 + \sum e_i^{\,2}$$

OLS estimates of $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$

$$\widehat{\beta}_1 = \frac{\left(\sum y_i x_{1i}\right)\left(\sum x_{2i}^{\,2}\right) - \left(\sum y_i x_{2i}\right)\left(\sum x_{1i} x_{2i}\right)}{\left(\sum x_{1i}^{\,2}\right)\left(\sum x_{2i}^{\,2}\right) - \left(\sum x_{1i} x_{2i}\right)^2} \quad x_{2i} = X_{2i} - \overline{X}_2$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}_1 - \hat{\beta}_2 \overline{X}_2$$

$$SE(\widehat{\beta}_1) = \sqrt{\frac{\sum e_i^{\,2} \big/ (n-3)}{\left(\sum x_1^{\,2}\right)\left(1 - r_{X_1 X_2}^{\;\;2}\right)}}$$

$$r_{X_1 X_2} = \frac{\sum \left(X_{1i} - \overline{X}_1\right)\left(X_{2i} - \overline{X}_2\right)}{\sqrt{\sum \left(X_{1i} - \overline{X}_1\right)^2 \left(X_{2i} - \overline{X}_2\right)^2}}$$

Some statistical measures

$$t = \frac{\widehat{\beta} - \beta_{H_o}}{SE(\widehat{\beta})}$$

$$DW\ d = \frac{\sum_{t=2}^{T} \left(e_t - e_{t-1}\right)^2}{\sum_{t=1}^{T} e_t^{\,2}}$$

$$F = \frac{ESS / K}{RSS / (n - K - 1)}$$

This formulas sheet will be supplied in the examination.